2020

An Analysis of Teacher Decision-Making in Grading 10th Grade Student Writing in English Language Arts

Guy Swenson
University of Central Florida

AN ANALYSIS OF TEACHER DECISION-MAKING IN GRADING 10TH GRADE STUDENT WRITING IN ENGLISH LANGUAGE ARTS

by

GUY SWENSON
B.S. Indiana University, 1994
B.A. Indiana University, 1997
M.Ed. Stetson University, 2005

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Education
in the Department of Educational Leadership and Higher Education
in the College of Community Innovation and Education
at the University of Central Florida
Orlando, Florida

Spring Term
2020

Major Professor: RoSusan D. Bartee

ABSTRACT

This qualitative study was conducted to investigate teacher decision-making while grading samples of 10th grade student writing in English language arts. Extensive research spanning 100 years has shown that inter-rater agreement of student work is weak at best (Ashbaugh, 1921; Brimi, 2011; Starch & Elliot, 1912). A cognitive laboratory interview method was chosen to focus on real-time teacher decision-making due to a discrepancy between teacher beliefs and teacher practices (Phipps & Borg, 2009). Qualitative data were gathered from 21 cognitive laboratory interviews in which the participants graded two samples of student writing while verbalizing their thoughts. The grading data revealed discrepant scores with a range of 40 points and 25 points for each student essay. The findings revealed that participants interrupted their reading of student work to consider the conventions of standard English, the thesis, or to ask themselves or the imaginary student questions about the writing. The differences were that participants' interruptions focused on the conventions or the thesis, but not both, and participants either made binary or quality decisions regarding the thesis and conventions. Furthermore, participants exhibited an evaluation focus either on the thesis or the conventions of standard English, but not both. A substantive grounded theory emerged from the qualitative data: The Theory of Disparate Purposes of Writing Assessment. This emergent theory states that teachers' grading practices indicate the purpose of student writing is for the student either to demonstrate the five-paragraph essay or for the student to express something important in their writing. The theory offers an explanation for the differential application of rubrics and for the lack of rater agreement in student writing. These findings may inform teachers, school district leaders, and teacher preparation programs in ways to improve writing assessment practices and instruction.

To my parents, Robin and Judy Swenson, who instilled in me

a strong work ethic and a desire to learn more each day.

ACKNOWLEDGMENTS

This milestone would not have been possible without the encouragement and support of many individuals. To my committee chair, Dr. RoSusan Bartee, thank you for your guidance and support and for your willingness to lead this committee already in progress. To my committee members, Dr. Brandon McKelvey, Dr. Elsie Olan, and Dr. Sheila Moore, thank you so much for your feedback and encouragement. And to Dr. Jerry Johnson who helped me begin this dissertation and took my small idea and gave it a strong direction. Thank you all.

To the teachers who opened their classrooms and shared their grading practices, thank you for your time and effort. Your eagerness to participate in my research is admirable and appreciated.

To my Cohort 8 members and especially the library crew who met nearly every Saturday during this process. Thank you Krissy, Maria, Mary, and Dave for your assistance, strength, camaraderie, and laughs. I will always remember those days with great fondness.

Finally, to my husband David Swenson, I would like to thank you for everything. Your infinite patience, understanding, and willingness to take on more than your share will forever be appreciated.

TABLE OF CONTENTS

LIST OF TABLES xii

CHAPTER 1 INTRODUCTION 1

Background of the Study 1

Statement of the Problem 3

Purpose of the Study 5

Significance of the Study 6

Theoretical Framework 7

Categorization Theory 7

Grounded Theory 12

Research Questions 13

Definition of Terms 14

Limitations of the Research Study 15

Delimitations of the Research Study 16

Assumptions of the Research Study 16

Organization of the Research Study 17

CHAPTER 2 REVIEW OF THE LITERATURE 18

Introduction 18

Historical Context of Grading 21

Writing Assessment 23
Holistic Rating 24
Assessment Literacy 26
Rater Agreement 28
Inter-Rater Agreement 29
Variation in Grading – The Beginning 30
Lack of Common Grading Practices and Scales 32
Variation in Grading – 100 Years Later 33
Intra-Rater Agreement 36
Causes of Inter-Rater Disagreement 38
Rater's Comments 38
The Rubric 41
Rater Effects 42
Teacher Decision-Making Research 45
Think-Aloud Method 46
In-Context and Out-of-Context Writing 47
Experienced vs. Inexperienced Raters 50
Assessment as an Instructional Tool 53
Summary of the Literature Review 54

CHAPTER 3 METHODOLOGY 57
Introduction 57
Research Design 58
Categorization Theory 58
Grounded Theory 59
Cognitive Laboratory Interview 60
Selection of Participants 62
Instrumentation 64
Writing Prompt 64
Researcher as Instrument 66
Data Collection 67
Pilot Study 67
Procedures 68
Data Analysis 70
Validation and Credibility 73
Peer Review 73
Negative Case Analysis 74
Rich, Thick Description 74
Respondent Validation 74

Summary of Chapter 3 75

CHAPTER 4 ANALYSIS OF THE DATA 76

Introduction 76

Participants 77

Grading Data 80

Coding 82

Research Questions 93

Research Question 1: Classical Categorization 94

Research Question 2: Prototype Categorization 96

Research Question 3: Exemplar Categorization 102

Research Question 4: Similarities and Differences in Grading 106

Similarities 107

Differences 117

Research Question 5: Differences in Grading High and Low Essays 137

Additional Analyses 146

Summary 153

CHAPTER 5 SUMMARY, DISCUSSION, AND CONCLUSIONS 156

Introduction 156

Summary of the Study 156

Discussion of the Findings 158
Grading Data 158
Research Question 1: Classical Categorization 160
Research Question 2: Prototype Categorization 161
Research Question 3: Exemplar Categorization 163
Research Question 4: Similarities and Differences in Grading 164
Similarities 165
Differences 168
Research Question 5: Differences in Grading High and Low Essays 172
Additional Analyses 173
Emergent Theory: The Theory of Disparate Purposes of Writing Assessment 174
Implications for Practice 176
High School Teachers 176
School District Leaders 177
Teacher Preparation Programs 178
Recommendations for Further Research 178
Conclusions 180
APPENDIX A MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM 2018 GRADE 10 COMPOSITION RUBRIC 184

APPENDIX B 2018 MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM STUDENT WRITING SAMPLE – GRADE 10 ENGLISH LANGUAGE ARTS STANDARD ENGLISH CONVENTIONS – SAMPLE 1 – SCORE 3 187
APPENDIX C 2018 MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM – SCORING GUIDE FOR STANDARD ENGLISH CONVENTIONS – SAMPLE 1 – SCORE 3 189
APPENDIX D 2018 MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM STUDENT WRITING SAMPLE – GRADE 10 ENGLISH LANGUAGE ARTS TOPIC/IDEA DEVELOPMENT – SAMPLE 2 – SCORE 5 191
APPENDIX E 2018 MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM – SCORING GUIDE FOR TOPIC/IDEA DEVELOPMENT – SAMPLE 2 – SCORE 5 194
APPENDIX F UNIVERSITY OF CENTRAL FLORIDA INSTITUTIONAL REVIEW BOARD APPROVAL 196
APPENDIX G SCHOOL DISTRICT NOTICE OF APPROVAL TO CONDUCT RESEARCH 198
APPENDIX H INFORMED CONSENT 201
APPENDIX I COGNITIVE LABORATORY INTERVIEW PROTOCOL 204
APPENDIX J DEMOGRAPHIC DATA COLLECTION FORM 207
APPENDIX K PRACTICE ESSAY 209
REFERENCES 211

LIST OF TABLES

Table 1 *Inter-rater Agreement Studies* 36

Table 2 *Teacher Decision-making Studies* 52

Table 3 *Research Questions, Data Source and Coding Type* 72

Table 4 *Demographics of 21 Participants* 79

Table 5 *Scores and Letter Grades for Essays 1 and 2* 81

Table 6 *Codes Converted from Topics to Gerunds* 83

Table 7 *Coding* 85

Table 8 *Emergent Themes* 92

Table 9 *Prototype Categorization Coding* 98

Table 10 *Participant's Years of Experience and Exemplar Categorization* 103

Table 11 *Focused and Thematic Codes for Interruptions* 109

Table 12 *Participant Interruptions to Consider Conventions or Thesis* 120

Table 13 *Participant Evaluative Statements Regarding Conventions or Thesis* 122

Table 14 *Evaluation Focus on the Thesis or the Formula* 128

Table 15 *Participant Thesis Evaluations: Binary or Quality* 130

Table 16 *Research Question 4: Similarities and Differences* 137

Table 17 *Participant Decisions about Addressing the Prompt* 139

Table 18 *Essay Scores Organized by Participant Evaluation Focus* 147

Table 19 *Evaluation Focus and Writing Prompt Decision-Making for Essay 1* 151

CHAPTER 1
INTRODUCTION

Background of the Study

Grades are an essential aspect of the secondary school classroom. Teachers instruct their students in the curriculum; students demonstrate their mastery on a set of skills; and then, teachers assess their students' work. Grading is important because of the "centrality of grades in the educational experience of all students" (Brookhart et al., 2016, p. 3). Whether the student is in kindergarten or graduate school, private or public school or in an urban center or a small rural school, each of those students will be subject to remarkably similar grading experiences. Yet, grading is most often rooted in tradition and our own schooling experiences rather than relevant research (Guskey & Brookhart, 2019). Grades are not just a report on past academic achievements. Rather, grades have become increasingly important as indicators of future success in post-secondary institutions (Sawyer, 2013). Even with the increased amount of achievement data from standardized testing, grades have been shown to be a better predictor of college success than standardized test scores (Atkinson & Geiser, 2009). The importance of the grades students receive is essential to their future success in college and beyond.

Guskey (2004) stated, "Grading is one of a teacher's greatest challenges and most important professional responsibilities. However, few teachers have any formal training in grading methods, and most teachers have limited knowledge about the effectiveness of various grading practices" (p. 31). This lack of training and understanding has led to what Brookhart (1991) described as a "hodgepodge grade of attitude, effort, and achievement" (p. 36). Combine the lack of training with the increased role grades play in a student's future, and the need for

more understanding in this area becomes clear. Yet, with the numerous educational reforms, grading is the last reform to be addressed because it means educators must tackle one of the longest held traditions in school (Guskey & Brookhart, 2019). One of the most powerful forces in a school is tradition and few educational leaders are willing to challenge it.

Complicating the issue even further, the assessment practices teachers use have not followed any accepted assessment and grading principles held by the assessment and measurement professionals (Cizek, Fitzgerald, & Rachor, 1996). In a study by Tierney, Simon, and Charland (2011), only a small number of teachers self-identified they had a high level of awareness and use of accepted grading principles. The results of this study further indicated that "Although many teachers in this sample reported at least some awareness and use of grading principles, they had difficulty identifying relevant principles, and the grading practices they reported suggested that underlying principles were not well-understood" (Tierney, Simon, & Charland, 2011, p. 222).

By gaining a better understanding of the assessment decision-making rationale and practices, measurement specialists could "reconceptualize assessment principles and suggested practice in ways that further teachers' goals for their students" (McMillan, 2003, p. 35). The decision-making process of teachers is important because assessment and instruction are inextricably linked with each other. Assessment informs instruction and, in turn, instruction informs assessment, both of which lead to increased student learning (McMillan, 2003). At the center of this circular motion between instruction and assessment are the decisions that teachers make in their classrooms daily.

The grading of student work is more than just a tallying of points. It is a decision-making process about the student's level of achievement (Newton, 2007). Brookhart (2003) and McMillan (2003) have indicated a need for a new assessment theory concept that bridges the gap between current assessment theory and the practices of classroom teachers. It is imperative to study and understand the grading decisions teachers make for this new assessment concept to be realized. According to McMillan (2003), "a greater understanding of this [grading] process may help measurement specialists adapt and apply important assessment concepts and principles to what teachers engage in on a regular basis" (p. 35). With new assessment concepts and principles, the grades teachers assign to student work could become more reliable in describing student achievement and predicting of student success in the future.

Statement of the Problem

Education reform has increased the focus on improving student outcomes through standards-based instruction accompanied by a significant reliance on high-stakes testing. Although these reforms have continued to develop and mature, grading reform has failed to materialize in any significant way inside the classroom (Guskey, 2009), and this may be a direct result of a lack of teacher education in assessment and measurement (Schneider & Bodensohn, 2017). Researchers have shown that teachers tend to replicate grading practices they experienced as students, leading to a continual reliance on traditional grading practices (Grainger & Addie, 2014; Guskey & Brookhart, 2019). Guskey (2011) indicated there are five grading concepts or practices included in traditional grading. They include (a) the purpose of grades is to differentiate the achievement of students, (b) grade distributions should be bell-shaped curves,

(c) a student's grade should indicate a relative standing to their peers, (d) poor grades motivate students to work harder, and (e) one grade should be assigned to a student for each course (Guskey, 2011). These traditional grading practices have changed little since the advent of grading research which began in the early part of the 20th century.

The genesis of this researcher's interest in teacher grading came from reading the results of an early study in which a group of secondary teachers graded two samples of student writing (Starch & Elliot, 1912). The grades given to the essays showed a variance of 34 points on one paper and 49 points on another. On a modern-day grading scale this would be the equivalent of one teacher assigning the essay an A grade and another teacher assigning the same essay an F. The question in this researcher's mind was "How can there be such variability in the grades given by education professionals?" Surely, the answer would lie in the varying grading practices and procedures of the early 20th century, but that was not the case. This study was replicated nearly 100 years later by Brimi (2011), who utilized a highly developed rubric and was supported by extensive training. The results were nearly identical to those found in the earlier study - there was a high degree of variability that included grades of A through F for the same student essay.

Ashbaugh, in a 1924 mathematics study, found even greater variances in grades given to student work. Teachers even gave different grades when asked to reevaluate the same student work several weeks later (Ashbaugh 1924; Eells, 1930). The variance in grading the same student work lends credence to the arbitrary nature of teacher grades. Furthermore, Diederich, French, and Carlton (1961) and Australian researchers (Cooksey, Freebody, & Wyatt-Smith, 2007) indicated similar variances in college freshman writing and 5th grade student writing

respectively. Given the varying decisions teachers must make when determining assessment grades, it is not surprising that researchers have shown that the inter-rater agreement of grades on individual assessments of student work is weak at best (Ashbaugh, 1921; Brimi, 2011; Starch & Elliot, 1912). It is this large degree of variability that led this researcher to investigate teacher decision-making on individual writing assessments in English language arts.

Twentieth century student assessment studies have shown a lack of inter-rater agreement in student assessment. Additionally, there has been a lack of understanding and a lack of research about the decision-making processes teachers use to grade samples of student work (Wyatt-Smith, 1999). The benefits of an increased focus on standards-based instruction cannot be fully realized if teachers are unable to accurately recognize and assess if students have met the standards and if teachers have not given the appropriate critical feedback in order to increase student outcomes on the assessment tasks. Wyatt-Smith (1999) noted the lack of research regarding the teacher decision-making process in assessment and recognized the need for more research in the area of teacher judgment. Thus, the focus in the present research was on the decision-making processes teachers use in grading student writing in 10th grade English language arts.

Purpose of the Study

The purpose of this qualitative study was to investigate the decisions teachers made while grading samples of decontextualized student writing in English language arts using a cognitive laboratory interview method. The data collected focused on the decisions teachers made as they assessed student writing, rather than the judgments recorded as scores or letter grades.

Furthermore, this study was conducted to determine if there were any commonalities in decision-making among teachers as well as if there were any differences in decision-making among the teachers sampled.

Significance of the Study

This study added to the literature regarding teacher decision-making in grading and assessment because it focused on real-time decision-making using a cognitive laboratory interview method to capture teachers' thoughts while grading samples of student writing in 10th grade English language arts. Harris (1977) found that the criteria teachers reported to be important for grading student writing were not the criteria the same teachers utilized when grading student writing. In fact, there was nearly an inverse relationship between the self-reported criteria and the criteria the teachers used. Furthermore, Phipps and Borg (2009) indicated that because of the discrepancy between teacher beliefs and teacher practice, "studies which employ qualitative strategies to explore language teachers' actual practices and beliefs will be more productive (than, for example, questionnaires about what teachers do and believe) in advancing our understanding of the complex relationships between these phenomena" (Phipps & Borg, 2009, p. 388). Due to this discrepancy between beliefs and actions, it is necessary to directly study teachers' actions while grading student writing. According to Ericsson (2003), data gathered in a cognitive laboratory interview as the teachers made their grading decisions in real-time provided a strong path for analyzing the cues used in the basis for their judgments. Therefore, the present study focused on investigating the actions teachers took while grading student writing without regard to their stated beliefs.

Great strides in the theory of standardized assessments have been made in the last 75 years, but those same advancements have not been seen in teacher-made classroom assessments. "The time has come to develop measurement theory for classroom assessment purposes and uses" (Brookhart, 2003, p. 5). By gaining a better understanding of the assessment decision-making rationale and practices, the results of this study could be used in designing new assessment principles and practices for the classroom.

Theoretical Framework

Theoretical frameworks have an important place within qualitative research in that they guide the research (Saldaña, 2014). Two applications of theory in qualitative research are (a) they provide a framework and method for research (Anfara & Mertz, 2015) and (b) they allow for building a theory as a result of data analysis (Jaccard & Jacoby, 2010). This research study used two theoretical frameworks: categorization theory and grounded theory. The purpose of using two frameworks was to analyze the data through the lens of categorization theory and also to construct a theory based upon the data using grounded theory. A discussion of both frameworks is included in the following sections.

Categorization Theory

Categories are essential to cognition because they are needed for analogy, causal reasoning, memory, imagination, creativity, generalization, and prediction (Yamauchi, Love, & Markman, 2002, p. 585). Categorization is a cognitive process and a form of decision-making in which the essential functions of decision-making are used (Seger & Peterson, 2013). Seger

(2009) stated that categorization is a process that involves viewing a stimulus and then determining the category in which the stimulus belongs. There are many similarities between decision-making and categorization, but they also differ in significant ways. Simple decisions result in an evaluation such as the desk is made of wood or the meeting occurred in the conference room. Yet other more complex decisions involve evaluating several types of stimuli and determining a course of action such as which contractor to hire or which major to choose (Seger & Peterson, 2013). Seger and Peterson further stated that categorization is similar to decision-making in that a stimulus is viewed and several candidate categories may be evaluated. Eventually, one of the categories is chosen.

The primary difference between decision-making and categorization is generalization which is defined as "any extension to a stimulus that is novel or changed in any way" (Seger & Peterson, 2013, p. 1188). These changes can be either incremental or significant, abstract changes to the stimulus. Simple categorization tasks such as determining if a piece of furniture is a chair require minimal generalizations (Seger & Miller, 2010). More significantly complex generalizations are needed to evaluate stimuli that are closely related but do not share all the characteristics of the previously studied stimuli. These new situations require a transfer of the knowledge and the use of generalizations to determine the category. This ability to apply knowledge to new situations is an extremely important skill (Casale, Roeder, & Ashby, 2012). Furthermore, the importance of learning to categorize is that this skill forms the foundation for inference. Knowing that an item belongs to a category allows for additional characteristics to be inferred (Yamauchi et al., 2002). These inferences are an important aspect of categorization since new stimuli are presented and do not necessarily match previously experienced stimuli.

Categorization theory (CT) is comprised of three sub-categories developed over various time periods by several philosophers and researchers. Haswell (2001) defined categorization theory as "the basic cognitive procedure of sorting things into conceptual boxes" (p. 57) and applied CT to the study of writing assessment. He further stated the act of assessing writing was in fact categorization because giving the student writing a score of 5 or "doesn't meet proficiency" is an act of placing the essay into a category (Haswell, 1998). Moreover, once the category has been assigned, inferences about the student writing or the student writer could be made (Yamauchi et al., 2002). Although Haswell (2001) applied CT to college placement writing in which a student sample of writing was used to place the student into the appropriate college writing course, Lynne (2004) stated that further study in other contexts is warranted. Haswell (1998) defined CT as having three types of categorization: (a) classical, (b) prototype, and (c) exemplar, each of which is discussed in the following paragraphs.

Haswell (1998, 2001) identified classical categorization as a model in which people analyze the features of new stimuli and match the features to the properties that define the category. Each category is well defined with a specific set of features in which each feature is a necessary component that defines the category. As an example, a table has a horizontal surface, legs, and is about waist high. Each of those features are necessary and are required for membership in the category. However, there is no reason to question the materials used to make the table or the number of legs. Those features do not define what a table is even though they can vary in an infinite number of ways. Haswell (1998) further stated it is not necessary to compare individual members within the category because each member shares the same characteristics.

Prototype categorization is the second of three types of categories within CT. Prototype categorization is comparing an object to "abstract schemas [people] have of the best example or most representative member [prototype] of possible categories. The prototype of a category is not a specific member but an idealized construction" (Haswell, 1998, p. 246). Although the primary attribute in classical categorization is correctness, in prototype categorization it is gradience. The prototype as an idealized version of the category is compared to the new instance and then evaluated on how closely the new member resembles the prototype. Unlike classical categorization, no member of the prototype category needs to contain any or all the features of that category (Haswell, 2001). For instance, the category of fruit is an example of a prototype category. Strawberries or blueberries are idealized prototypes of the fruit category and they would be considered closer versions to the prototype than tomatoes or pumpkins even though they are equal members of the category of fruit.

The third type of categorization is exemplar categorization in which people categorize a new stimulus by comparing it to something in their recent memory (Haswell, 1998). As people gain more experience, they have more exemplars from which to draw. In a task such as choosing a movie that a spouse would like, there may be instances of classical categorization such as science fiction, or a foreign film. There could also be prototype categorization in which the movie could be funny or scary. And finally, a person could be categorizing the film as one his or her spouse would like to see based on past exemplars (e.g., spouse did not like the movie with subtitles, but did like the movie with a favorite actor). Ashby and Alfonso-Reese (1995) proposed prototype categorization as one single representation that includes all items; on the

opposite end of the spectrum is exemplar categorization which includes multiple representations held in memory.

It is worth noting that although these categories seem well defined with distinct boundaries, people often apply a prototypical response to an obvious classical category (Armstrong, Gleitman, & Gleitman, 1983). Armstrong et al. studied categorization of odd and even numbers which are classical in nature. A number is either odd or even because a number cannot be "more odd" or "more even" than another number. Yet people judged 3 as a better example of an odd number than 501. Both numbers are clearly odd in a classical category sense, yet people applied a prototype categorization to the numbers. In another similar instance, the word "mother" was judged as a better example of a female than the word "comedienne" (Armstrong et al., 1983). Haswell asserted that, as Armstrong et al. found, people could not maintain classical categorization with even and odd numbers, it was unlikely that raters would be able to do so with something as complex as student writing.

When grading writing, a rubric most closely resembles classical categorization (Haswell, 2001). The rubric is sectioned into various score levels with each level containing a list of attributes that would be expected from a sample of student writing for that sample to be a member of that category. However, Haswell maintained it was impossible to find anchor papers that match all the attributes on any level of a rubric, stating: "It is easy to find an essay one scale-point better than another essay in two or three of the subskills, but almost impossible to find one better across the board" (Haswell, 1998, p. 242). Moreover, because of the complex nature of writing samples and their inability to be consistent members of one category, raters have tended to use prototype and exemplar categorization while they are assessing samples of

student writing. Therefore, raters with more experience, and thus more exemplars at their disposal, would be able to make better assessment decisions (Lynne, 2004). The importance of an expansive catalogue of exemplars is essential to the accurate assessment of student writing in English language arts.

Broad (2000) noticed an issue regarding the "weird cases" (p. 232) that represent specific types of problems or are "too perfect" (p. 236) in some way. This suggests that perhaps even experienced raters may have difficulty matching student samples of writing to exemplars within their memories or to an anchor paper.

Grounded Theory

It was necessary to include an additional conceptual framework to allow for the construction of an original analysis to the data collection along with the categorization theory. The second conceptual framework is grounded theory which is a qualitative analysis method that derives an explanation of a process based upon the data (Corbin & Strauss, 2008). According to Charmaz (2014) "grounded theory methods consist of systematic, yet flexible guidelines for collecting and analyzing qualitative data to construct theories from the data themselves" (p. 1). Furthermore, Charmaz stated the data is gathered qualitatively through observations, interactions, and materials. Beginning with the data allows the researcher to construct a theory that is based upon the data. This openness to the situation the data presents also allows the researcher to develop themes inductively rather than only using a predetermined framework upon the data. Charmaz (2014, p. 10) describes a finished grounded theory as,

> Thus, for [Glaser and Strauss], a finished grounded theory explains the studied process in new theoretical terms, explicates the properties of the theoretical categories, and often demonstrates the causes and conditions under which the process emerges and varies, and delineates its consequences.

Therefore, grounded theory is the construction of a new theory based upon the observation of events and situations that cannot be described within existing theory.

Research Questions

In order to understand teachers' grading decisions better, this study focused on the following research questions.

1. In what ways and to what extent do teachers use classical categorization in their grading decision-making process when grading samples of student writing in English language arts?
2. In what ways and to what extent do teachers use prototype categorization in their grading decision-making process when grading samples of student writing in English language arts?
3. In what ways and to what extent do teachers use exemplar categorization in their grading decision-making process when grading samples of student writing in English language arts?
4. To what extent, if any, do similarities or differences exist in teachers' decision-making processes when grading samples of student writing in English language arts?

5. To what extent, if any, do teachers' decision-making processes differ when grading student writing samples of high and low performance levels in English language arts?

The research questions were developed using the two theoretical frameworks utilized in this study. Categorization theory provides the basis for the first three research questions and grounded theory provides the basis for questions four and five.

Definition of Terms

Assessment. A large-scale measurement used for institutional purposes such as exit and placement examinations (Speck & Jones, 1998).

Classical categorization. Sorting objects into rigid, clearly defined categories based on rules (Haswell, 2001).

Exemplar categorization. Comparing the object to a recent memory of an example or examples in the category (Haswell, 2001).

Grading. The process of judging the quality of piece of student work and assigning a value in either numerical, letter grade, or descriptor form (Guskey, 2009). Furthermore, grading is specific to the classroom and measures the performance on a single assignment (Speck & Jones, 1998).

Generalization. Any extension to a stimulus that is novel or changed in any way (Seger & Peterson, p. 1188).

In-context writing. Writing samples created in the teacher's classroom in which the students are known to the teacher (Cooksey et al., 2007).

Inter-rater agreement. "The degree to which a rater assigns scores to a set of examinee responses that are consistent with scores assigned to those responses by other raters" (Wolfe, Song, & Jiao, p. 2).

Interview probes. Probing questions used in a cognitive laboratory interview to prompt the subject to elaborate, explain, or clarify a response (Willis, 2015).

Out-of-context writing. Writing samples created outside of the teacher's classroom in which the students are not known to the teacher (Cooksey et al., 2007).

Prototype categorization. Classification of objects based upon how similar they are to a mental image of a prototype of that group (Haswell, 2001).

Rater effects. The patterns within the scores assigned by the rater which exhibit a degree of predictability which contribute to low rater agreement (Wolfe et al., p. 2).

Traditional grading. Grading concepts or practices including (a) differentiation of students, (b) bell-shaped curves, (c) relative standing to their peers, (d) poor grades as motivators, and (e) one letter grade assigned for each course (Guskey, 2011).

Limitations of the Research Study

This study contained four limitations worth noting. The qualitative research design did not allow causal conclusions to be drawn nor did it allow for generalizability outside of the subjects being studied (Fraenkel et al., 2015). Furthermore, as this researcher coded and interpreted the data alone, there was a possibility of researcher subjectivity in the findings (Creswell, 2007). To minimize this subjectivity, this researcher took several steps to increase trustworthiness and limit subjectivity in the findings. Additionally, participants may reveal only

what they are willing to reveal in a cognitive laboratory interview (Alshenqeeti, 2014). Finally, language will be viewed as a neutral means to "capture accurate responses from participants" (Hennink, 2008, p. 23). Therefore, cultural nuances regarding language and communication will not be explored.

Delimitations of the Research Study

In order to focus on grading practices teachers use while grading samples of student work in English language arts, this study was delimited to include English language arts teachers who were teaching 9th or 10th grade English during the 2019–2020 school year. By focusing on 9th and 10th grade teachers, there were specific writing standards and processes teachers taught during this course. Therefore, there was a common curriculum followed by each teacher that included writing. The teachers in the study were selected from the traditional high schools within a large urban public school district in the Southeastern United States.

Assumptions of the Research Study

This study included the following assumptions. Teachers had a sincere interest in participating in the study without any other motivation for personal gain. The selected teachers were truthful in their responses and the data accurately represented decision-making the teachers used while grading samples of student writing in 10th grade English language arts.

Organization of the Research Study

This research study has been organized into five chapters. Chapter 1 describes the background of the study, statement of the problem, purpose of the study, and the significance of the study. The theoretical framework consisted of categorization theory and grounded theory. Also included were the research questions, definition of terms, limitations, delimitations and assumptions of the study. Chapter 2 is a review of the relevant literature used to inform the research study. It includes a discussion about the historical context of grading, writing assessment, rater agreement, causes of inter-rater disagreement, teacher decision-making research, and assessment as an instructional tool. Chapter 3 describes the methodology including the research design, instrumentation, data collection, data analysis, and validation and credibility. A discussion of the results of the study is presented in Chapter 4. Chapter 5 concludes with a discussion of the findings including recommendations for further research.

CHAPTER 2
REVIEW OF THE LITERATURE

Introduction

Teachers use many tools in their classrooms including standards, technology, textbooks, content-specific pedagogy, to name a few. However, there is another tool that often goes unmentioned, yet it provides the foundation needed to use the aforementioned tools in the best way possible. That tool is a teacher's professional judgment. In numerous ways it is the most important and least understood aspect of a teacher's job.

> Professional judgment is central to grading students' various performances, but professional judgment often seems "mysterious" or unpredictable, not only to the laity but even to professionals themselves. Because it is based on interpretation of a highly complex object – such as a piece of writing…professional judgment can appear to be unreliable. (Speck, 1998, p. 17)

It is the mystery surrounding the decision-making and professional judgment of teachers as they grade student writing that led the researcher to conduct the present study

The purpose of this qualitative study was to investigate the decisions teachers made while grading samples of decontextualized student writing in English language arts. Five research questions were created to focus the study on teacher decision-making and were thoroughly considered during the review of the literature. The research questions were as follows:

1. In what ways and to what extent do teachers use classical categorization in their grading decision-making process when grading samples of student writing in English language arts?

2. In what ways and to what extent do teachers use prototype categorization in their grading decision-making process when grading samples of student writing in English language arts?
3. In what ways and to what extent do teachers use exemplar categorization in their grading decision-making process when grading samples of student writing in English language arts?
4. To what extent, if any, do similarities or differences exist in teachers' decision-making processes when grading samples of student writing in English language arts?
5. To what extent, if any, do teachers' decision-making processes differ when grading student writing samples of high and low performance levels in English language arts?

Two theoretical frameworks formed the basis of the research questions. Categorization theory provides the basis for the first three research questions by investigating whether teacher decisions while grading writing are rooted in classical, prototype, or exemplar categorization. The final two research questions use grounded theory to guide an investigation in the similarities and differences in teacher decision-making processes.

The literature utilized in this review represents relevant literature from 1888 to the present day. The study of teacher grades and teacher decision-making has spanned more than 131 years. The literature was chosen specifically to represent each of the relevant time-periods and eras of research into teacher decision-making in grading student writing. Although many sources of the variability in grades have been found, many questions remain unanswered (Brookhart et al., 2016). The pervasiveness of grading variability and the apparent lack of

definitive answers served as the basis for this researcher's decision to specifically search for and include research throughout the last 131 years.

A systematic literature search was conducted using the resources at the University of Central Florida to find peer reviewed studies and articles, as well as books written in English that included any year of publication. The search procedures consisted of a keyword search of electronic databases and backwards and forwards snowball searches using the citations and reference list to identify additional sources. The databases used were: Education Full Text, ERIC – EBSCOhost, ProQuest, PsycInfo, Sage Journals, Dissertation and Thesis Full Text, Taylor and Francis, Sage Premier, and JSTOR. The keywords used to search the databases were: assessing writing, grading writing, grading essays, writing rubrics, teacher decision-making, teaching writing, inter-rater reliability, intra-rater reliability, inter-rater agreement, assessment literacy, categorization theory. Online journals consisted of *Assessing Writing, Journal of Educational Research, Educational Researcher, Practical Assessment, Research & Evaluation, Educational Measurement: Issues & Practice, Educational Assessment,* and *Assessment in Education: Principles, Policy and Practice.* Additionally, the internet was used to access websites for the Massachusetts Department of Elementary and Secondary Education and the National Association of Educational Progress.

The review of the literature consists of six sections. The first section provides a historical context to writing assessment dating back to the beginning in 1888. The second section focuses on the overall issues surrounding writing assessment including holistic rating and assessment literacy. The nature of rater agreement, including inter-rater and intra-rater agreement, are discussed in the third section. The fourth section investigates the causes of inter-rater

disagreement. The fifth section provides insight into teacher decision-making research including the initiation of modern methods of research such as the think-aloud method. A variation of this method called a cognitive laboratory interview that was utilized in this study. The final section contains a discussion of assessment as an instructional tool which tied into the central question of the study. That question is, if teachers were unable to agree upon whether a student had met the standards for writing, then how were teachers able to give the appropriate feedback to improve student writing?

Historical Context of Grading

Grading has been a topic of study for over 100 years. The first study of teacher grades by Edgeworth (1888) was conducted to investigate the varying degree to which teachers graded student work. Although the 5-point A-F scale or the 13-point scale with additional +/- included were, at the time of the present study, considered commonplace, this was not the case in the 19th and early 20th centuries. In the 19th century, teachers presented student progress orally to parents during a home visit which later became written reports (Guskey & Bailey, 2001). High schools preferred to report grades in percentages because they often believed that written reports were too time consuming (Farr, 2000). Brookhart et al. (2016) argued the change from written reports in favor of percentages eliminated any communication of the academic progress of students in both skills and knowledge.

It was not until after the turn of the century that a move away from percentages to a standardized A-F measurement scale was proposed by Starch (1915). This letter grade scale gave meaning to the percentages by placing them in a range of ten percentage points per letter

grade. By the 1940s, the A-F scale was used by 80% of schools across the country (Brookhart et al., 2016). The A-F scale developed in the 1940s is now commonplace and a part of the culture of the modern American high school.

In 1921, Campbell studied teacher marks and decried a lack of thorough study at the time, which despite many decades of research since then, has remained an area in which more understanding is needed (McMillan, 2001). Campbell further set the context for additional research by arguing a student's grade

> …simply registers relative standing with respect to other pupils in the class. It can be said to give, at most, a general diagnosis of the pupils' relative condition; it certainly does not furnish a prescription for the teacher to follow. (Campbell, 1921, p. 510)

A century ago, it was clear the primary limitation of grades was only to describe the past and not to prescribe anything that needs to occur in the classroom setting in the future.

Although grades may be arbitrary, they need to be clear in their meaning so that future employers will understand the work candidates may bring to their workplace (Campbell, 1921). This sentiment was echoed in 2016 by Brookhart et al. who further stated that grades are a better predictor of post-secondary success than standardized testing because they are multi-dimensional measures that are comprised of achievement and the classroom effort needed to achieve those results. Moreover, high school grades measure effort and achievement over time; a standardized test measures a student's best effort during a few hours (Atkinson & Geiser, 2009). A key distinction here is that standardized testing only measures academic achievement at a single point in time but high school grades include other factors such as effort and endurance. Therefore, the purposes of these two measurements are different as well.

The historical correlation between student grades and achievement test scores for the last 100 years has been shown to be about .5. If it is accepted that the achievement score measures academic achievement, then 25% of teacher grades are also measuring academic achievement. That means the other 75% is measuring something else (Bowers, 2011). Bowers further explained that a .5 correlation is neither weak nor strong, indicating teacher grades are neither a measure of pure academic merit nor do they demonstrate a lack of any academic merit.

Writing Assessment

In the early years of assessment research, researchers identified a strong link between the effects teacher assessment practices had on student achievement. Hartog and Rhodes (1936) studied classroom assessment and noted the importance it plays not only in the classroom but also in the trajectory of a student's life.

> No element in the structure of our national education occupies at the present moment more public attention than our system of examinations. It guards the gates that lead from elementary education to intermediate and secondary education, from secondary education to the universities, the professions, and many business careers, from the elementary and middle stages of professional education to professional life. (Hartog & Rhodes, 1936, p. 1)

Indeed, the importance of classroom assessment is paramount because of the effects that it has on students' futures. However, with the amount of research that has been conducted over the last 100 years, teaching methods, standards, accommodations, and technology have changed dramatically yet assessment practices have remained nearly constant. There is a hypocritical

quality to striving for innovation in every other way except for innovation in grading (Ferriter, 2015). This lack of innovation in grading has also pervaded writing assessment as well.

Huot (1996) argued that a theory of writing assessment is lacking because the focus has been primarily on developing processes and procedures for assessing student writing. These processes include a collection of systems, rubrics, and beliefs about assessment of student writing, but there is no theoretical basis for any part of this collection. Huot posited that because of this focus on processes, all theoretical considerations have been left behind.

Holistic Rating

The process of assessing student work is a highly individualized one in which even teachers themselves have been seemingly unable to explain or give a rationale for their decision-making (McMillan, 2003). When teachers rate essays, they are not merely assessing the skill of the writer, but rather they are using "observation, interpretation, and perhaps most importantly, the exercise of personal and professional judgment" (Myford & Wolfe, 2003, p. 389). It is this professional judgment that leads to the variability in scores that has been shown to occur in numerous studies in the first half of the 20th century (Healy, 1935; Hulten, 1925; Rugg, 1918; Starch, 1913, 1915; Starch & Elliott, 1912). Yet, professional judgment is at the heart of holistic rating and continues to be an accepted practice in grading student writing.

Holistic rating is scoring the student's overall proficiency in writing using a single scale and is more suitable to large-scale assessment where ranking is the focus rather than detailed feedback (Smith & Dunstan, 1998). The counterpart to holistic writing is analytic scoring which then scores students' writing proficiency on multiple components where each component is

scored independently and detailed feedback is given to the student (Rosen, Ferrara, & Mosharraf, 2016). The modern rubric which is commonplace now is based upon analytic scoring.

Numerous studies have indicated that when teachers are rating essays, they tend to focus too closely on aspects of writing that have a quality of correctness such as grammar, spelling, punctuation, word choice, and mechanics (Stern & Solomon, 2006). This is likely due to the nature of scoring such traits because they are easy to identify and easy to correct. However, some other scholars have noted that focusing mostly on macro-level issues may not be helpful either. Straub (1997) suggests that the primary focus of the assessment should be a blend of micro and macro level issues to include claims, development of ideas, support and evidence, and paragraph style and structure.

Because the assessment of student writing is such an individual process and also lacks an underlying theory of assessment, there have been no agreed upon processes or procedures for assessing student work. An ethnographic study by Kalthoff (2013), in which he observed five high school teachers in Germany grading the work of their own students, highlighted some common findings. Kalhoff (2013) identified individual evaluation and collective evaluation as two separate procedures. After evaluating the individual student, teachers compared each student's score to the collective, adjusting as necessary. He further concluded that the teachers in the study needed to distribute the scores along the entire length of the scale because there were good and bad students. In Kalthoff's findings (2013, pp. 93–96), the following assessment procedures were observed (a) teachers constructed model answers before they began grading; (b) teachers took notes in the margins of the papers but did not necessarily assign points; (c) teachers verbally engaged the student as if the student were in the room; (d) teachers took the student, as a

person, into consideration which may have positively or negatively affected the score; and (e) good students functioned as an "alarm system" with teachers using them to gauge the quality of their assessments. These findings indicate the pervasiveness of the variability in grading procedures that are not easily attributable to one single issue.

Regardless of the procedures or the focus on local vs. global concerns, the common variable in the assessment of student writing has been the rater. Even 70 years ago, it was realized by Guilford (1954) that the human rater is ultimately flawed. Guilford also discussed assumptions about a rater: that the rater is objective, has the content knowledge, and the quantitative observational skill to judge the essay accurately. Brimi (2011) observed, however, that most teachers are not well trained in assessment, and it is because of this lack of training that teachers often neglect the teaching of writing in their classrooms (Dempsey, PytlikZillig, and Bruning (2009). However, Brimi did not suggest an exact type of training needed to bring the essay scores closer in agreement.

Assessment Literacy

In order for teachers to assess student work accurately, they need to be adequately trained in assessment. The lack of pre-service and in-service training for assessment has been well documented (McMillan, 2003). This lack of training has led to an entire teaching corps who lack knowledge in assessment. The phrase *assessment literacy* originated with Stiggins (1991) and it describes one who understands the difference between a high and a low-quality assessment and can apply that knowledge to increase student outcomes. Being literate in assessment is essential because without that form of literacy, a teacher is bound to make incorrect interpretations which

then could lead to misguided decisions within the classroom and negatively affect student achievement (Purpura, 2016). Researchers in writing assessment have defined assessment literacy as "technical know-how, practical skills, theoretical knowledge, and understanding of principles" (Taylor, 2009, p. 27). Taylor further stated that these skills and attributes need to be accompanied by an understanding of the role that assessment has in learning.

The use of rubrics assumes that teachers have the content knowledge and practice to be able to score student samples of writing effectively (Crusan, Plakans, & Gebril, 2016). This assumption has been incorrectly generalized to include new teachers who have not completed a pre-service teacher preparation program (Weigle, 2007). Rubrics alone are not enough of a scoring guide unto themselves. Appropriate training and practice are needed to ensure that the rubrics are used appropriately and consistently.

The American Federation of Teachers [AFT], the National Council on Measurement in Education [NCME], and the National Education Association [NEA] (1990) wrote seven standards for teacher development in the area of assessment. These standards are:

1. Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.
2. Teachers should be skilled in developing assessment methods appropriate for instructional decisions.
3. Teachers should be skilled in administering, scoring, and interpreting the results of both externally produced and teacher produced assessment methods.

4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and improving schools.
5. Teachers should be skilled in developing valid pupil grading procedures which use pupil assessment.
6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.
7. Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information (AFT et al., 1990, para 13).

Despite teachers spending between 25% and 30% of their professional time on assessment activities and the standards being in existence for 30 years, teachers have not been instructed in assessment principles (White, 2009). This lack of assessment training continues to exacerbate the variability in teacher grades for student writing samples.

Rater Agreement

Given the variable nature of the composition of teachers' grades, the natural question arises regarding the reliability of grades between teachers. The reliability question has been a mainstay for most of the 20th century. Finkelstein (1913) was an early identifier of the issue.

> When we consider the practically universal use in all educational institutions of a system of marks, we can but be astonished at the blind faith that has been felt in the reliability of

> the marking systems. School administrators have been using with confidence an absolutely un-calibrated instrument. (p. 1)

Finkelstein's words read as if they were written today which highlights the significance of the continued reliability problem.

It is worth noting the preceding discussion of reliability focused on inter-rater agreement between teachers. Wolfe et al. (2016) defined inter-rater agreement as "the degree to which a rater assigns scores to a particular set of examinee responses that are consistent with scores assigned to those responses by other raters" (p. 2). Several researchers highlighted the lack of agreement among teacher scoring of samples of student work.

Inter-Rater Agreement

Studies in the first half of the 20th century have consistently indicated a lack of inter-rater agreement. Each study in this section is noteworthy because they establish the context for the lack of inter-rater agreement that framed the basis of the present study.

Over 130 years ago, Edgeworth (1888) studied the nature of what is presently referred to as inter-rater agreement. He devised the theory of errors, which described humans as prone to perceive the world around them inaccurately, elaborating further on human inaccuracy, "The observations of the senses are blurred by a fringe of error and margin of uncertainty" (Edgeworth, 1888, p. 600). He distilled the sources of errors to three primary areas: (a) chance, (b) differences among the raters, and (c) the rater's error in judgement becomes the estimation of the examinee's actual proficiency. Edgeworth gave recommendations to mitigate the rater errors when it came to nominating a student for honors. He recommended that students who had

scored just above and just below the threshold for honors should be given to a rating supervisor for a final rating. His reasoning for re-grading scores above and below the threshold was that the rater errors could equally as likely resulted in a higher score as a lower score (Edgeworth, 1888). This recommendation for a third scorer would become commonplace in the second half of the 20th century in large-scale writing assessments for state and national level tests.

Variation in Grading – The Beginning

Although Edgeworth was considered one of the first to study the concept of inter-rater agreement, Starch and Elliot (1912) were among the earliest researchers to conduct empirical studies of the problem. Starch and Elliot analyzed the scores received by 142 teachers who graded two high school English papers. The directions given were to grade paper A and paper B on a 100-point scale according to the standards and practices of the respective school.

Starch and Elliot found the results varied by 34 points for paper A and by 49 points for paper B. These point values on a modern 5-point scale would give the student work a range from A-D for paper A and from A-F for paper B. The severe variability of the results startled the researchers because "teachers usually state, when asked about differences in marking, that the grades of the same paper assigned by different teachers might differ at the most by 10 points" (Starch & Elliot, 1912, p. 454). This study was one of the first to document the imprecise nature of teacher's grades. However, one of the complicating factors was that the score needed to pass also varied from school to school. Some of the participating teachers worked in schools where 70% was considered to be a passing score; others considered 75% to be a passing score. Schools that used 60% and 80% to determine passing scores were discarded from the study due to the

difficult nature of the comparison. However, these findings highlighted the highly variable nature of the grading scales at each school (Starch & Elliot, 1912, p. 449). The varying nature of the measurement tool itself made the study of teacher's grades difficult in the beginning of the 20th century.

The researchers also investigated the relative standings of paper A and paper B. The median score of paper A was 8 points higher than paper B, effectively meaning that paper A was the better paper. Starch and Elliot proposed that if teachers were consistent with themselves, a teacher who gave paper A one of the highest grades would have also given paper B one of the highest grades. They found that not to be the case. Nineteen of the 142 teachers scored the reverse, with paper B judged to be the better paper.

The most concerning of the variations was the number of teachers who scored the same paper either just above the passing score or just below. This, according to the researchers, has serious consequences for the student, because it is not just a variation in scores, but also rather a variation as to whether the student will pass the course. The consequences of Starch and Elliot's findings were significant: "Therefore, it may be easily reasoned that the promotion or retardation of a pupil depends to a considerable extent upon the subjective estimate of his teacher" (Starch & Elliot, 1912, p. 454). Students' success in a course relies heavily upon the ability of the teacher to assess the work accurately. But perhaps more importantly, their success is also determined by the teacher to which they have been assigned.

In another content area, Ashbaugh (1924) found similar variation in grades with 50 pre-service mathematics teachers grading a seventh-grade mathematics test. The range of scores was

even greater than in the study by Starch and Elliot (1912) in that the grades on the tests ranged from 29 to 80 of 100 points possible.

What distinguished Ashbaugh's 1924 study relative to Starch and Elliot's (1912) findings was that after the mathematics test was graded, the group of teachers discussed the best way to grade each mathematics problem. A unique issue in grading mathematics is that one teacher may grade an answer as either completely correct or incorrect. Another teacher may award points for the correct process even if it does not lead to the correct answer. In Ashbaugh's study, the teachers reached a consensus to grade for the correct processes as well as for the correct answers. Therefore, the teachers awarded partial points to an incorrect solution. After a consensus was reached, the papers were graded again and the range of scores narrowed from 51 points to 18 points. Although there was still variation in the grades, the range decreased significantly after agreeing on a common set of grading procedures. Ashbaugh concluded, "An agreement among those who are to do the marking upon the values to be given to certain phases of the work will result in reduced variability" (Ashbaugh, 1924, p. 197). Ashbaugh's study was one of the earliest studies to suggest training, or at least agreeing upon a common set of grading procedures.

Lack of Common Grading Practices and Scales

With the early work of Edgeworth (1888) and the subsequent studies of Starch and Elliot (1912), and Ashbaugh (1924), it was difficult to compare across schools or even across teachers because common grading practices did not exist at the time (Brookhart et al., 2016). Regarding the grading variation in Starch and Elliot's (1912) study, Starch (1913) argued there were several

reasons why the papers were scored with such a high variability in his previous study with Elliot. He reasoned [there are]:

> Four major factors enter into the problem which, I believe, fully account for the situation: (1) Differences among the standards of different schools, (2) Differences among the standards of different teachers, (3) Differences in the relative values placed by different teachers upon various elements in a paper, and (4) Differences due to the pure inability to distinguish between closely allied degrees of merit. (Starch, 1913, p. 630)

In many ways it is this variability that led Starch to his next and perhaps most important contribution. He then proposed a nine-point scale of A+, A-, B+, B-, C+, C-, D+, D-, and failure. It is important to note that the middle of each grade, namely A, B, C, and D were not included in the scale. Only the +/- variants were part of the nine-point scale (Starch, 1913, p. 633). The purpose of this scale was twofold. The first was to bring about a common grading scale for all teachers. The second was to help reduce the variability in the scores of student work. Before this time, many schools and teachers were using a 100-point numerical scale and reporting the score in percentage form. Starch proposed the letter grade scale to group the scores into larger units that would result in less variation. This was an incredibly simple solution because instead of attempting to reduce the teacher's score variability, he suggested changing how the scores were reported so that the variability was less obvious.

Variation in Grading – 100 Years Later

After nearly 100 years of progress toward common grading practices, standards-based instruction, writing instruction, introduction of rubrics, and writing assessment, the lack of inter-

rater agreement remained nearly identical, as shown in a replication study by Brimi (2011). In Brimi's study, one anchor paper was scored by 73 English teachers in the same school district in Tennessee who completed two days of training in a district-wide writing assessment framework, 6+1 Traits of Writing. The traits that were scored were ideas, organization, voice, word choice, sentence fluency, conventions, and presentation (Culham, 1995). Teachers were instructed to grade the paper using the framework and then to assign a score based upon a 100-point scale. The range of scores varied by 47 points, from a score of 50 to a score of 96. Ten teachers gave the paper an A, 18 a B, 30 a C, nine a D, and six an F (Brimi, 2011, p. 6). Similar to the findings of Starch and Elliot (1912), the grades assigned by this group of teachers varied from A to F for the same piece of student work.

In his analysis of the graded papers, Brimi (2011) identified a few trends. He found that even after detailed training on assessment procedures that the teachers within the same district using the same assessment framework: (a) graded the writing differently and (b) a wide range of scores resulted (Brimi, 2011). Furthermore, outside of the scoring data, Brimi found that some teachers showed evidence in their own lack of writing knowledge and their inability to teach and assess beyond the five-paragraph essay. Others showed an unwillingness to follow the new grading framework or change their old assessment methods. Finally, he found some teachers to be "assessment illiterate" (Brimi, 2011, p. 7). The implications of this study were clear. The grade a student received for an assessment relied largely upon which teacher was grading it. Brimi concluded,

> It does indicate that an "A" in one class may not be an "A" in another class or at another school. And the Advanced Placement student who received an 83% …in my class might

> expect anything from an "F" to an "A," depending on who grades the paper and what the grader knows about writing and assessment. (p. 8)

Brimi's statement succinctly describes the problem in which a significant portion of a student's academic success is not dependent upon their own achievement, effort, or the ability of the teacher to teach writing. Rather, the student's success is dependent upon which teacher to which he or she has been assigned.

These three seminal studies of the variability of teacher grades in the early 20th century laid the foundation for the study of lack of inter-rater agreement. The degree of variability was far greater than believed. Starch and Elliot (1912) argued "marks are far less precise than the majority of teachers and pupils believe" (p. 456). What is also evident is that there was no attempt at this point in time to understand why the inter-rater agreement was so poor. The purpose of the studies was to determine if the subject matter made a difference or if 100 years of advances in education had made a difference. It did not make a difference and the one constant among findings in the studies was the great amount of variability in teachers' grades.

Table 1 displays the research studies, method, sample, and findings of the inter-rater agreement studies. Table 1 is as follows:

Table 1

Inter-rater Agreement Studies

Study	Method	Sample	Findings
Ashbaugh (1924)	Descriptive statistics	55 pre-service teachers grading one seventh grade mathematics test three times with interval of four weeks in-between each scoring	Variability of 51 points. Successive grading resulted in reduction of variability to 18 points
Brimi (2011)	Descriptive statistics	73 high school English teachers graded one paper using 6+1 Traits of Writing	Variability of 46 points
Starch & Elliot (1912)	Descriptive statistics	142 high school English teachers graded two papers from two different students using grading procedures from their own school	Variability of 34 points for paper A, 49 points for paper B

Contained within Table 1 are the three seminal studies that form the basis of this researcher's research study. Ashbaugh (1924) and Starch and Elliot (1912) indicate the highly variable nature of teacher's grades when grading the same work. Nearly 100 years later, the results were replicated by Brimi who found identical results indicating the problem of grade variability had not changed and continues to be an issue.

Intra-Rater Agreement

The previous section described studies establishing the concept of inter-rater agreement or consistency in scoring of the same work by different raters (Wolfe, Song, & Jiao, 2016). By extension, intra-rater agreement is the "consistency of grading a given writing by the same rater

twice" over time (Rezaei & Lovorn, 2010, p. 21). While inter-rater agreement has been studied extensively, intra-rater agreement lacks the same amount of investigation. Perhaps the problem of intra-rater agreement is more vexing because it points to a deeper, more problematic issue within teacher grading. Hulten (1925) recognized this problem with intra-rater agreement in his study of high school English teachers grading student writing. The variability between the ratings resulted in 20% of the papers swapping their pass/fail status from the first to the second rating. Hulten's strong words were particularly pointed and indicate the real problem with teacher grades. "Teachers' marks are mere guesses, some good, some poor, some indifferent. Since they are mere guesses, they are not sufficiently reliable to be used for promotion purposes" (Hulten, 1925, p. 54). Even though describing teacher's grades as "mere guesses" could be considered shocking especially written nearly 100 years ago, without standard procedures and a theory of writing assessment, Hulten's words are accurate.

As previously discussed, Ashbaugh's study (1924) included a look into the intra-rater agreement by giving teachers the same mathematics tests to grade two more times with several weeks in between grading sessions. Although the variability of scores was reduced after the teachers agreed on a common scoring procedure, a lack of inter and intra-rater agreement existed. Rugg (1918) noted a similar conclusion regarding teachers' agreement with themselves. He stated, "…as one examines the grades given by an individual teacher to the same piece of student work graded at two different times there is 'distinct evidence of unreliability of marking'" (p. 703). The unreliability of grading with themselves added to Hulten's assertion that grades are mere guesses rather than a solid assessment of student work.

One of the first large scale intra-rater agreement studies was conducted by Eells (1930) in which 61 teachers were given a set of written responses to questions about elementary school level geography and history. The teachers were asked to grade the responses and then grade the same responses 11 weeks later. Eells found a great amount of variance from the first grading to the second grading by the same teacher. Eells, like Hulten (1925), strongly worded his conclusion when he determined the intra-rater agreement was just above "sheer guesses" and that "the fallibility of human judgment, even when it is the same human judging the same material, is strikingly demonstrated" (Eells, 1930, p. 52). When raters are not agreeing with themselves in successive ratings of the same material, the conclusion can only be that there is no conceptual or practical basis to the raters' assessment.

Causes of Inter-Rater Disagreement

By the middle of the 20th century, the majority of the research regarding the assessment of student writing and more specifically the inter-rater agreement of writing assessment has focused on the degree to which the lack of agreement exists. Few researchers investigated why or how the lack of agreement existed. The following sections explore the causes of inter-rater disagreement, including rater effects and teacher assessment decision-making research.

Rater's Comments

Several studies in the first half of the 20th century laid the groundwork for the lack of inter-rater agreement in the grading of student work. Subsequently, several researchers began studying the decisions that go into making those judgments. Diederich, French, and Carlton

(1961) are considered to have begun the modern era of writing assessment research. As stated in their abstract, the purpose of their study was to "…serve as a stepping stone toward closer agreement among judges of student writing…by revealing common causes of disagreement [among those judges]" (p. 1). In their study, 300 samples of college freshman writing representing two different writing prompts were given to a group of university professors representing several different departments as well as lawyers, business leaders, and newspaper editors. This study was unusual in that individuals from professions other than English teachers and English professors were included as raters. It is unclear in the study as to why that choice was made by Diederich et al. (1961), and this decision alone may have jeopardized the reliability of the study. However, the results of the study were similar to the findings in many other studies.

The raters were given no specific directions other than to assess the works based upon what they liked and what they did not like, write comments on each paper, and then sort the papers into one of nine piles representing grades ranking the papers from best to worst. One specific direction was that each of the nine ranking scores had to be used and that no less than six papers could have any one score.

The results showed that 94% of the papers received seven or more different grades and no paper received less than five different grades. The researchers went further and examined the written comments using factorial analysis. The most significant finding of Diederich et al. (1961) related to the types of raters. Five "types" of raters were identified based on the comments they wrote in response to the directions they were given: to write anything they liked or disliked about each paper. The five types, which emerged based on the concepts used in

raters' written comments, were: (a) ideas, (b) mechanics, (c) analysis and organization, (d) style, interest, and sincerity, and (e) choice and arrangement of words (Diederich et al., 1961, pp. 51–55). This analysis formed the beginning of the first rubric. Diederich et al. (1961, "Abstract") arranged the rater categories into the following five schools of thought:

- Ideas: relevance, clarity, quantity, development, persuasiveness;
- Form: organization and analysis;
- Flavor: style, interest, sincerity;
- Mechanics: specific errors in grammar, punctuation, etc.;
- Wording: choice and arrangement of words

Rubrics that remain largely in use today rely heavily on at least a few of these five schools of thought.

Diederich et al.'s 1961 study represented the first investigation into what factors led teachers to make their assessment decisions while grading writing. In 1974, Diederich further developed the schools of thought into what has been regarded as the birth of the analytical rubric. A 5-point rating scale comprised of two categories organized using eight criteria: (a) the category of general merit with criteria established for ideas, organization, wording, and flavor; and (b) the category of mechanics with criteria established for usage, punctuation, spelling, and handwriting (Diederich, 1974, pp. 53–58). Although the five schools of thought from their previous study have been rearranged slightly, the general basis of the analytic rubric remains the same.

The true significance of the Diederich et al. (1961) study is that it was the first investigation into what raters were thinking while they were grading papers. However, there were some weaknesses in the study. Relying solely on comments written on the essay is

problematic because those comments did not represent all the thinking of the raters. Rather, comments represent what raters were willing to write. In addition, it was unclear if the comments were written as quasi-feedback to the imagined student or if the feedback was provided to justify the score for the study. Furthermore, it was also difficult to analyze the number of comments a rater had written because there were more opportunities to comment on mechanical errors than there were opportunities to comment on ideas. To conclude, after having observed more comments on mechanics (than on ideas) that raters were more concerned with mechanics was a difficult conclusion to make.

The Rubric

With the advent of the rubric, more and more educators are using a rubric to assess writing. However, the usefulness of rubrics as an instructional tool and validity of assessments using a rubric has been studied with conflicting results (Broad, 2003). On one hand, rubrics focus the rater in such a way that they can concentrate more on the substance and argument of the essay rather than on the micro-level topics of mechanics and style (Rezaei & Lovorn, 2010). Rezaei and Lovorn further stated that an additional benefit to rubrics is that they allow for more consistent feedback from the teacher to the student. Other researchers have found that even with a rubric, raters apply it inconsistently when assessing the same piece of student writing (Hunter & Docherty, 2011). Therefore, a rubric is not the complete solution to the variability in teacher's grades.

The commonality in the variation in scores is the rater. Raters rely upon collective academic and non-academic experience and training to assess the student work. As human

beings, the raters are subject to bias that influences their rater decisions (Spool, 1978). These biases simply do not allow a rater to be "neutral and objective recorders of some physical reality" (Hill, O'Grady, & Price, 1988, p. 346). In the next section, biases known as rater effects will be discussed.

Rater Effects

One of the most researched causes of inter-rater disagreement is the area of rater effects. Rater effects are defined as a "broad category of effects [that result in a] systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the ratee" (Scullen, Mount, & Goff, 2000, p. 957). According to Myford and Wolfe (2003), numerous rater effects exist. They include logical error, contrast error, influences of rater biases, beliefs, attitudes, and personality characteristics, influences of rater/ratee background characteristics, proximity error, and order effects (Myford & Wolfe, 2003). More specifically, rater and ratee background characteristics including students' first names (Garwood, 1976), gender and attractiveness (Krawczyk, 2018), handwriting (Huot, 1988), and the order of grading have been shown to influence the rater's grading (Wade, 1978). Myford & Wolfe (2003, p. 393) identified four rater effects that have been researched extensively: (a) leniency/severity effect, (b) central tendency effect, (c) restriction of range effect, and (d) halo effect. These four effects are discussed in the following paragraphs.

The rater effect known as leniency or severity is one in which the rater tends to rate the assessments higher than the average (leniency) or lower than the average (severity) of all the scores (Saal, Downey, & Lahey, 1980). There is a noticeable pattern that a rater consistently

scores too high or scores too low compared to other raters. Cronbach (1990) considered severity and leniency to be the most serious of errors and even after extensive training, though the degree of the effect may be lessened, the raters' tendencies may continue (Weigle, 1998). This proves to be another example in which training may not assist in the consistency of teacher's grades.

The central-tendency effect and the restriction of range effect are closely related. The central-tendency effect in many respects is the opposite of the severity/leniency effect in that the rater tends to score assessments around the midpoint and avoids the extreme ends of the rating scale (Myford & Wolfe, 2003). This rater does not make extreme distinctions about the assessment and, therefore, each of the scores is tightly wound around the median. Although the scores are close to the median, extreme scores on the high and low end may also be present but to a much lesser degree. Myford and Wolfe observed that this effect is harmful because the lack of discrimination in performance can hamper decisions about the assessment. Another way to view this effect is that when the assessments are graded, the result of the ratings is that "everybody is average" (Cascio, 1982, p. 393). Since the raters are reluctant to score at the extreme ends of the scale, crowding the scores toward the median makes the group of students all appear to be average.

The restriction of range effect involves the clustering of scores around a central point, but that point is not necessarily the midpoint. The point could be above or below the median (Saal et al., 1980). The natural question that arises if the scores are clustered above the median is how this restriction of range differs from the leniency effect? In the leniency effect, the scores tend to be above the mean, yet the full range of scores is used. In the restriction of range effect, all of the scores tend to focus around one of three points, below the mean, around the mean, or above

the mean (Saal et al., 1980). In this case, the full range of scores is not used and as the title of the effect suggests, the scores are restricted around one of the three points.

The halo effect occurs when the rater tends to rate each of the components of the writing with the same score as the holistic score regardless of the merit of the individual components (Saal et al., 1980). For example, if the overall piece of writing is successful and scores well, then the individual components, such as mechanics, word choice, etc., will also score well even if one of those individual components is relatively poor. Wells (1907), who has been considered the first to identify this effect (Saal et al., 1980), indicated the rater allows the overall impression of the piece of writing to influence the rater's impression of the individual components of the writing (Wells, 1907). Therefore, the individual components create a halo that carry over to other components and to the work as a whole.

The key factor of each of the rater effects is that the ratings on the assessments vary not because of the merits of the writing, but because of who the rater is and his or her tendencies when rating student assessments. Guilford (1954) noted there is an assumption regarding the abilities of a rater to be objective and serve as "a good instrument of quantitative observation" (Guilford, 1954, p. 278). However, the rater is a human being who is subject to error and biases. These biases became evident in a study by Sakyi (2000) who found that raters bring their own expectations to the rating process. Sakyi expressed the belief that raters do not necessarily focus on the rubric. Rather, it appears they have a personal dialogue with the piece of writing as they are rating it.

Teacher Decision-Making Research

The study of writing assessment in the first half of the 20th century focused on investigating the assessment outcomes quantitatively or investigating the assessment procedures themselves (Cooksey et al., 2007). This line of inquiry focused solely on objective procedures but not the people who were enacting the procedures. The inter-rater disagreement has had underlying causes, but the investigation into these causes of disagreement was lacking and has not been thoroughly studied (Branthwaite, Trueman, & Berrisford, 1981). Later in the 20th century, researchers began studying the underlying causes of inter-rater disagreement using varying techniques.

Beginning in the late 20th century and the start of the 21st century, the writing assessment studies began to use different research techniques to determine the causes of inter-rater disagreement. Studies by Cooksey et al. (2007), Huot (1993), Pula and Huot (1993), and Wyatt-Smith and Castleton (2005) were different in a few important aspects. The first difference is that researchers began to study grading of writing within the context of the teacher's classroom. In-context writing is that written by teachers' students within the context of the current classroom environment (Cooksey et al., 2007). Teachers assign the writing prompt as part of the curriculum and grade assignments using their usual grading procedures. Prior to these in-context writing studies, the focus was on out-of-context writing (Starch & Elliot, 1912; Ashbaugh, 1924; and Brimi, 2011) in which teachers were given stand-alone writing samples that were not linked to their classroom or curriculum.

The second and perhaps most important aspect differentiating these studies was the use of think-aloud methods. A think-aloud method allows for participants to verbalize their thoughts

(Leighton, 2017), and the method also will "permit the identification of additional unanticipated cue information, as well as facilitating the unpacking of the interior dynamics of teachers' thinking on a text-by-text basis" (Cooksey et al., 2007, p. 429). Think-alouds have since proven to be an effective method for studying teacher decision-making behavior (Wiseman, 2012). Previous studies have been conducted to investigate how the raters (as persons) affected the way in which they scored the assessment.

In the studies described in the following section, the investigation moves beyond rater attributes, focusing on the thought processes of teachers. Teacher decision-making was compared in grading in-context and out-of-context writing (Cooksey et al., 2007; Wyatt-Smith & Castelton, 2005). Two other studies which focused on how teacher thought processes differed for experienced vs. novice raters are also described (Huot, 1993; Pula & Huot, 1993). These studies will be discussed in the following sections.

Think-Aloud Method

Vaughn (1991) was one of the earliest studies to delve into teacher thought processes and decision-making by using a think-aloud method. Nine raters graded six essays in a university writing course placement context in which their verbalizations were recorded and transcribed. The most common verbalizations were "content unclear or weak" and "handwriting hard to read" (Vaughn, 1991, p. 114). It is of special interest that many researchers have concluded that handwriting not only detracts from the assessing writing process but also influences the outcome in positive and negative ways (Huot, 1988; Johnson, Penny, Gordon, Shumate, & Fisher, 2005;

Wiseman, 2012). Therefore, to reduce that rater effect in studies, it would be advantageous to ensure that student responses are typed vs. handwritten.

Vaughn (1991) found that raters had their own reading style (p. 118). Styles included a *single-focus* style in which raters would search for specific aspects that would make the essay a passing or failing essay. Another reading style was a *two-category* strategy in which the raters focused on content and mechanics and weighed the writer's competence in both. A third reading style was described by Vaughn as the *laughing rater* who interacted with the imaginary student rather than the writing itself. One rater commented, "Well, I just don't like this student at all" (Vaughn, 1991, p. 120). Vaughn's conclusion was that each of the nine raters focused on different aspects of the essay, despite the training they received. Regardless of training or a rubric, raters decided for themselves what was important and what was not when grading a student essay. Vaughn summarized with "each rater comes to rely on his own method" (Vaughan, 1991, p. 121). Regardless of the rubric or the training, teachers develop their own method of grading through their experience.

In-Context and Out-of-Context Writing

In an Australian study, Cooksey et al. (2007) studied the decision-making process of 20 fifth-grade teachers assessing 25 in-context student essays supplied by the teachers and 25 out-of-context essays provided by the researchers. The teachers evaluated the 50 samples of student work using two assessment frameworks. The first framework was a five-point rating scale with a "1" indicating poor performance and a "5" indicating excellent performance. This rating scale matched the scales the teachers used in their own classrooms on a regular basis. The second

framework was a three-point benchmark rating scale used in Australian standardized testing. A think-aloud protocol was employed to capture the teachers' thoughts in real-time as they were grading the essays. The researchers determined the teachers did not use the same cues to make their judgments and that their scores often varied across four of the five available scores. Only in the classification of the very best or very worst of samples were the teachers able to achieve reliable scoring patterns. Furthermore, they determined more research was needed to explore teachers' judgments of out-of-context essays (Cooksey et al., 2007). The out-of-context essays represented a standardized assessment context in which the rater does not have background information in which to make additional judgments about the student work. Therefore, the cues teachers use within the essays become even more critical when determining a grade.

In another Australian study, Wyatt-Smith and Castleton (2005) studied two teachers who team-taught their fifth-grade class. A think-aloud protocol was used to capture the decisions both teachers made together as they graded 25 in-context essays the teachers provided and 25 out-of-context essays the researchers provided. The researchers found the think-aloud data to be striking. For the in-context essays, the teachers focused on local attributes rather than standards or curriculum. The local attributes included how early it was in the school year, whether the student came from a low or high socio-economic status family, or what the teachers expected of fifth grade students after their own years of teaching experience. Their years of experience were a resource for their judgment. For these two teachers "an externally-defined, stable standard was not a relevant point of reference; they chose instead to develop a site specific, locally-relevant standard…" (Wyatt-Smith & Castleton, 2005, p. 138). The standard to which the teachers were

grading the student changed based upon the various attributes of the individual student. This suggested that these teachers were grading the student, not the essay.

For the out-of-context essays, the teachers had a difficult time scoring the essays because they did not have any of the locally derived attributes from which to draw. During the grading, the teachers spent much time discussing what the classroom context was, the student's background, and even wondering how long the students were given to complete the writing. What was also notable was that the teachers spent considerable time attempting to guess the gender of the student. This indicated the heavy reliance on knowing the student when grading in-context essays. Moreover, this was evidence of earlier research regarding rater effects in which the characteristics of the student, in this case the student's gender, could have been an influence upon the teacher if the gender were known (Krawczyk, 2018). Knowledge of the student and the context of the instruction was a significant factor in teachers' decision-making in grading student work.

In contrast to the in-context essays, the out-of-context essays were graded with more of a reliance on the relevant writing standards for fifth grade in Australia. The researchers concluded that for these teachers, they graded their students more on their individual progress over time rather than comparing their achievement to a standard (Wyatt-Smith & Castleton, 2005). It is interesting to note that whether the student met the standard was not nearly as important to the teacher as the student's overall positive progress in writing.

Experienced vs. Inexperienced Raters

Furthering the study of raters' cognitive processes was Huot (1993) who identified differences in the procedures experienced raters rated student writing vs. the procedures novice raters rated student writing. Huot selected four raters with holistic scoring training and experience, and four raters considered to be novice raters. A total of 24 college level essays were rated by each rater utilizing a think-aloud method. The findings were similar to Vaughn (1991) in that they indicated several differences in the processes the raters used in scoring the essays. The novice raters were more likely to start and stop during the rating process of one essay to comment or discuss the portion they had just read. This made the rating session more inefficient and distracted novice raters. Furthermore, the novices interacted with the rubric while rating the essay and focused more on the writing quality (Huot, 1993). Novice raters appeared to access assessment supports and refer to them often while assessing student work.

The experienced raters, on the other hand, were more likely to read the essays without interruption, determining their rating based on a general impression at the end. Instead of interacting with a rubric, Huot (1993) contended that the experienced raters were not interacting with the text as much as they were interacting with imaginary students in their classes. Vaughn (1991) found similar results with experienced raters. Additionally, the experts appeared to have a rating style rather than a rubric. This style included reading rapidly to gain a quick impression, guarding against biases, being fair, and conversing with the text directly (Huot, 1993). This rating style was developed by the expert raters after many years of experience in grading numerous samples of student writing.

Pula and Huot (1993) conducted a validation study of Huot (1993) in which novice and experienced raters rated 21 essays and then participated in a focused interview with retrospective probes after the ratings were completed. Personal and professional reading and writing experience tended to be the most influential aspects in rating the student writing. Other influences were teaching experience and holistic scoring experience. Pula and Huot expressed the belief that holistic scoring experience developed over reading thousands of essays which created an internal rubric or the "real rubric" (Pula & Huot, 1993, p. 252). Indeed, according to Smith (1993), the extensive reading and grading gave a rater an internal compass to find the good qualities of writing, and it was this experience, which was more powerful than any rubric or training provided.

Table 2 displays the research studies, method, sample, and findings of the teacher decision-making studies. Table 2 is as follows:

Table 2

Teacher Decision-making Studies

Study	Method	Sample	Findings
Cooksey et al. (2007)	Idiographic-Statistical and Think-Aloud	20 Australian fifth-grade teachers graded 25 in-context papers and 25 out-of-context papers	Teacher assessment is an exercise in teacher judgment
Huot (1993)	Protocol Analysis	24 college level essays rated by 4 raters utilizing a think-aloud method	Novice and experienced raters grade writing in vastly different ways
Pula & Huot (1993)	Protocol Analysis	21 essays rated by novice and experienced raters	Personal and professional reading and writing experience were the most influential factors
Wyatt-Smith & Castelton (2005)	Think-aloud case study	Two Australian fifth-grade teachers grading 25 in-context papers and 25 out-of-context papers	Highlights the different judgments made between in-context and out-of-context student work
Vaughn (1991)	Think-aloud study	Nine university professors scored six essays using 6-point rating system	Raters have different reading styles and focus on different essay elements

Table 2 shows the major studies involving teacher decision-making. These studies investigated the decisions teachers made while grading student writing. The investigative technique of protocol analysis and think alouds revealed that novice and experienced raters grade in different ways (Huot, 1993), writing experience were the most influential factors (Pula & Huot, 1993), in-context and out-of-context grading produced different judgments (Wyatt-Smith & Castleton, 2005), and raters focus on different essay elements when grading student writing (Vaughn,

1991). The significance of these studies was in the techniques used to investigate the causes of inter-rater by observing the raters while rating student writing.

Assessment as an Instructional Tool

The research discussed previously has shown there was a lack of accuracy and reliability in the assessment of student writing in English language arts. The purpose for striving for inter-rater agreement was so that the assessment was fair (Huot, 2002) and not based upon who was grading the assessment. The assessment should have been based upon the qualities of the writing, yet research results continually indicated that many other factors influenced the raters' decision-making. Huot also stated that for the assessment to be fair, it not only needed to be consistent (inter-rater agreement) but it also needed to include the basis for the decision (Huot, 1990). If the basis for the decision is not included, then the rating provides little use for the student.

The basis for the assessment can shared with students as feedback and can also serve as a formative assessment. An assessment is formative when the results are used to inform future instruction (Cizek, 2010). Students would be better served and learning would be accelerated if the process of assessment and feedback for learning became a focus of the classroom (Wylie & Lyon, 2015). Furthermore, it is important to note that assessment has an instructional purpose and that, along with the proper feedback, it can be more than an administrative task (Nagin, 2003). Nagin believed that it was not enough simply to give students more practice at writing. A systematic focus on instruction and feedback needs to take place in order to increase student achievement.

The essential question is: How can teachers give appropriate feedback for student growth in writing if they are grading student writing with such high variability? As Starch and Elliot demonstrated in 1912 and Brimi replicated in 2011, teachers were grading student work and assigning a grade of F at the same time another teacher's assigned grade was an A. In this example, it is important to consider the feedback both teachers might have shared. It is likely that the A grade teacher reflected the essay was excellent with few notes of critical feedback. In contrast, the other teacher, assigning an F grade, would have indicated the student had failed at the task and included critical feedback indicating what needed to be improved, providing a dilemma for the student. The increased focus on standards-based achievement cannot be realized if the teachers were unable to recognize accurately if the student had met the standard.

Summary of the Literature Review

In modern societies, writing is an essential form of expression and is considered the foundation of communication and literacy (Behizadeh & Engelhard, 2011). Writing is taught in every grade level, and there has been much teacher preparation focus on the process of teaching writing. Though writing assessment is essential to a comprehensive program of writing instruction, it has not received the same amount of attention (Myers et al., 2016). Without quality writing assessment, the growth of the students' skills and abilities will be hampered (White, 2009). Therefore, further study in teacher assessment of student writing is needed.

The review of the literature for this study established a foundation for further study regarding teacher decision-making while grading student writing in English language arts. More specifically, there is a need to investigate the real rubric (Pula & Huot, 1993) teachers used while

grading student work. Most importantly, it is essential for researchers to understand the cognitive processes teachers used while grading student work and the biases that affected their ratings (Cronbach, 1955). It is imperative that these cognitive processes are studied in more detail in the future.

Three major eras of writing assessment studies have been discussed in this literature review: (a) inter-rater and intra-rater agreement, (b) rater effects, and (c) teacher decision-making. The first era of inter-rater and intra-rater agreement was studied thoroughly to establish a consistent lack of consistency in grading student work in a variety of contexts (Ashbaugh, 1924; Brimi, 2011; Eells, 1930; Starch & Elliot, 1912). These studies focused only on the degree of the scoring inconsistency, not on how or why the discrepant scores existed.

The second era began in the mid-20th century when Diederich et al. (1961) studied why the inter-rater disagreement occurred by analyzing the written comments of raters while grading student writing. This study led to what has been considered the first assessment rubric. However, further inter-rater studies have shown that a well-constructed rubric does not seem to make any difference. Teachers tend to ignore the rubric and use their own internal rubric (Pula, & Huot, 1993). What that internal rubric contains has not been thoroughly studied.

Furthering the work of Diederich et al. (1961) were the concepts of rater effects. The background characteristics of the rater and of the ratee influenced the scores the raters gave for the assessment of writing (Scullen, Mount, & Goff, 2000). These effects also laid the groundwork for the many factors that influenced the score a student receives on an assessment, regardless of the rubric. Hodges et al. (2019) determined that the issue is much more complex

than previously understood and that simply requiring more rubric training may not lessen the amount of inter-rater disagreement.

The third era of writing assessment study involved investigating the teacher decision-making processes. Several studies were conducted to investigate the reasons why discrepant scores exist, using think-aloud methods to reveal the thoughts and decisions teachers made while grading samples of student writing (Cooksey et al., 2007; Huot, 1993; Pula and Huot,1993; Wyatt-Smith & Castleton, 2005). These studies began to uncover teacher thought processes but also focused on the different ways in which in-context and out-of-context writing was assessed and whether experienced raters grade differently than novice raters.

Throughout the study of writing assessment, the studies focused on the following areas: (a) inter-rater agreement, (b) writing for placement in college level composition courses, (c) rubrics, (d) rater effects, and (e) teacher decision-making. These studies were primarily concerned with the degree of the inter-rater disagreement and how to lessen the disagreement. Furthermore, the studies for why the disagreement exists have centered solely on who raters are as people and the biases they bring to the assessment process. At the time of the present study, there was a lack of sustained effort to study the teacher judgement process regarding assessment (Wyatt-Smith, 1999). Pula and Huot (1993) observed that even when teachers use a rubric to aid in their decision-making, they tend to rely on their own "real rubric" (p. 249) based upon their own reading, writing, and assessing experience. Uncovering this internal rubric is an essential next step in the study of writing assessment.

CHAPTER 3
METHODOLOGY

Introduction

The purpose of this qualitative study was to investigate the decisions teachers made while grading samples of decontextualized student writing in English language arts using a cognitive laboratory interview method. In order to understand teachers' grading decisions better, this study focused on the following research questions.

1. In what ways and to what extent do teachers use classical categorization in their grading decision-making process when grading samples of student writing in English language arts?
2. In what ways and to what extent do teachers use prototype categorization in their grading decision-making process when grading samples of student writing in English language arts?
3. In what ways and to what extent do teachers use exemplar categorization in their grading decision-making process when grading samples of student writing in English language arts?
4. To what extent, if any, do similarities or differences exist in teachers' decision-making processes when grading samples of student writing in English language arts?
5. To what extent, if any, do teacher decision-making processes differ when grading student writing samples of high and low performance levels in English language arts?

The methodology utilized in this study is presented in this chapter and has been organized into seven sections: (a) research design, (b) selection of participants, (c) instrumentation, (d) data collection, (e) data analysis, (f) validation and credibility, and (g) summary.

Research Design

A qualitative research design employing a cognitive laboratory interview was utilized for this study. When teachers are grading student writing, the decision-making processes needed are rather complex. In order to study a complex issue, the methodology is also complex, involving the researcher's stepping into the environment and interacting directly with the participant to see the world from the participant's perspective (Corbin & Strauss, 2008). Furthermore, a qualitative research design allows for the intimate discovery of variables rather than the distant study of them. By interacting directly with the participants, the participant's environment can be studied at a significant level of detail.

> This detail can only be established by talking directly with people, going to their homes or places of work, allowing them to tell the stories unencumbered by what we expect to find or what we have read in the literature. (Creswell, 2007, p. 40)

It is this level of detail in interacting directly with the participants that served as the basis of this researcher's decision to choose a qualitative methodology.

Categorization Theory

The method of inquiry for this study was two-fold. The first method was to use categorization theory (CT) as a basis for a priori coding. CT is "a cognitive procedure of sorting

things into conceptual boxes" (Haswell, 2001, p. 57). There are three types of categorization in CT: (a) classical, (b) prototype, and (c) exemplar (Haswell, 1998). Haswell defined the three types as follows:

- Classical Categorization. Sorts objects into rigid, clearly defined categories based on rules.
- Exemplar Categorization. Compares the object to a recent memory of an example or examples in the category.
- Prototype Categorization. Classifies objects based upon how similar they are to a mental image of a prototype of that group (Haswell, 2001).

Grounded Theory

The second portion of inquiry for this study was grounded theory. Grounded theory is a qualitative research design that derives from an explanation of a process based upon the data (Corbin & Strauss, 2008). Moreover, grounded theory serves as a method of analysis for the qualitative data. In the present study, once qualitative data were gathered, they were analyzed using open coding to form the basis of the explanation. Qualitative coding is an iterative process in which themes, categories, and codes are developed in an inductive manner through a constant comparative method (Glaser, 1992). The basis for Research Questions 4 and 5 was grounded theory.

The choice to use an observational research methodology was deliberate. Hook and Rosenshine (1979) surveyed teacher beliefs regarding their grading procedures for student writing in English language arts, observing the same teachers while grading student writing. The

results indicated that self-reported beliefs and actions were incongruent at best and in some cases were nearly inverses of each other. Therefore, the researchers concluded that self-reported beliefs could not be accurately used to represent actual behaviors. It was this finding that led the researcher to choose an observational methodology for the present study to gather qualitative data in the classroom setting.

Cognitive Laboratory Interview

The qualitative methodology for this study was a cognitive laboratory interview. A cognitive laboratory interview is one in which the participant completes a task while thinking aloud in order to generate verbal reports (Leighton, 2017). The transcription of the verbal reports allows the researcher to study a cognitive process (Ruiz-Primo, 2014). Although a cognitive laboratory interview is similar to another research method called the think-aloud method, the two differ in purpose and in procedure (Leighton, 2017). The purpose of a think-aloud is to study a participant while solving a problem. The problem could be a mathematical problem or some other type of problem that requires analysis, steps to derive the solution, and the solution itself. In procedure, the researcher remains silent while the participant thinks aloud, uninterrupted during the problem solving process.

Unlike think-alouds, the purpose of a cognitive laboratory interview is to study a participant's comprehension of written text and to analyze the process of completing the task (Ruiz-Primo, 2014). More specifically, a cognitive laboratory interview is used to "study the manner in which target audiences understand, mentally process, and respond to the materials we present" (Willis, 2005, p. 1). Willis further stated that the cognitive laboratory interview aimed

to study the processes of (a) comprehension, (b) recall, (c) decision and judgment, and (d) response (Willis, 2005, p. 4). The task in the present study was to analyze the decisions teachers made while grading student samples of writing in English language arts. In the case of grading student writing, the teacher must comprehend the student sample, recall exemplars, rubrics, or whatever information the teacher used to grade the sample. Teachers must use their judgement to make decisions about the samples and create a response in the form of a grade or feedback for their students.

The procedures for a cognitive laboratory interview differ from a think-aloud in one aspect, which is the use of interview probes. An interview probe is a statement used by the researcher during the interview to elicit an explanation, clarification, or elaboration of a cognitive process (Willis, 2015). Interview probes can be concurrent, i.e., as the interview is occurring, or retrospective, i.e., after the interview has concluded (Ruiz-Primo, 2014). In contrast, a think-aloud interview leaves the participant to verbalize thought processes unfettered by researcher participation.

Willis (2015) identified seven primary interview probes in a cognitive laboratory interview. They are as follows:

1. Meaning-oriented probe to help clarify interpretation of specific terms in an item or task.
2. Paraphrase-type probe to uncover genuine understanding.
3. Process-oriented probe to uncover the rationale underlying a given response.
4. Evaluative-type probe to explore participants' assessment of the item or task of interest.

5. Elaborative-type probe to uncover the rationale or reasons underlying a given response.
6. Hypothetical-type probes to explore possible responses that the participant has not provided to a given item or task.
7. Recall-type probes to determine the parameters for a given response. (p. 37)

These interview probes are the primary distinction between a think-aloud interview and a cognitive laboratory interview. In order to uncover the teacher's grading process, it was important to probe their verbalizations further if they were unclear. For example, if teachers annotated the essay using their own annotation scheme, it was important to use a meaning-type probe to understand the meaning of the annotation. Furthermore, if teachers determined that the student essay should receive a score of 85 of 100 points, it was necessary to use a process-type probe to understand the process the teacher used to determine that score.

Each of the interview probes were utilized during the cognitive laboratory interviews. The probes used most frequently were meaning-oriented probes, elaborative-type probes, and process-oriented probes. The probes used the least in the interviews were hypothetical-type probes because they did not provide any value to the interview.

Selection of Participants

The participants for this study were selected using a purposive, criterion sampling technique (Miles & Huberman, 1994). The criteria for selection were 9th and 10th grade English language arts teachers who taught during the 2019–2020 school year in a large urban public school district in the Southeast United States. Teachers of all genders, ethnicities, and

experience levels were included as participants because the English language arts faculties included a wide range of demographics. It was important to ensure that the diversity of the sample represented the diversity of the population.

Both of these grade levels were chosen because the graduation requirements for the large urban public school district included passing an English language arts examination in 10th grade which included a significant writing portion. With this graduation requirement, there was a common focus on writing instruction and assessment which was present within the state standards, curriculum, and practices of 9th and 10th grade English language arts.

The sample size for this study was 21 participants. Creswell (2007) recommended a sample size of 20–30 participants and Mason (2010) found in his study that the most common sample size for qualitative dissertations was 20–30 participants with the average being 28 participants (Mason, 2010). However, Mason was concerned at the lack of empirical evidence to support the use of such a sample size not only in dissertations but among experts as well. Although these are recommended guidelines, the important consideration was "whether there is sufficient representation in the sample (and, by extension, the data) to inform and support the investigator's conclusions" (Leighton, 2017, p. 78). Thus, the volume of data needed to be sufficient enough to justify the conclusions by which the data collection continued until saturation occurred (Charmaz, 2014). In the case of this research study, the data being collected by the 21st interview did not lead to new themes thereby indicating data saturation.

Instrumentation

In this study, two primary instruments were utilized. The first instrument was the English language arts writing prompt and the second instrument was the researcher. Both instruments are discussed in the next two sections.

Writing Prompt

The first instrument was comprised of the 2018 10th grade English language arts student writing samples from the Massachusetts Comprehensive Assessment System (MCAS) which were obtained from the Massachusetts Department of Elementary and Secondary Education (DESE). The Massachusetts DESE was selected because Massachusetts consistently led the nation in reading at the eighth and 12th grade levels and in writing at the eighth-grade level in the National Assessment of Educational Progress [NAEP] (www.nationsreportcard.gov, n.d.). NAEP has not tested 12th grade writing. Furthermore, MCAS writing prompts were chosen because it was not likely that teachers in the large urban public school district in the Southeast would have encountered either the prompts or the sample student responses.

The Massachusetts DESE provided a scoring rubric (Appendix A) and a scoring guide. The scoring guide included previous years' writing prompts with genuine handwritten sample student responses and the accompanying rationale for the score the student response earned in that testing year (www.doe.mass.edu, n.d.).

Sample student responses scored at a medium score (Appendices B and C) and a high score (Appendices D and E) on the MCAS rubric were selected to provide a variety of student work that elicited a sufficient volume and quality of responses from the participants in the study.

Student writing samples with the highest and lowest scores possible were not selected to avoid any issues with the writing being too perfect or too problematic to generate high quality data from the participants.

The student essays provided by the Massachusetts DESE were digitally scanned versions of the original hand-written responses. Student handwriting has been found to have a significant impact upon the grade the student writing received (Huot, 1988; Johnson et al., 2005). Namely, a positive correlation exists between neat handwriting and a high score. To reduce this effect, the sample student responses used in this study were transcribed verbatim by the researcher in typewritten form.

The 2018 MCAS writing prompt was as follows:

> Often in works of literature, a character is influenced by another person or factor. From a work of literature you have read in or out of school, select a character who is influenced by **one** of the persons or factors listed in the box below.
>
> - a friend
> - a family member
> - a spiritual belief
> - society
>
> In a well-developed composition, identify the character, describe how the character is influenced by the person or factor, and explain how the character's experience is important to the work as a whole. (Massachusetts Department of Elementary and Secondary Education, 2018)

The 2018 MCAS writing assessment (Massachusetts Department of Elementary and Secondary Education, 2018) was scored using a rubric in two domains: (a) topic/idea

development and (b) standard English conventions. One sample student response (Appendix B) was selected from the 2018 MCAS testing administration with a score of 3 on the 4-point scale for standard English conventions because it represented student writing in which "errors do not interfere with communication and/or few errors relative to the length of the essay or complexity of sentence structure, grammar and usage, and mechanics" (Appendix C). The second student writing sample (Appendix D) was selected with a score of 5 on the 6-point scale rubric for topic/idea development because it represented "Full topic/idea development, logical organization, strong details, [and] appropriate use of language."

Researcher as Instrument

The second instrument used in this study was the researcher. "Because the researcher is the instrument in semi-structured or unstructured qualitative interviews, unique researcher attributes have the potential to influence the collection of empirical materials" (Pezalla, Pettigrew & Miller-Day, 2012, p. 2). At the time of this study, the researcher was a 47-year-old male educator with 23 years of experience at the secondary level. He had taught high school mathematics, served as a mathematics instructional coach, assistant principal, and a principal of a large urban public high school for eight years. His experience as a principal led him to identify an arbitrary nature to grades teachers assigned to individual assignments and to the overall course grade. This belief was developed while discussing grading philosophy with individual teachers, providing grading professional development and managing conflict in numerous parent conferences regarding grades. Throughout the eight years of his principalship, a curiosity was developed as to the decision-making process teachers used to assign grades to student work.

This curiosity was then sharpened by the research of Starch and Elliot (1912) and later replicated by Brimi (2011) in which teachers graded the same piece of student writing with the scores varying by at least 37 points. The researcher in the present study wondered how experienced classroom English language arts teachers could read the same student writing and make such varying determinations as to its quality. It is this wondering that led him to pursue this study.

Data Collection

Pilot Study

A pilot study was conducted with four teachers who were interviewed in their classrooms. Pilot studies are a smaller version of a full-scale study, as well as the specific pre-testing of a particular research instrument such as a questionnaire or interview protocol (van Teijlingen & Hundley, 2001). The purpose of this pilot study was to test the cognitive laboratory interview protocol, receive feedback from the pilot study participants, and make any adjustments necessary to ensure the success of the study.

The cognitive laboratory interviews were audio recorded and transcribed by an online audio transcription service to ensure the correct usage of the recording device and the accuracy of the transcripts. During the pilot interviews, attention was paid to the types of interview probes, the participant responses, body language, and non-verbal cues. Through the pilot study, feedback was given regarding the order of the writing samples and the instructions within the interview protocols. The most significant feedback was regarding the desire of the participants to annotate the text. The participants expressed that their grading practices always included annotating the text either by hand on paper or using digital word processing programs. Thus, the

protocols were amended to include protocol instructions allowing for the annotation of the text and the verbalization of the annotations so that they would be captured by the audio recording device. After the changes to the protocols were made, a fourth and final pilot study was conducted to test the changes. No more revisions were needed at that point.

Procedures

The University of Central Florida required approval by the Institutional Review Board (IRB) before the collection of data could begin. Approval to conduct this research study was granted on June 18, 2019 (Appendix F). The large urban public school district in the Southeastern United States required approval to conduct research at one of its comprehensive high school campuses. Approval to conduct research at three campuses was received on July 30, 2019 (Appendix G).

The principals of each campus were contacted and provided the university and school district notices of approval to conduct research. The principals were asked to identify teachers who taught 9th or 10th grade English language arts during the 2019–2020 school year. Each identified teacher was contacted and an appointment was scheduled to meet on campus at their convenience. The cognitive laboratory interviews occurred during the months of September and October of 2019. Each interview was privately conducted inside the teacher's classroom during their planning period or after school. Before an interview began, field notes were taken to describe the classroom environment which included a diagram of the classroom.

Participants were made aware that their participation was completely voluntary and that they could withdraw from the study at any time and for any reason. Participants were also

informed that their identity would be kept confidential and that the data collected during the cognitive laboratory interview would remain secure. They were also informed of the following data security measures: (a) no identifiable data would be collected, (b) voice recording data and transcription data would be maintained for five years in a secured, locked cabinet according to university policy. An informed consent statement was signed by the participant (Appendix K) and field notes were taken indicating the participant understood and agreed to give consent.

The cognitive laboratory protocol (Appendix F) was read to the participant and any questions were answered. The protocol included directions to read the student essays aloud, verbalize all thoughts, and grade the essay as if it were an assignment in the participant's own classroom. The participant gave the essay a score out of 100 points and it was recorded in the field notes. Demographic information was gathered using the demographic information collection form (Appendix J). It was explained to the participant that the interview would consist of one practice essay (Appendix K) as well as the two essays included in the study.

The interview was recorded using a Sony ICDUX560BLK hand-held digital recording device. The device was placed between the researcher and the participant and was set to record after a brief recording and sound level check. The interview began by giving a paper copy of the first essay to the participant to read and score. Participants were encouraged to make annotations on the essays while also verbalizing their annotations for the purposes of audio recording. During the interview, the researcher used interview probes "designed to have the participant elaborate, explain, and/or clarify his or her response" (Leighton, 2017, p. 82). The purpose of the interview probes was to encourage the participant to verbalize concurrently rather than retrospectively. Willis (2015) identified seven distinct interview probes: (a) clarifying probes,

(b) paraphrase probes, (c) process probes, (d) evaluative probes, (e) elaborative probes, (f) hypothetical probes, and (g) recall probes. At the conclusion of the interview, the digital recorder was stopped and the participant was thanked. The researcher spent time after each interview completing field notes to include any non-verbal cues, contexts, behaviors, or situational factors that were not captured on the recording device.

Data Analysis

Qualitative analysis was performed on the data collected during the 21 cognitive laboratory interviews. The interviews were recorded on a digital recording device, transcribed verbatim by a professional audio transcription service, and uploaded to a qualitative data analysis tool, ATLAS.ti.

The 21 transcribed interviews were coded using a process whereby data are reduced into segments. The name given the segment is the code (Creswell, 2007). The first round of coding identified topics teachers verbalized in their cognitive laboratory interviews. After the initial codes had been identified, the codes were converted from topics to gerunds. Glaser (1978) indicated that gerunds, rather than topics, allow the researcher to identify processes. Furthermore, utilizing codes written as gerunds encourages the researcher to gain a strong sense of action that allows analysis from the perspective of the participant (Charmaz, 2014). Throughout the coding process, codes were combined into axial codes and emergent themes to make comparisons utilizing a constant comparative method. The constant comparative method is a method of analysis that makes comparisons through inductive processes, resulting in themes

and concepts (Charmaz, 2014). After the initial rounds of coding concluded, axial coding was utilized to identify the initial codes that were central to the emergent themes (Charmaz, 2014).

A priori coding was utilized stemming from the theoretical framework, categorization theory (CT). Haswell (2001) defined categorization theory as "the basic cognitive procedure of sorting things into conceptual boxes" (p. 57) and applied CT to the study of writing assessment. He further stated the act of assessing writing was in fact categorization because giving the student writing a score of 5 or "doesn't meet proficiency" is an act of placing the essay into a category (Haswell, 1998). Haswell (1998) defined CT as having three types of categorization: (a) classical, (b) prototype, and (c) exemplar. Classical categorization represents the ideal of the rubric. The student writing fits exactly into one of the categories of the rubric, and the rater follows the rubric with fidelity. Prototype categorization is matching the current stimulus to an idealized version of the category and measuring how closely it fits that category. Lastly, exemplar categorization in writing assessment is using exemplars from a rater's memory to compare to the current writing sample. The rater accesses an extensive personal catalogue of exemplars to compare to the current sample of student writing being assessed. Even though rubrics represent classical categorization, researchers have found that teachers most closely follow prototype and exemplar categorization when assessing writing (Haswell, 1998). This is due to the rarity that any one sample of student writing fitting exactly within the confines of a rating category within a rubric.

An additional round of coding consisted of open coding stemming from grounded theory. Grounded theory is a method of analysis that intends to generate or develop a theory as a result of the research (Strauss & Corbin, 1990). The process of developing the open codes included (a)

reading through the audio transcripts several times, (b) developing a list of about 10 to 12 codes, and (c) reviewing the codes and the transcripts and reducing the categories down to five or six (Creswell, 2007). Table 3 contains the research questions and coding scheme for each research question. Table 3 is as follows:

Table 3

Research Questions, Data Source and Coding Type

Research Question	Coding Scheme
1. In what ways and to what extent do teachers use classical categorization in their grading decision-making process when grading samples of student writing in English language arts?	A priori coding
2. In what ways and to what extent do teachers use prototype categorization in their grading decision-making process when grading samples of student writing in English language arts?	A priori coding
3. In what ways and to what extent do teachers use exemplar categorization in their grading decision-making process when grading samples of student writing in English language arts?	A priori coding
4. To what extent, if any, do similarities or differences exist in teachers' decision-making processes when grading samples of student writing in English language arts?	Open coding
5. To what extent, if any, do teachers' decision-making processes differ when grading student writing samples of high and low performance levels in English language arts?	Open coding

The first three research questions were based upon categorization theory and utilized a priori coding. The final two research questions were based upon grounded theory and utilized initial open coding and focused coding.

Validation and Credibility

Validation in qualitative research is an issue that needs special attention and methodical processes and procedures (Patton, 1999). Creswell (2007) indicated that validation is an attempt to determine the accuracy of the research findings using validation strategies as a process rather than a mere verification of the results. He further stated, "Any report of the research is a representation by the author" (p. 207). Indeed, qualitative research relies heavily on the researcher as an instrument, and it is important to include procedures to identify and account for the subjectivity of the researcher. Hammersley (1992, p. 69) stated, "An account is valid or true if it represents accurately those features of the phenomena that it is intended to describe, explain or theorise." In order to establish credibility, the validation strategies recommended by Lincoln and Guba (1985) are discussed in the following sections.

Peer Review

Peer review is a technique in which the researcher meets with a disinterested peer who serves as an external check on the research process (Creswell, 2007). The peer acts as a "devil's advocate" whose role is to "be sure that the investigator is as fully aware of his or her posture and process as possible" (Lincoln & Guba, 1985, p. 308). Lincoln and Guba further stated that peer review is essential because it provides researchers a means to clear their minds of thoughts and feelings that may be clouding their minds and hampering the judgement process.

A colleague in the same doctoral program as the researcher served as the peer in the peer review. The data collection process, analysis procedures and conclusions were discussed thoroughly with the peer to ensure fidelity in the qualitative research process.

Negative Case Analysis

Negative case analysis requires searching for and discussing instances in the data which appear to contradict or conflict with the patterns emerging from the data (Creswell, 2007). To account for these negative cases, it was important to revise the working hypotheses throughout the data analysis.

Rich, Thick Description

Rich, thick description is a method of writing that describes in detail the setting, participants, and behaviors that allow readers to draw conclusions as to the transferability to other situations (Creswell, 2007). The detail provided in the descriptions gives the context necessary so that people outside the study can make meaning of the behavior of the participants.

Respondent Validation

Respondent validation is a method in which the transcript and themes are returned to the participants to determine if they accurately represent their understanding of the topic under study (Long & Johnson, 2000). Furthermore, the purpose is not to take the raw data back to the participant or to validate the accuracy of the data. Rather, the purpose is to determine if the findings and interpretations are accurate from the perspective of the participant.

Summary of Chapter 3

This chapter presented the qualitative research design employing a cognitive laboratory interview. A purposive, criterion sample of 9th and 10th grade teachers was chosen to grade two samples of 10th grade English language arts writing. The participants verbalized their thoughts while grading the samples as they were digitally recorded. The recordings were professionally transcribed and coded. A priori and open coding were employed to identify emergent themes. Validation strategies included peer review, negative case analysis, rich, thick description, and respondent validation. Results of the data analysis are presented in the following chapter.

CHAPTER 4
ANALYSIS OF THE DATA

Introduction

The purpose of this qualitative study was to investigate the decisions teachers made while grading samples of decontextualized student writing in English language arts using a cognitive laboratory interview. The research questions that guided the study were:

1. In what ways and to what extent do teachers use classical categorization in their grading decision-making process when grading samples of student writing in English language arts?
2. In what ways and to what extent do teachers use prototype categorization in their grading decision-making process when grading samples of student writing in English language arts?
3. In what ways and to what extent do teachers use exemplar categorization in their grading decision-making process when grading samples of student writing in English language arts?
4. To what extent, if any, do similarities or differences exist in teachers' decision-making processes when grading samples of student writing in English language arts?
5. To what extent, if any, do teacher decision-making processes differ when grading student writing samples of high and low performance levels in English language arts?

A cognitive laboratory interview was utilized to collect qualitative data regarding the decisions teachers made while grading student writing in English language arts. A purposive, criterion sample of 9th and 10th grade English language arts teachers was chosen from a large

urban public school district in the Southeastern United States to grade two samples of 10th grade English language arts writing. While being digitally recorded, the participants read the student writing samples aloud and verbalized their thoughts regarding the writing as the samples were being graded. The participants also annotated the samples and read the annotations aloud to be captured by the digital audio recording device. The recordings were professionally transcribed and coded using a priori and open coding to identify emergent themes. Validation strategies included peer review; negative case analysis; rich, thick description; and respondent validation.

Chapter 4 begins with a description of the relevant demographic characteristics of the study participants, followed by descriptions of the coding process. Next, the initial open coding, focused coding, and thematic coding data are presented. Following the coding data, the presentation of findings has been organized around each of the five research questions which guided the study. The chapter concludes with a discussion of two additional analyses of the grading data.

Participants

A purposive, criterion sample was utilized to identify classroom teachers who participated in the study. Participants were full-time teachers in a large urban public school district in the Southeastern United States who taught at least one 9th grade or 10th grade English language arts class at the time of the study. A sample size of 21 teachers was chosen because data saturation (Charmaz, 2014) was achieved with the 21st participant. Data collected in the 20th and 21st cognitive laboratory interviews did not bring to light any new information that had not already been discussed by previous participants nor did they reveal any new insights.

Furthermore, the final interviews did not lead to any new themes, and the volume of data collected was sufficient enough to justify the conclusions of the study.

Participant's demographic data were collected to determine the characteristics of the teachers participating in the study. The demographic data included: gender, years of teaching experience including the year of the study, and course schedule. Table 4 presents the demographic data.

Table 4

Demographics of 21 Participants

Participant	Gender	Experience	Course Schedule	
			Course 1	Course 2
T1	F	15	ELA 10 Honors	ELA 10
T2	F	24	ELA 10 Honors	ELA 10
T3	F	11	ELA 10 Honors	ELA 10
T4	F	21	ELA 9 Honors	ELA 9
T5	F	19	ELA 9	
T6	M	3	ELA 10 Honors	
T7	M	10	ELA 10	
T8	F	16	ELA 9 Honors	ELA 9
T9	F	3	ELA 9 Honors	
T10	M	2	ELA 10 Honors	ELA 10
T11	F	4	ELA 10 Honors	
T12	F	3	ELA 9 Honors	ELA 9
O1	F	9	ELA 9 Honors	ELA 9
O2	M	16	ELA 10 Honors	AP Literature
O3	F	12	Debate	ELA 9
O4	M	1	ELA 9 Honors	ELA 9
O5	F	20	ELA 10 Honors	AP Literature
P1	F	14	ELA 10 Honors	ELA 10
P3	F	5	ELA 10 Honors	ELA 10
P4	M	22	ELA 10 Honors	AP Literature
A1	F	3	ELA 10 Honors	ELA 10

Note. Participants taught multiple sections of one or two courses in the 2019–2020 school year.

The first three columns in Table 4 represent the participant code, gender, and years of experience. The years of experience included the 2019–2020 school year so that a new teacher indicated one year of experience. The final two columns represent the teaching schedule of the English language arts teachers. Teachers taught multiple sections of one or more English Language Arts courses during the 2019–2020 school year.

The years of experience ranged from a new teacher (Participant O4) to a veteran teacher with 24 years of experience (Participant T2). Furthermore, 17 of the 21 teachers who taught more than one course, taught the same grade level (participants T1, T2, T3, T4, T5, T6, T7, T8, T9, T10, T11, T12, O1, O4, P1, P3, A1). The remaining four teachers taught two unrelated courses such as debate and ELA 9 (Participant O3) or ELA 10 Honors and AP Literature (Participant O5).

Grading Data

During the cognitive laboratory interviews, participants were asked to score each essay based on a possible 100 points. Instructions were given to grade the essays as if the participants gave the assignment themselves in the fourth marking period, using whatever grading policies and procedures they typically used in their classrooms (Appendix I). Table 5 presents the numerical scores and letter grades for Essays 1 and 2:

Table 5

Scores and Letter Grades for Essays 1 and 2

	Essay 1		Essay 2	
Participant	Score	Grade	Score	Grade
T1	70	C	85	B
T2	73	C	92	A
T3	65	D	80	B
T4	60	D	95	A
T5	50	F	80	B
T6	90	A	100	A
T7	80	B	95	A
T8	65	D	88	B
T9	63	D	85	B
T10	70	C	90	A
T11	70	C	80	B
T12	50	F	80	B
O1	50	F	90	A
O2	65	D	88	B
O3	70	C	90	A
O4	75	C	90	A
O5	70	C	94	A
P1	70	C	95	A
P3	85	B	98	A
P4	75	C	88	B
A1	75	C	75	C

Note. Participants scored essays based on 100 possible points using the grading procedures of their choice.

Table 5 shows the numerical scores for Essay 1 which varied from a low score of 50 points to a high score of 90 points, representing a range of 40 points. The letter grades for Essay 1 represented all five letter grades including one A, two Bs, 10 Cs, five Ds, and three Fs. The numerical scores for Essay 2 varied from a low score of 75 points to a high score of 100 points, representing a range of 25 points. The letter grades for Essay 2 included 11 As, nine Bs, one C, zero Ds, and zero Fs.

Coding

Qualitative analysis was performed on the 21 transcribed cognitive laboratory interviews using the ATLAS.ti qualitative data analysis software program and a constant comparative method. Data collection and coding were conducted simultaneously, with each process informing the other. After each interview concluded, in vivo coding was conducted utilizing grounded theory to identify topics that emerged from the grading of two samples of student writing in English language arts. As more interviews were conducted, new topics emerged and further rounds of coding were performed on the previous interview transcripts utilizing the additional codes. After the initial group of codes had been identified, the topics were converted to gerunds to identify processes. Then, codes were grouped into focused codes and emergent themes. Finally, a priori coding was utilized stemming from the theoretical framework, categorization theory. The three a priori codes related to categorization theory were, (a) *Categorizing-Classical*, (b) *Categorizing-Prototype*, and (c) *Categorizing-Exemplar*. In vivo coding combined with a priori coding provided the basis for the qualitative analysis for this study.

The initial rounds of open coding identified topics mentioned by the participants in the cognitive laboratory interviews. Through the constant comparative method, a total of 35 initial topic codes were developed. Examples of the initial codes included the following: (a) *rubric*, (b) *claim/thesis*, (c) *comma*, (d) *exemplar*, (e) *prototype*, (f) *grading*, and (g) *prompt*. To better identify grading processes, the topics were converted to gerunds. Converting topics to gerunds revealed an important grading procedure; Participants were interrupting their reading of the text to consider certain grading decisions, ask questions of the imaginary student, ask questions of

themselves, or wonder aloud about certain text features. This was in contrast to the process of assessing the work as a whole at the conclusion of reading the text. Therefore, it was necessary to subdivide several codes into two codes for "interrupting" and "assessing" which also necessitated an additional round of coding to delineate between the two. Table 6 presents examples of initial codes converted into gerunds.

Table 6

Codes Converted from Topics to Gerunds

Initial Code	Gerund
Prompt	Assessing-Addressing the prompt
Prompt	Referring back to the prompt
Grammar	Interrupting-Grammar
Claim/Thesis	Assessing-Claim/Thesis
Claim/Thesis	Interrupting-Claim/Thesis
Level of Student	Considering the level of student

The column on the left in Table 6 represents the initial topic codes. Through the constant comparative method, it became clear to this researcher that the *prompt* code was too general and did not accurately describe what participants were doing during the grading process. There were two types of participant actions involving the essay prompt which were, (a) the teacher assessing if the student had addressed the prompt in the writing sample and (b) the teacher referring back to the prompt to clarify the purpose of the writing task. Therefore, the initial code was converted

from *prompt* to two codes, (a) *Assessing-Addressing the prompt* and (b) *Referring back to the prompt*.

Furthermore, the code *Claim/Thesis* did not adequately represent participants' grading processes. The first process participants exhibited was interrupting their reading aloud of the text to ask questions of themselves or the imaginary student regarding the direction and completeness of the claim/thesis. The second process was assessing the claim/thesis after the reading of the text had been completed. Thus, two codes in gerund form were developed to represent the processes of the participant interrupting their reading to consider the direction of the claim/thesis and a second code to represent the participant assessing the overall claim/thesis after reading the entire writing sample. These codes were *Assessing-Claim/Thesis* and *Interrupting-Claim/Thesis*.

Once the topics were converted into gerunds, a final list of 29 initial codes was utilized throughout the coding process for a total of 1,606 observations of the recurring phenomena. Four codes were discarded because they lacked prevalence throughout the interviews and also did not represent an important grading process. Table 7 presents the 29 codes used in initial open and a priori coding:

Table 7

Coding

Code	Description	Example	Participant	*f*
Assessing-Addressing the prompt	Participant assesses if or how well the student addresses the prompt.	"I think the biggest problem in this essay…um, that this student only marginally addresses the prompt."	T4	81
Assessing-Claim/Thesis	Participant assessed the quality of the claim or thesis.	"It does show the power of Allie's influence. It would be better if he ended with the point of Allie's influence rather than just restating that Allie has influenced him."	O5	39
Assessing-Conventions	Participant assessed the quality of the conventions.	"On the conventions of standard English, this student, uh, does have some problems, but consistently…it's got a decent readability. I'm probably not gonna take off any points whatsoever on conventions of standard English."	T4	43
Assessing-Evidence/ Elaboration	Participant assessed the quality of the evidence and/or elaboration.	"So it's really lacking in elaboration and I think he's also kind of making an assumption that his reader read *The Giver* and there's no assumptions made here."	T3	115
Assessing-Sophistication	Participant assessed the sophistication of the student's writing	"Got choppy sentences here. It's not very sophisticated writing."	O5	94

Code	Description	Example	Participant	*f*
Categorizing-Classical	Participant's decision-making exhibited classical categorization.	"…do you have a topic sentence? Do you have evidence? Do you explain your evidence to tie back to your topic sentence? Um, do you have a concluding sentence? Do you have transition words? So it would definitely be a very detailed rubric that I could check off as I went."	T5	25
Categorizing-Exemplar	Participant's decision-making exhibited exemplar categorization.	"Now this one I would say probably was written by a high schooler, oh, it could be an advanced middle schooler."	T2	25
Categorizing-Prototype	Participant's decision-making exhibited prototype categorization.	"So I would say the structure would be like a three out of four because I think that again, they could have done more with their thesis regarding the second part of the prompt. Um, the elaboration is, is a true struggle. So I would say, um, maybe [inaudible] 20%, so are two out of four. And then I would say overall his spelling and grammar was good. He just had the one issue that he needs to work on."	T3	56
Considering level of student	Participant considered the level of the student.	"Um, that was a good conclusion. Um, I don't, I think this is a student who seems pretty advanced in their writing."	T10	24
Deciding the grade	Participant explained their grading decision-making for the writing as a whole.	"But other than that, I mean I would grade this novel, I mean this essay literally between, I would say, yeah, that 90 or 95."	T2	97

Code	Description	Example	Participant	*f*
Decision making-Binary	Participant did not consider the quality of the student's thesis, rather determined the student either had a thesis or did not have a thesis.	"Okay. So now that I've read the whole thing, I would go back and look at the prompt, um, and I'll see what it's asking for. So it's asking you in a well developed composition identify a character. He did that describe by the characters influenced by the person or factor. Okay. It was by society. Got that. How the characters experience is important to the work as a whole. So it looks like I can check off all those boxes."	T6	28
Discussing rubric	Participant discussed a rubric to grade the student writing.	"I would use a rubric. Um, like I'm like, because this is such an AP prompt, I'm, I'm instinctively reading it like an AP prompt and, and applying the AP rubric and on the AP rubric, um, this would be insufficient."	O2	39
Evaluating-Conventions	Participant evaluated the student's use of conventions of standard English.	"He does have somewhat complex sentence structure, so kudos to him or her. Um, for the most part I didn't see lots of comma splices."	T3	123
Evaluating-Thesis Binary	Participant evaluated the student's thesis for the existence of a thesis. Also considered if there were evidence and elaboration present. No consideration was given to the quality of the thesis, evidence, or elaboration.	"I do believe that the student did a good job identifying a character and I think that they conveyed their ideas about how the character was influenced by society. I think the ideas are here."	T11	54

Code	Description	Example	Participant	*f*
Evaluating-Thesis Quality	Participant evaluated the quality and completeness of the thesis and the quality of the evidence and elaboration to support the thesis.	"I'm glad you're going here and I wish this was where you really put your focus on instead of the pain and loss element. Again, that's not wrong, but I think it would be a better, more cohesive essay if this was your focus throughout. It took a while for you to finally get here and I wish it was more consistently developed throughout."	O2	146
Interrupting-Addressing the prompt	Participant interrupted the reading of the essay to consider if the student addressed the prompt adequately.	"If I go back to the prompt, um, I'm not feeling like the prompt is really being addressed right now. Um, all right."	T5	18
Interrupting-Claim/Thesis	Participant interrupted the reading of the essay to consider the student's claim or thesis.	"Um, this essay does a good job of taking these details and, and explaining the significance of them. So giving the, the meaning behind them, how does this relate to what they're trying to say? So they're saying, um, Holden hates change, resist. Growing up they talk about where do the ducks go? They're using dialogue, they're elaborating on that."	A1	173
Interrupting-Comma	Participant interrupted the reading of the essay to consider the student's usage of commas.	"Okay. I would tell them to put in the auxiliary comma because it makes that sentence a lot clearer to me."	T3	35
Interrupting-Grammar	Participant interrupted the reading of the essay to consider the student's usage of grammar.	"Okay, so other than some grammar errors, that paragraph is not bad."	T12	65

Code	Description	Example	Participant	*f*
Interrupting-Mechanics	Participant interrupted the reading of the essay to consider the student's usage of mechanics.	"This is still awkward. This construction."	O3	78
Interrupting-Spelling	Participant interrupted the reading of the essay to consider the student's usage of spelling.	"…missing an 'e'."	T6	40
Interrupting-Structure	Participant interrupted the reading of the essay to consider the structure of the student's essay.	"Um, so maybe I would say, um, separate his paragraph into two paragraphs, one about hating change and the other about growing up."	O3	64
Interrupting-Summary	Participant interrupted the reading of the essay to consider if the student is over-utilizing the summarization of the plot in the essay.	"I am so confused by this entire paragraph. It's all summary about how he fell in love with this girl and not actually about the impact that the, it's not tying into the topic sentence 'society influenced him to not be able to love.' So it's not proving that."	O2	22
Interrupting-Transitions	Participant interrupted the reading of the essay to consider the student's usage of transitions.	"So we've got a good transition. Moving onto the next paragraph and now we have the focus in the right direction."	P4	40
Interrupting-Visual Structure	Participant interrupted the reading of the essay to look at the essay visually and made grading decisions from that impression.	"Oh, it dismays me to see that the entire first page is one paragraph. Are there any paragraph breaks? Yes. Okay. There are paragraph breaks, but the entire first page is one paragraph. Uh, generally my expectation is that one paragraph pages are generally not a good idea."	O5	4

Code	Description	Example	Participant	*f*
Interrupting-Vocabulary	Participant interrupted the reading of the essay to consider the student's usage of vocabulary.	"The domain vocabulary is excellent."	T10	18
Referring back to prompt	Participant referred back to the prompt to remind himself of the content of the prompt.	"Okay, the prompt is to um, explain how a character's influenced by another, um, in the story of that friend family member or by a spiritual belief our society."	T8	23
Reluctance to evaluate	Participant exhibited a reluctance to evaluate the essay.	"Okay. Well, honestly, I feel like I would have a hard time grading this without a proper rubric and I feel like it would be unfair to the kids to just assign a grade without a rubric."	T3	13
Subtracting points	Participant considered subtracting points for conventions.	"So my biggest issue that I would, uh, take points off for in this, in this, in this piece would be, um, some of it like the grammar errors, um, the way that they wrote some of these sentences."	T6	24

In Table 7, the left column lists the initial code in gerund form. The next two columns contain the description of the code and an example spoken by a participant during the cognitive laboratory interview. For example, the code *Interrupting-Addressing the prompt* was defined as the participant interrupted the reading of the essay to consider if the student addressed the prompt adequately. *Interrupting-Addressing the prompt* code identified the statement spoken by Participant T5, "If I go back to the prompt, um, I'm not feeling like the prompt is really being addressed right now."

The next step in the coding process was to identify patterns among the 29 initial codes. Several themes began to emerge from the codes in gerund form. Interruptions fell into two distinct areas: (a) interruptions to consider the conventions of standard English and (b) interruptions to consider the claim/thesis. An additional theme was the assessment of conventions of standard English and the assessment of the thesis after the participant had finished reading the essay. The third major theme that emerged from the data was the binary decision-making regarding the claim/thesis.

Table 8 presents the coding data combined to form themes.

Table 8

Emergent Themes

Code	*f*	Thematic Code	*f*
Assessing-Addressing the prompt	81	Assessing-Thesis	235
Assessing-Claim/Thesis	39		
Assessing-Evidence/Elaboration	115		
Assessing-Conventions	43	Assessing-Conventions	137
Assessing-Sophistication	94		
Interrupting-Addressing the prompt	18	Interrupting-Thesis	191
Interrupting-Claim/Thesis	173		
Interrupting-Comma	35	Interrupting-Conventions	366
Interrupting-Grammar	65		
Interrupting-Mechanics	78		
Interrupting-Spelling	40		
Interrupting-Structure	64		
Interrupting-Summary	22		
Interrupting-Transitions	40		
Interrupting-Visual Structure	4		
Interrupting-Vocabulary	18		
Decision making-Binary	28	Decision making-Binary	28
Reluctance to evaluate	13	Reluctance to evaluate	13

The left column in Table 8 represents the codes that were combined to form the thematic code in the right column with the corresponding recurrence of the phenomena. *Assessing-Addressing the prompt*, *Assessing-Claim/Thesis*, and *Assessing-Evidence/Elaboration* were combined to form

the thematic code *Assessing-Thesis*. In contrast, the final two codes, *Decision making-Binary* and *Reluctance to evaluate* stood on their own when converted to a thematic code.

Research Questions

The research questions were developed to gain further understanding of the grading decisions teachers made while grading student writing in English language arts. The research questions were established using the two theoretical frameworks utilized in this study. Categorization theory provided the basis for the first three research questions, and grounded theory provided the basis for the last two research questions.

Categorization theory (CT) is "a cognitive procedure of sorting things into conceptual boxes" (Haswell, 2001, p. 57). The three types of categorization in CT are: (a) classical, (b) prototype, and (c) exemplar (Haswell, 1998). Haswell defined the three types as follows:

- Classical Categorization. Objects are sorted into clearly defined, rigid categories based upon rules.
- Exemplar Categorization. Objects are compared to a recent memory or memories of examples that fit the category.
- Prototype Categorization. Objects are classified based upon how similar they are to the most ideal version of that category (Haswell, 2001).

According to Haswell (2001), rubrics utilized to grade student writing most often resemble classical categorization. However, because the categories in a rubric are so strict and student writing is much more complex than a rubric allows, teachers most often use prototype or exemplar categorization. Therefore, when teachers are grading samples of writing, they either

think of similar writing features, essays, or students and use those memories as a guide (exemplar categorization) or they think of the ideal version of writing features or essays and measure how closely the student samples compare to the ideal (prototype categorization).

Research Question 1: Classical Categorization

The first research question was, in what ways and to what extent do teachers use classical categorization in their grading decision-making process when grading samples of student writing in English language arts? To answer Research Question 1, a priori coding based upon categorization theory was used to code the data. A *categorizing-classical* code was created to identify each instance in which a participant exhibited classical categorization. These decisions manifested themselves in the grading of writing such that participants identified specific features of the essay in a binary way, meaning the student sample either included the feature or did not include the feature. For example, Participant T4 said, "As far as addressing the prompt, this student hit everything that the prompt asks them to hit. I mean one, two, three, first this [evidence/elaboration], second this [evidence/elaboration], the transitions were very clear." This quotation from Participant T4 exemplified classical categorization in that the participant identified that the student addressed the prompt and had a list of evidence and elaboration, including transitions, but did not in turn address the quality of the thesis or how well the student's evidence supported the thesis.

The *categorizing-classical* code was coded with 25 occurrences throughout the 21 cognitive laboratory interviews. The 25 occurrences were concentrated to only eight of the 21 participants. Those participants were as follows: Participant T1, Participant T2, Participant T4,

Participant T5, Participant T6, Participant T7, Participant T8, and Participant T10. The remaining 13 participants did not exhibit any classical categorization decision-making.

Each of the eight participants who used classical categorization discussed the features of the two student essays in a binary, either/or, manner. Participant T1 exhibited classical categorizing when she mentioned, "This is a summary of the novel…, not addressing the prompt." She categorized the student writing as a summary, noting that it did not address the writing prompt which was to give a literary analysis of the novel. A summary is an account of the main points of a story, and Participant T1 determined that Essay 1 fit that definition. Therefore, this statement by Participant T1 was an example of classical categorization. Participant T8 also made a classical categorization when she said, "Um, it doesn't address the prompt in the beginning. So, um, address prompt." Her statement exemplified classical categorization because to address the prompt, the student's essay would have to include specific features, which it did not, in her estimation. Participant T8 further noted about Essay 1,

> And generally I try to focus on one thing like for this one, because I can already see that the student, I mean all the evidence and elaboration doesn't mean anything if you haven't addressed the prompt. So I would just kind of keep focusing on the prompt and almost say, well forget the rest of it right now because if you don't have this, you're already going to score low. (Participant T8)

Participant T8 categorized the essay as one that did not address the prompt. In her decision-making, essays either addressed the prompt or they did not, and this essay was an example of that category to her.

Participant T10 focused on the features as if he had a checklist. "Going back to uh, intro conclusion, they have a strong intro, a pretty decent conclusion…well there's a conclusion as well. So there are a couple, um, transitions. I'll give them that." Participant T10's checklist included whether or not the essay possessed a set of features such as an introduction, conclusion, and transitions. Furthering the concept of a check-off list, Participant T6 even mentioned "checking off the boxes" as if he also had a checklist of items that he identified. Participant T6 said, "He did [identify a character]…Okay. It was [influenced] by society. Got that. How the character's experience is important to the work as a whole. So it looks like I can check off all those boxes." The checklist that both Participants T10 and T6 used was another example of classical categorization in that there are specific features that a category must include.

In each of the classical categorization instances, the participants did not consider the quality of the conventions nor did they consider the quality of the thesis. The fact that the essay included a thesis, evidence, and transitions (as just a few examples) was enough for those participants to be satisfied with their choice of the category. Their decision-making exhibited classical categorization.

Research Question 2: Prototype Categorization

The second research question was, in what ways and to what extent do teachers use prototype categorization in their grading decision-making process when grading samples of student writing in English language arts? To answer Research Question 2, an a priori code *categorizing-prototype* was used to code the participant data. A prototype categorization within the context of grading student writing is one in which the participant thinks of the ideal version,

or prototype, of the essay and compares the student's essay to the ideal. The ideal version is not necessarily a perfect essay because there could be ideal versions at each level of a rubric or at each letter grade. For instance, an ideal version of a B paper may be one with a fully developed thesis with enough of a pattern of grammatical errors so as to distract from the reading of the essay. Therefore, at each letter grade, or at each level of a rubric, there is a prototype in the participant's mind that they are comparing to the student's essay. Moreover, prototype categorization can manifest itself with the participant using a grade level, the time of the school year, or the level of the class (honors vs. regular) as a prototype. The decision the participant makes when using prototype categorization is how close or how far the student's essay is from the prototype, whatever that prototype may be.

Prototype categorization was coded 56 times within the 21 cognitive laboratory interviews. Unlike classical categorization in which those codes were clustered in only eight of the 21 participants, prototype categorization decision-making was found at some level within each of the 21 participants' grading decisions. Table 9 shows the recurring phenomenon of prototype categorization for each participant.

Table 9

Prototype Categorization Coding

Participant	*Categorizing-Prototype* Code *f*
T1	3
T2	5
T3	3
T4	2
T5	3
T6	2
T7	3
T8	4
T9	2
T10	5
T11	2
T12	1
O1	5
O2	4
O3	1
O4	2
O5	3
P1	2
P3	2
P4	1
A1	2

The column on the left in Table 9 represents the participant and the column on the right represents the recurring phenomenon of prototype categorization for participants' grading decisions.

Participants exhibited prototype categorization decision-making when they considered which letter grade to assign to the student's essay. This type of decision exemplifies prototype categorization because there are ideal versions of essays in the participant's mind that represent the letter grades A-F. Participant T8 said, "I would say it's like a high B, high B, low A. I mean

it's not like an A yet. So, I'm going to go high B because I can't give it an A." Her reluctance to give the essay an A was because it did not meet her criteria for an A essay which indicated she had a prototype of what an A essay looked like and the features it contained. An additional example of prototype categorization occurred when Participant T1 observed,

> So, in looking at this because of the simple sentences, because of the errors, there are some typos because of the contractions that the student used. I would say it's almost an 80. So I would probably give this like a 78 or 79 right under, um, the B, so like a high C. (Participant T1)

This example by Participant T1 highlighted the difficulty in placing the student writing on the letter grade scale according to the prototypes she had in her mind for a B and C essay. In her assessment, this essay did not fit the prototype of a B essay and more closely fit the prototype of a C essay.

Participant T10 also displayed prototype categorization while considering where to score the student's essay on the letter grade scale. Participant T10 said, "So I'm going to give this student a 90. It was close to an 80. I think that the evidence and the conventions were borderline…but I decided to give it to them because the essay does hit the marks." Participant T11 discussed, "In general, I would probably give this…a 70% with the opportunity for revision. I think the ideas are there…. But I think for sophistication and conventions and development at the end of the year, it could be stronger." Participant T11 suggested that the prototype for a higher grade changes as the school year progresses. This could be because of increased expectations as writing instruction and assessment progress throughout the year.

Another example of prototype categorization came from Participant T11 who said, "I would be torn between giving this I think a 75 or an 80…[because] it shows a good command of transitional strategies and conventions. But I feel like it's just not there yet with organization and structure." In this statement by Participant T11, she noted that the conventions were close to the ideal essay for a C, yet the organization had qualities that were closer to a B; and she was, therefore, conflicted as to which letter grade better represented the student's work. Furthering this idea of being conflicted between two prototypes, Participant O3 mentioned, "Um, I would probably give this essay…a 95%. But I would say that this is an A essay…um, and that's really the upper edge. If anything, I might go down to a 94, but that's, it's in that range." And finally, Participant P3 offered, "It's confusing to me. Maybe an 85 to be completely honest, an 85 because now that I'm thinking at an 85 or an 80 it kind of borders on that high C." This statement indicated the participant had an ideal version of a B essay and an ideal version of a C essay in her mind and she was having difficulty determining the ideal to which the essay was more closely aligned. Ultimately, she decided that it was closer to an essay that represented a B than a C.

An additional manner in which participants used prototype categorization was to think about what they would expect from a grade level, or from a student in an honors or advanced class. Participant T10 expressed, "Some of them, I would assume that by now the students should understand how to use commas. Therefore I will deduct, um, by the end of this at least 10 points." Participant T5 also compared the student's writing to the prototypical ninth-grade student when she stated, "I'm thinking we in ninth grade teach [writing for standardized testing], so everything would be citing textual evidence with parenthetical explanation with elaboration,

with tying it back together. And I don't feel like this essay totally did that either." Furthermore, Participant T5 also compared the student's writing to the prototype of her expectations for advanced students. She explained,

> I did teach…the higher level up until this year, so that's probably not a good thing either because I'm grading it on the aspect of having [higher level] kids. I haven't seen that much writing from honors yet, so that could be part of it. I might be in for a shock as what I'm thinking right now to be honest, because I had the highest level kids up until this year. Um, and yeah, I would expect more. This could be actual honors. I don't know yet. (Participant T5)

Participant T5 in the preceding two quotations illustrated that there are multiple types of prototypes in which to compare writing including her prototype of the level of student. Finally, Participant O5 stated simply, "This is clearly a pretty strong essay for a 10th grader." Clearly, her expectation for 10th grade student writing functioned as her prototype which served as her basis for assessing the quality of the student writing.

Additionally, one participant used a different type of prototype categorization in which she compared features of the student's writing to a negative prototype of what a student should not do. This was in the form of *canned transitions*. The participant gave modest praise to the student for exceeding the expectations of this prototype when Participant T3 explained, "Like I said, he does have some transitions here in there. So that was good. And for the most part they didn't feel that canned. I would say in conclusion, [they are] somewhat canned but it's okay." This participant's version of a prototype differed from the ones presented earlier but still exemplified the nature of comparing the writing to an ideal, even if the ideal is a negative one.

The participants who exhibited prototype categorization to grade student writing compared the essay to the prototype letter grade, grade level, or writing features students should not replicate. The participants envisioned the prototype in their minds, made the comparison, then categorized the writing based upon the comparisons.

Research Question 3: Exemplar Categorization

The third research question was, in what ways and to what extent do teachers use exemplar categorization in their grading decision-making process when grading samples of student writing in English language arts? To answer Research Question 3, an a priori code *categorizing-exemplar* was utilized to code the participant data. Participants categorized features of an essay or the entire essay by comparing the features to examples (exemplars) of essays they had read recently or had catalogued in their vast experience as teachers. Furthermore, the exemplars also could be types of students they had encountered in the past during their teaching career. Participants assessed the student writing based upon the comparison between the writing samples and the exemplars.

The *categorizing-exemplar* code was used in 25 instances of the cognitive laboratory interviews. Similar to classical categorization, the *categorizing-exemplar* code was concentrated to just a handful of participants. Twelve of the 21 participants exhibited exemplar categorization. The remaining nine participants did not display any exemplar categorization. It is important to note that this researcher did not find evidence in this study of a connection between the years of teaching experience and the participants who exhibited exemplar categorization. Participants of all experience levels used or did not use exemplar categorization.

Table 10 displays the participants' use of exemplar categorization and their years of teaching experience.

Table 10

Participant's Years of Experience and Exemplar Categorization

Participant	Experience	*Categorization-Exemplar* code *f*
T1	15	1
T2	24	4
T3	21	2
T4	19	0
T5	3	3
T6	10	0
T7	16	0
T8	2	2
T9	11	0
T10	3	0
T11	4	0
T12	3	0
O1	9	1
O2	16	2
O3	12	4
O4	1	1
O5	20	2
P1	14	2
P3	5	0
P4	22	1
A1	3	0

Note. Years of experience included the year of the study so that a new teacher was listed with one year of experience.

The left column in Table 10 represents the participant code with the second column being the teacher's years of experience including the year of the study. The column on the right is the

recurring phenomenon of exemplar categorization in participants' grading decisions. Although the definition of exemplar categorization includes a catalogue of previous essays or a catalogue of students in a teacher's memory gained through experience, there did not appear to be a relationship between the use of exemplar categorization and years of experience.

Participants exhibited exemplar categorization when they recalled previous essays in their recent memory and compared the student writing to the memories of those essays. Participant P4 stated, "And in my years of experience, I'd like to think I pay attention. This is the average essay of a student who's going to hand it in. This is going to be in that realm." It should be noted that Participant P4 had an understanding of the "average" essay based upon his years of experience. Even though he did not explicitly state that he recalled memories of previous essays, the connection Participant P4 made was clear.

Furthermore, Participant O4 also used memories of essays as exemplars when she stated, "And that's why I don't waste a lot of time with these, like, minuscule errors here and things like that because I know what good writing should be based on just my past experience." Her statement was another example of an indirect reference to the many exemplars she acquired during her teaching. Yet, Participant O2 used his vast catalogue of essays to compare the student's writing. He said,

> That was where I think the essay really did very well because…that's where people struggle with this kind of question…is dealing with the work as a whole and the theme and this student doesn't really struggle with that at least compared to most essays that I've read. This is a better essay regarding that. (Participant O2)

Participant O2 directly referenced his experience when he described the student's writing in relation to his catalogue of exemplars.

An additional type of exemplar did not come from the participants' experiences with students, rather from a catalogue of exemplars outside their classrooms. Participant O5 used Ernest Hemingway and Charles Dickens as exemplars to compare the student's use of variation to their work.

> The thing about sentence patterns, some authors write with very short sentences like Hemingway. His sentences are very short. Dickens writes with these long, elaborate, convoluted sentences. But even with someone who uses all short sentences and someone who uses all long sentences, there's variation…otherwise it's just putting one foot in front. It's, it's just plotting. And this [Essay 1] does not have a plotting sense to me. (Participant O5)

Although the use of important authors was an unusual exemplar choice to compare the student writing, Participant O5 indicated the student's essay used variation rather well. Participant O3 also used an alternative catalogue of exemplars when she accessed her collective memory of an online study guide website as an exemplar, observing, "This is sounding to me like they got this off of SparkNotes. So, I might go and Google search…just to make sure this is their own work. Because it's weird how much of a summary it is." Both Participants O5 and O3 used their experience of non-student writing to assess the essay.

Moreover, some participants used previous students, or knowledge of how previous students wrote as exemplars to grade the essays. Participant O3 stated, "So I'd, I would definitely think this is an honors kid who just can't, I can picture of the type of kid it was, [who]

just can't stop talking. So, they need to work on the organization here." Unlike previous examples, Participant O3 compared the essay to a specific student when discussing the essay. Participant O1 compared the writing to that of students she had in elementary school when she said, "So overall the grammar structure was not what should be at ninth grade. Even when I've taught elementary, like I feel like they did a better job at this." Additionally, Participant T2 said, "That's why…this is probably a freshman. I mean this is something I would get probably from my 9th graders or even my 10th graders." Participant T2 imagined the work she was grading as a member of one of the many students she has had in 9th or 10th grade in her experience.

The participants who used exemplars to grade the student writing compared their students to specific students in their memory, prototypes of previous students, and outside sources. In each instance, the memory was brought forth and was used to categorize the writing and determine the grade.

Research Question 4: Similarities and Differences in Grading

The fourth research question was, to what extent, if any, do similarities or differences exist in teachers' decision-making processes when grading samples of student writing in English language arts? To answer Research Question 4, open coding was performed on the 21 transcripts of the cognitive laboratory interviews. The open codes were then combined to form focused codes, converted into gerunds, and themes began to emerge. The processes the participants used to grade the student writing samples became clear early on in the interviewing and coding process and were confirmed through the constant comparative method of coding.

Two similarities were revealed through the qualitative data. The first similarity that emerged regarding the participants' decision-making processes was that each of the participants interrupted their reading of the student essays during the grading process. The interruptions occurred quite frequently. The participants would stop their reading multiple times during one paragraph or even during a single sentence to make various comments about the essay. Three types of interruptions were noted and they are as follows: (a) participants asked questions of themselves or the imaginary student, (b) participants considered and commented on what they just read, and (c) participants gave preliminary evaluations of the writing. The second similarity in participants' grading decision-making was that participants expressed a strong desire to use the state standardized writing rubric to evaluate the writing samples.

Two differences in decision-making emerged from the participant interviews. First, participants focused their interruptions, evaluative statements, and their overall evaluation focus either on the conventions of standard English or on the thesis in the essay. The second difference in the decision-making processes was that the participants either made binary decisions regarding the thesis or made quality decisions regarding the thesis. The following sections are divided into two parts. The first section focuses on similarities followed by the second section, focusing on differences.

Similarities

The coding data revealed that each of the 21 participants interrupted their reading of the two samples of student writing to ask themself a question or to pose questions to the imaginary student. Furthermore, they commented on various features of the student writing or wondered

aloud regarding the direction, completeness, or quality of the student's thesis. Additionally, participants expressed their desire to use a rubric to grade the student writing a total of 39 times.

Interruptions

The initial coding identified 557 open codes regarding interruptions that were then focused into 11 types of interruptions. The 11 focused codes regarding interruptions were further combined into two themes: *Interrupting-Thesis* and *Interrupting-Conventions*. Table 11 presents the coding data for the 11 focused codes and the two thematic codes for interruptions.

Table 11

Focused and Thematic Codes for Interruptions

Focused Code	*f*	Thematic Code	*f*
Interrupting-Addressing the prompt	18	Interrupting-Thesis	191
Interrupting-Claim/Thesis	173		
Interrupting-Comma	35	Interrupting-Conventions	366
Interrupting-Grammar	65		
Interrupting-Mechanics	78		
Interrupting-Spelling	40		
Interrupting-Structure	64		
Interrupting-Summary	22		
Interrupting-Transitions	40		
Interrupting-Visual Structure	4		
Interrupting-Vocabulary	18		

In Table 11, the left column represents the focused code for interruptions with the number of observations in the next column. The third column from the left represents the thematic codes that combine the interruptions into two types: *thesis* and *conventions* with their corresponding number of observations of the recurring phenomena.

Through the constant comparative coding method and the multiple readings of the cognitive laboratory interview transcripts, it became apparent that there were two primary processes the participants were using in their interruptions. First they were addressing the thesis

in their interruptions and in the second process the participants were addressing the conventions of standard English in their interruptions. Within these two themes the participants either made general remarks, asked themselves a question or asked the imaginary student a question.

The first type of interruption was a general remark regarding the thesis. Participant T11 made the general remark about the thesis when she said, "I'd like to know who Allie is in relationship to him there, but um, that's a good intro." Participant T4 said simply regarding the thesis, "So vague." Participant T12 remarked that the student did not address the prompt when she said, "This part [of the prompt] is specifically what I feel like the student didn't address." Participant T5 expressed confusion regarding the direction the student was taking. She said, "I'm actually a tad confused with this paper. I'd probably have to call the child over at this point going, uh, okay. What was your thesis?" Each of these quotations exemplifies the interruptions participants made to give a remark about the student's thesis.

Other participants interrupted their reading to consider or question the direction of the student's thesis. Participant O1 questioned the accuracy of the student's thesis when she said,

> Um, you know, they're saying society then he keeps saying society, but they're not mentioning society. They're mentioning direct things. They're saying like he was forced to lie because of his job, but that's not really society. That's his job. (Participant O1)

In the preceding quotation, Participant O1 commented that the student's thesis regarding society was inaccurate because the support the student provided was not about society, but about his job. In another example, Participant T10 praised the student when he said, "All right, so the student is focused on, you know, the influence of Allie's death and how this death has affected Holden in different ways. So that's excellent." Conversely, Participant O2 expressed some doubt about the

student's interpretation of the novel when he said, "That's an interesting interpretation. But sure." And finally, Participant T11 stated, "I underlined that [sentence] because I'm starting to see what their true focus of the essay is. I think is Allie and Holden's relationship." This comment by Participant T11 showed her need to interrupt her reading, annotate, and consider the direction of the thesis before moving on to the rest of the essay.

Another type of interruption regarding the thesis occurred when a participant interacted with the imaginary student. This interaction occurred as if the student was in the room having a conversation with the teacher about the essay. Participant P1 stated, "This needs more development so that you understand what society's explanation is. Our society's influence is on the line. Why does he have to lie? Um, this just says that he does it doesn't say why he has to." The use of the word "you" by Participant P1 showed that he was speaking directly to the imaginary student, rather than the researcher in the interview. Participant A1 interacted with the imaginary when she asked, "How is a character influenced by a person or factor? So, they have a friend, family member, spiritual belief or society. So, is it family though? It's the loss of his family member." And finally, Participant T8 asked a series of questions of the imaginary student. "This is really, I might say, is this your thesis? Is this your claim?" These questions asked by Participant T8 are examples of a direct line of inquiry towards the student.

The final type of interruption regarding the thesis occurred when the participants asked themselves a question or pondered to themselves regarding the thesis to help process what they were reading in order to evaluate the quality of the writing. Participant T3 openly discussed her thoughts about the direction of the student's thesis when she said,

> So, I'm assuming that that's going to be relevant to how it's important to the work as a whole. However, I'm not sure. I don't feel, I feel like maybe he could have done more with the thesis so far, but I'm going to continue to read to see if it's more of an open thesis and if he can definitively answer the question. (Participant T3)

This remark by Participant T3 occurred towards the beginning of the essay and indicated her willingness to hold off on evaluating the completeness of the thesis until she read further.

Similar to Participant T3, Participant T4 also wondered aloud about the direction of the thesis when she said, "That has a claim there. Now I'm hoping to read at this point since he actually wrote a claim and probably should have been a paragraph break there. I'm hoping there is evidence to support that claim." Participant T9 also wondered to herself if the student's writing would follow through by supporting the claim. She said, "We'll see how it goes if he follows that last one correctly. Then we should be on the right track. But then I'm with it just moving directly into the next paragraph." Participant O2 noticed the incompleteness of the student's thesis part way through the reading when he mentioned, "So if the question is how is he influenced and how is it important to the work as a whole? While so far we haven't discussed anything about how it's important to the work as a whole." Lastly, Participant O4 discussed confusion as to the point the student was attempting to convey in his writing. "This doesn't make any sense to me. I'm not understanding what this has to do with *The Giver*, the man himself. So, this seems out of place. Not, well…if it's relevant, it's not very well explained."

The preceding quotations from the participants were just a sample of the 191 instances in which they interrupted their reading of the text to interact with and actively engage in the student

writing. Each of the 21 participants interrupted their reading of the student writing multiple times to consider, question, and preliminarily evaluate the student work.

The second type of thematic interruption involved the conventions of standard English. Participants interrupted their reading of the student texts 366 times to consider writing conventions compared to 191 instances of interrupting to consider the thesis. Although this may have appeared quantitatively that participants were more concerned with conventions than with the thesis, it was difficult to make that connection. Each essay had only one thesis and perhaps a topic sentence for each paragraph. The opportunities for a participant to interrupt their reading and discuss the thesis and topic sentences were relatively limited. In comparison, the opportunities were far greater for a participant to interrupt their reading to discuss grammar, spelling, and punctuation, let alone other mechanics and usage issues.

Quite often participants interrupted their reading to notice or comment on errors in conventions as they read the samples of student writing. The next several quotations are comments merely noting the error occurred without any evaluative statements or interactions with the text or the imaginary student. Participant T12 said, "…comma, should be a comma…", and Participant T6 noticed "…that needs to be two words." Participant T3 stated, "So there's…a grammatical error here" and she also stated later regarding the same essay, "…so he's got a transition here." Lastly, Participant T5 noted poor vocabulary when she said, "'shut out?' Word choice." None of the interruptions by these participants indicated an evaluation of the student work. Rather, the participant verbally noted the issue without any further consideration.

Although many interruptions regarding the conventions of standard English were remarks noting that the error occurred, many other interruptions considered the effect the error had on the

overall quality of the essay. Participant A1 explained her grading process after reading a lengthy passage in Essay 2 that inhibited her ability to understand the thesis. She stated, "When I have an essay like this where it's fairly long without any breaks in between, I'll usually read through the whole long chunk and then go back to try to figure out what the main…idea is." Participant T10 offered a more thorough evaluation,

> Excellent. So, when students use elaboration, techniques like transitions, this shows that this is the explanation but should be followed by a conclusion or at least the end of the idea. The student does end the idea and the sentence, [but] there's no conclusion. Sentencing, therefore this is what is happening because there is no, there's no structure to this essay. It's structured and…there's a claim, there's evidence, there's elaboration, but the physical structure that allows the reader to kind of, uh, you know, flow easier and the cues of the transitions from paragraph to paragraph will allow the reader to, you know, have an easier and clearer read. (Participant T10)

In the preceding quotation, Participant T10 interrupted his reading and noted that the lack of structure to the essay was impeding his understanding and affecting the overall quality of the essay.

Participant O5 also noticed Essay 2 included a paragraph that lasted the entire first page. She said, "Oh, it dismays me to see that the entire first page is one paragraph. Are there any paragraph breaks?" The participant flipped through additional pages of the essay. "Yes. Okay. There are paragraph breaks, but the entire first page is one paragraph." In another example, Participant T10 was particularly concerned with comma usage. His remarks about commas resulted in 12 instances of the 35 *interrupting-comma* codes. The following is an example of a

comment about commas and his process for determining any necessary point deductions. Participant T10 said,

> Interesting. There's a list here that also has a comma that could possibly be missing. It depends on if the student uses the Oxford comma or not. So, we can leave this alone because some teachers tend to teach kids to not use that comma. Some teachers tend to tell kids to use a comma. So, so far in the first paragraph, I maybe see a pattern in punctuation errors. But if I see a pattern with a specific punctuation, then that's when I will begin deducting points. (Participant T10)

Finally, Participant O1 pointed out conventions in a positive way unlike the majority of participants who noted only errors. She said, "So I circled *symbolize* because again they're using, you know all these like figurative language devices to explain things. Meanwhile, there's another transitional word that most people don't use." This alternative use of the interruption by Participant O1 was a rare example of the participant noticing a proper use of conventions by the student.

The preceding quotations illustrate the numerous ways in which participants interrupted their reading of the student writing in order to consider conventions. Although many of the interruptions were just to note that an error occurred, numerous other interruptions were more evaluative in nature and described the effect poor conventions had on the overall quality of the essay.

Need a Rubric

The second similarity to emerge from the data was the expressed desire to use the state standard writing rubric to grade the student writing samples. The participants were instructed to use the grading method they would normally have used in their own classroom (Appendix I). Although the participants were not asked to use a rubric to grade the essays, they mentioned quite often that they needed access to the state writing rubric or that it was unfair to grade student work without any rubric at all. The code *Discussing Rubric* was developed to capture these statements and was coded 39 times in the 21 cognitive laboratory interviews.

Participant T3 wondered which method she would use to grade the essays when she said, "So am I going to be using the [state standardized writing] rubric for this?" Participant T8 expressed that she preferred to have a rubric when grading student writing. She said, "I've changed my mind. I got a little teary eyed. I love that book and I haven't read it in a while and I'm getting um, ugh but woosh, I like having a rubric." Participant T11 echoed those statements when she said, "So I feel like I would give this one, I'm sorry, I'm so used to having a rubric already. Like pre-calculated. I feel like I should give this one an 80, like a very low B." This discomfort at the lack of a rubric did not stop those participants from using it from memory.

The discomfort at a lack of a rubric continued with Participant T3 who said, "I wouldn't really feel comfortable just giving him a grade. I don't think that that's good teaching. I would definitely want a rubric and…I would teach the students, uh, how to use the rubric." Participant T5 added that she felt a rubric was not just a grading tool but a teaching tool as well. She said, "I believe in all grading, writing should have a rubric where the kids have the rubric in advance…to practice and then so they know by the time you got to the ninth grade it would be a given."

Whether the rubric is a teaching tool or an assessment tool, the participants felt it was a necessary companion to grading writing.

Finally, Participant O2 discussed the need for a rubric and also expressed difficulty in translating the score from a rubric to a score based on 100 points as the cognitive laboratory instructions asked the participants to do. Participant O2 said,

> I would use a rubric…because this is such an AP prompt, I'm instinctively reading it like an AP prompt and applying the AP rubric and on the AP rubric, um, this would be insufficient. So, on the scale of one to nine using the old rubric, which they don't do anymore. To me this would be a four, maybe a three and five and five and up is passing. When I say passing, I mean on college level. So, using that logic 75 would be inappropriate cause it's three is a C in AP size. So really there should be a D. So, if I'm really thinking about that then really 75 would be incorrect. I would have to say in the 60s somewhere because a 60 would be a D maybe 65. (Participant O2)

This difficulty in translating the rubric score to a score out of 100 was noteworthy because student grades on report cards and transcripts are reported in percentages with letter grades.

Differences

The participants exhibited two differences in their decision-making processes while grading student writing in English language arts. Those differences were, (a) the participants focused their interruptions, evaluative statements, and their overall evaluation either on the conventions of standard English or the student's thesis and (b) the participants either made

binary decisions regarding the thesis or made decisions regarding the quality of the thesis. These differences in participant grading decision-making are discussed in the following sections.

Focus: Conventions or Thesis

A distinct theme emerged regarding the interruptions a participant made while grading the student essays as well as the evaluative statements the participant expressed regarding the essays. The participant either made interruptions and evaluative statements based upon the conventions of standard English or they made interruptions and evaluative statements based upon the thesis. Namely, the participant focused either on what the student wrote, or on how the student wrote it. Furthermore, the overall evaluation focus was also either on the conventions of standard English or on the thesis.

An *interruption* was anytime a participant stopped reading the text aloud to verbalize comments, questions, or evaluations regarding the essay. Interruptions could be brief, such as "Spelled wrong" (Participant T4) or could be longer and preliminarily evaluative such as the interruption Participant T3 expressed regarding the potential direction of the thesis. She interrupted her reading and stated, "So I'm assuming that's going to be relevant to how it's important to the work as a whole. However, I'm not sure" (Participant T3). In contrast, an *evaluative statement* is a comment that was expressed either while participants were interrupting their reading or it could be after the reading of the essay concluded and the overall evaluation was ocurring. Regardless of when the comment was made, the purpose of the evaluative statement was to give an evaluation either of the student's use of the conventions of standard English or of the thesis.

The coding data for *interruptions* indicated two distinct paths the participants took as they interrupted their reading while grading the student writing samples. These paths were evident for 16 of the 21 participants. Within those 16 participants' coding data, participants focused more on conventions or they focused more on the thesis. More specifically, the participants focused on and interacted with how the essay was written, or focused on and interacted with the content of the essay. Table 12 shows the relevent data regarding participant interruptions to consider the conventions or the thesis of the student writing sample.

Table 12

Participant Interruptions to Consider Conventions or Thesis

Participant	Interrupting-Conventions *f*	Interrupting-Thesis *f*
	Conventions *f* > Thesis *f*	
T1	9	0
T2	17	0
T3	20	6
T4	19	13
T5	31	8
T6	12	3
T7	21	0
T9	19	12
T10	26	6
T12	30	3
O1	14	4
O3	21	6
	Conventions *f* < Thesis *f*	
O4	6	11
P1	6	15
P3	0	13
A1	7	15
	Conventions *f* ≈ Thesis *f*	
T8	12	13
T11	8	8
O2	17	16
O5	19	16
P4	11	12

Table 12 is divided into three sections. The first section represents the participants whose interruptions regarding conventions of standard English outnumbered the interruptions regarding the student's thesis. The second section represents the participants whose interruptions regarding the thesis outnumbered the interruptions regarding the conventions of standard English. The third section indicates the participants whose interruptions for both were approximately equal to

each other. The coding for the 16 participants who exhibited a focus on either the conventions or the thesis ranged from a difference of 6 codes (Participant T4) to as many as 27 instances (Participant T12). It is important to note that three Participants, T1, T2, and T7 did not interrupt their reading at all to consider the thesis of Essay 1 or 2.

Similarly, the evaluative statements participants made regarding conventions and the thesis indicated a difference as well. An *evaluative statement* could be an interruption during the reading of the essay or it could be a statement that was made after the reading of the essay had concluded. In both cases, an evaluative statement either focused on the thesis or the conventions of standard English and this difference was evident in the coding data for 14 of the 21 participants. Table 13 presents the coding data for the evaluative statements.

Table 13

Participant Evaluative Statements Regarding Conventions or Thesis

Participant	Evaluating-Conventions *f*	Evaluating-Thesis *f*
	Conventions *f* > Thesis *f*	
T2	7	0
T7	7	3
T10	17	9
O1	12	2
	Conventions *f* < Thesis *f*	
T3	7	10
T8	8	11
T11	5	9
O2	2	21
O3	2	12
O4	1	9
O5	6	17
P1	3	8
P3	1	13
P4	3	16
A1	4	22
	Conventions *f* ≈ Thesis *f*	
T1	4	4
T4	10	9
T5	9	9
T6	5	5
T9	10	11
T12	0	0

Table 13 is divided into three sections. The first section represents the participants whose evaluative statements regarding conventions of standard English outnumbered the evaluative statements regarding the student's thesis. The second section represents the participants whose evaluative statements regarding the thesis outnumbered the evaluative statements regarding the

conventions of standard English. The third section indicates the participants whose evaluative statements for both were quite similar to each other.

Three Participants, O5, O2 and T6, illustrated the difference between focusing on the thesis or focusing on the conventions. It is worth delving into their interviews more deeply to investigate this difference. Participant O5 focused primarily on the thesis, or the *point* as she described it in her interview. She interrupted her reading 16 times over the two essays to interact with the thesis and interrupted her reading 19 times to consider conventions. However, the content of the interruptions clearly indicated her focus was on the thesis. Regarding Essay 1, she made a series of interruptions that indicated a sustained effort at interacting with the thesis. She said,

> So, the first essay always takes me the longest to grade because I have to refer back to the prompt and let me go back because I know he's mentioned Jonas and mentioned the giver. I don't have a clear sense of whom is being compared to what. So, I'm going to go back and reread the first paragraph. (Participant O5)

In this quotation from Essay 1, Participant O5 opened with a misunderstanding of the purpose of the first paragraph. She did not remark about conventions but was concerned with the meaning the student tried to convey in his writing. She continued with more concern about the direction of the thesis, saying:

> All right, so Jonas is affected by the giver. Um, uh, I'm a little bit muddled on the direction of this essay. It might become clearer if I weren't so rusty on The Giver. I might be clear, but let me keep going. Um, my first thought is there's not a very clear thesis. I'm not exactly sure where this essay is going. (Participant 05)

This further illustrated her primary concern about what the student was trying to express in Essay 1, not how he was expressing it. She continued along the same line of inquiry when she said, "Well, how does that relate to the line and fake life? Oh, well, okay. I see how it relates to a fake life, but we don't have much continuity here." These three quotations from Participant O5 were more than comments regarding the thesis. Rather, they were indicative of an interaction with the thesis as she graded the essay.

In contrast, Participant O5 made numerous statements regarding the conventions of standard English in Essay 1, but they were transitory and not of a nature in which she was interacting with the conventions. In the same essay, she said, "…apostrophe 's'" and "misspelled." She further made a more significant interruption when she said, "Got choppy sentences here. It's not very sophisticated writing." This interruption, although transitory in one sense, also represented more of an evaluative statement when she commented about the lack of sophisticated writing.

Given the transitory interruptions regarding conventions and the interactive focus on the thesis, her true evaluative focus was revealed in the evaluative comments stated at the conclusion of Essay 1. This exchange between the researcher and Participant O5 highlighted her evaluative focus.

O5 Um, this essay lacks focus and precision. I'm trying to separate a 10th grade student out of a senior student. The point lacks clarity. It's a passing essay but it's not a very good essay. I would probably give it, this is the fourth quarter, 70%.

Researcher: What about it makes it a passing essay? Because that was the first thing you said, "It's a passing essay."

O5: It's a passing essay. I try to keep in mind, uh, what it takes to be a high school graduate. The thinking is rather muddled. I can see that the thinking is muddled and it lacks clarity but it's not so bad as to be nonsensical. If it were nonsensical, I would have to give it a lower score.

Researcher: So if I were just to repeat back to you, the student is making a point?

O5: The student is making a vague, he's making an unclear point. He's not making "no point."

Researcher: So a really clear point is going be in the A range, a much fuzzier one, that's going to be in the C/D range. No point is going to be the, the F range and that all of the other mechanics, sophistication and whatnot can help it get a little bit higher or maybe a little bit lower than where the clarity of the point is. But that's all fixable with feedback.

O5: Yes.

Researcher: Um, but [the grade] really is the clarity of the point?

O5: There has to be a clear point, there has to be a clear point and there has to be some level of clean writing and sophisticated writing and sometimes that's easily fixed.

In this exchange, Participant O5 evaluated the student for making an *unclear point*. There is no mention of conventions, mechanics, or other features she expected to see. The basis for the grade was the unclear point in Essay 1 with no mention of the conventions of standard English.

In contrast to the evaluation focus on the thesis for Participant O5, Participant O2 discussed the *formula* in which students were taught to use the conventions of standard English in a specific way to construct an essay. He said, regarding Essay 1,

> So, if I look back at this, I still feel like we've never actually touched on how it's important to work as a whole. We have just summarized summary, summary, summary. Um, the, they've got the formula down, they have a beginning, middle and end. That's good. They have topic sentences, they got the thesis. Clearly they've had a teacher who, you know, figured...drill that into them. Um, but there's, there's very little elaboration, uh, beyond summary.

This statement by Participant O2 is significant for two reasons. He defined the formula in detail and also included that although a thesis existed, it was lacking, did not adequately address the prompt, and was not supported by elaboration.

Participant T6 focused almost exclusively on small grammatical errors and the formula as defined by Participant O2. Participant T6 began his interruptions with, "Jonas's should be a plural or like you should have the um, apostrophe." He then pointed out a spelling error with the word *therefore* when he said, "missing an 'e'." He continued to notice more grammar errors with, "apostrophe missing there" and he noticed a word choice error when he said, "I think you mean the great beyond. Maybe I'd underline that and put a question mark over it." Participant T6 concluded his analysis after he read the entire essay with the following discussion:

> Okay. So now that I've read the whole thing, I would go back and look at the prompt, um, and I'll see what it's asking for. So it's asking you in a well developed composition to identify a character. He did that. Describe the characters influenced by the person or

factor. Okay. It was by society. Got that. How the character's experience is important to the work as a whole. So it looks like I can check off all those boxes. So my biggest issue that I would, uh, take points off for…in this piece would be, um, some of it like the grammar errors, um, the way that they wrote some of these sentences. (Participant T6)

The final sentence indicated the focus on the grammar for the evaluation of the essay for Participant T6. This was significant because Participant T6 was grading the essay for the formula, rather than for the quality and depth of the thesis in the essay.

At the conclusion of reading each individual cognitive laboratory transcript and studying the codes and themes, the data suggested an overall evaluation focus of each participant. Participants either based their grading decisions upon the *formula* or based their grading decisions upon the *point* (thesis). Participant O2 discussed the writing formula which included topic sentences, thesis, beginning, middle, and end (Participant O2). In contrast, Participant O5 used the term *point* to describe the thesis. She further elaborated the differences between a strong point, weak point, and no point (Participant O5). To this researcher, these terms align with the primary assessment focus of the participants. This researcher categorized all 21 participants as either focused on the formula or focused on the thesis. Although 19 participants included a secondary focus on the thesis or the writing formula to some degree, the degree to which each participant primarily focused on one or the other became clear through a study of the qualitative data. The data used to categorize each participant were as follows: (a) coding data regarding the interruptions, (b) coding data regarding the conventions vs. the thesis, (c) coding data regarding the binary thesis decision or the quality thesis decision, and (d) the overall researcher's impression of each participant's interruptions and evaluative statements spoken

during the entirety of the cognitive laboratory interview. Eleven of the participants focused on the thesis and 10 participants focused on the formula as their overall evaluation focus. Table 14 presents the relevant data regarding the evaluation focus of each participant.

Table 14

Evaluation Focus on the Thesis or the Formula

Participant	Evaluation Focus
T1	Formula
T2	Formula
T3	Thesis
T4	Formula
T5	Formula
T6	Formula
T7	Formula
T8	Formula
T9	Thesis
T10	Formula
T11	Thesis
T12	Formula
O1	Formula
O2	Thesis
O3	Thesis
O4	Thesis
O5	Thesis
P1	Thesis
P3	Thesis
P4	Thesis
A1	Thesis

In Table 14, the column on the left represents the participant code and the column on the right indicates the evaluation focus of the participant being either the writing formula or the thesis.

Binary or Quality

An additional theme that emerged was the type of decisions participants made regarding the student's thesis. Either participants made a binary decision regarding the thesis or they evaluated the quality of the thesis. A *binary decision* is a choice between two alternatives. Within the context of writing assessment decisions regarding the thesis, the participant's decision rested upon whether the student expressed a thesis or did not express a thesis. More specifically, the participant noted the essay had a thesis without considering the quality, completeness, or elaborative support given to the thesis. In contrast, a *quality decision* is one in which the participant evaluated the quality of the thesis. Was the thesis complete, specific, debatable, and was the evidence effective at strengthening the thesis? Or were the thesis, support, and elaboration weak? In a quality decision, the participant made a judgement about the thesis determining if the thesis was specific, debatable, and supported by evidence.

Upon further examination of the *Evaluating-Thesis* codes, an additional round of coding was used to identify which evaluative statements regarding the thesis were binary and which statements assessed the quality of the thesis. Of the 200 codes for *Evaluating-Thesis*, 54 indicated binary decisions and 146 indicated decisions for thesis quality. These codes were concentrated within certain participants which suggested the participants either noted and gave credit because the student expressed a thesis or evaluated the quality of the thesis based upon its specificity, strength, and elaborative support. Table 15 presents the coding data for evaluative comments regarding the thesis.

Table 15

Participant Thesis Evaluations: Binary or Quality

Participant	Evaluating-Thesis Binary *f*	Evaluating-Thesis Quality *f*
	Binary *f* > Quality *f*	
T1	4	0
T4	7	2
T5	9	0
T6	3	2
T8	10	1
T10	9	0
T11	7	2
O1	2	0
	Binary *f* < Quality *f*	
T3	0	10
T7	1	2
T9	2	9
O2	0	21
O3	0	12
O4	0	9
O5	0	17
P1	0	8
P3	0	13
P4	0	16
A1	0	22
	Binary *f* = Quality *f*	
T2	0	0
T12	0	0

Table 15 is divided into three sections. The first section contains the participants whose evaluative statements regarding the thesis were binary. The second section contains the participants whose evaluative statements regarding the thesis were for quality. The final section

indicates the two participants who made no evaluative statements regarding the thesis for either essay. Additionally, the first column on the left indicates the participant code. The middle column presents the recurring phenomenon binary thesis decision-making with the last column displaying the recurring phenomenon of quality decision-making regarding the thesis.

The participants' statements regarding the thesis illustrated more clearly the difference between a binary evaluation and a quality evaluation of the student's thesis. Participant T4 exhibited binary thesis evaluation when she remarked that the student described how the character was influenced by a friend but did not explain how the experience was important to the work as a whole. Participant T4 said,

> Okay, so if I go back to the three things identified in the prompt, this student did identify the character…He does mention the character a lot, so he definitely, this person would definitely get good points on identifying the character. This particular essay does not describe how the character is influenced by their, a friend, a family member or spiritual belief for this society. It does not identify how the character's experience is important to the work as a whole, um, until the very, very end of the conclusion. (Participant T4)

It is important to note that Participant T4 used the three portions of the writing prompt as a checklist and indicated that the student identified the character, but did not complete the final two portions of the prompt. This was an example of binary decision-making regarding the thesis because Participant T4 did not address the quality of the thesis. Rather, she only discussed that the student did not address the prompt adequately.

Participant T5 also made binary decisions as if the thesis was another text feature to be included in the essay when she said,

> Having that clear cut thesis, an intro paragraph where I could count and say, okay, here's your thesis topic, sentence, topic, sentence, topic, sentence, conclusion. I should be able to see that. At this point, I should be able to say, here's my evidence, here's my explanation. I'm elaborating on it and tying it back to the topic sentence. (Participant T5)

This participant also compared the student's writing to a checklist of text features. Her statement did not include any evaluative assessments of the quality of the thesis or the quality of the text features that she listed. Rather, she merely noted that the essay lacked the text features she expected to see in the specific order in which she expected to see them. Participant T5 discussed in even more specific terms including the phrase *check off* when she said,

> So I would be looking for word choice with points. I would be looking at sentence structure with points. Um, do you have it, you know, the structure? Do you have an intro with an attention, um, or you know, some kind of hook with an intro, tying into a thesis statement that would be worth so many points, you know, do you have a topic sentence? Do you have evidence? Do you explain your evidence to tie back to your topic sentence? Um, do you have a concluding sentence? Do you have transition words? So it would definitely be a very detailed rubric that I could check off as I went. (Participant T5)

In the previous two quotations by Participant T5, it is clear that she did not make any decisions about the quality of the thesis or the other text features of the essay. To her, a rubric was a checklist of items to be included in an essay.

Furthering the concept of a thesis as a part of a checklist, Participant T6 specifically mentioned *check off all those boxes*. He said,

> So it's asking you in a well-developed composition identify a character. He did that. Describe the characters influenced by the person or factor. Okay. It was by society. Got that. How the characters experience is important to the work as a whole. So, it looks like I can check off all those boxes. (Participant T6)

The preceding quotation by Participant T6 was direct evidence of the checklist he used to evaluate the student's essay. Later in the same essay, Participant T6 continued with binary thesis decision-making when he said,

> He identifies a character. He describes how the characters influenced by the story he explained. This is how the experience, he explains how the experience is important to the work as a whole. So, he does, he crosses off all those boxes. So, to me, even though I think the writing could be better, ultimately, he did follow the prompt to the 'T'. (Participant T6)

The preceding four quotations by Participants T4, T5, and T6 indicated the thesis was evaluated in a binary manner similar to the binary decisions regarding the conventions of standard English. The thesis and conventions of standard English were either present in the essay or they were not present.

Additional participants exhibited similar binary decisions regarding the thesis. Participant T8 said, "Um, I think it addresses the prompt. I think it needs some tweaking, but it has the evidence." Although the phrase "needs some tweaking" could have indicated a quality decision regarding the thesis, the next phrase "has the evidence" is evidence of a binary decision because there was no mention of the quality of the evidence. Moreover, Participant T10 indicated a binary decision-making succinctly when he said, "…there's a claim, there's evidence,

there's elaboration." Indeed, Participant T10 analyzed the student's essay for the existence of those text features, not if the text features were of any particular quality.

On the other hand, the evaluative statements regarding the quality of a student's statement were in stark contrast to the binary statements made by the previous participants. Participant T11 delved into the need for the student to develop the thesis more when she said,

> I'd like to see more development on how that supports the idea of him having to lie and how he actually lies to his parents or why that is significant. I feel like it's connected, but I'd like to see the student's reasoning behind it. Um, influence him not to be able to love. I think that his ideas or her ideas are communicated well here. Um, but I feel like it could use more development sophistication. (Participant T11)

The difference between the binary thesis decision-making quotations and the preceding quotation by Participant T11 was significant. Participant T11 did not evaluate the essay for the presence of a thesis; rather, she acknowledged the presence of the thesis and also evaluated that the thesis needed more development.

Participant O2 had a series of four statements that pointed towards an active interaction with the text and with the imaginary student. Participant O2 started with, "All right. But your entire argument here is that Allie is influencing this. So, you need to bring this back to Allie and you, we've moved beyond Allie, I think you forgot about Allie." Participant O2 further discussed that the student needed to be clearer and more explicit with the thesis and elaborative support when he said,

> So if we're trying to talk about our ability, how we handle pain and loss, that seems to be their thematic interpretation of the book. Uh, it's, I want that to be clearer and more

> explicitly addressed in making that connection to that thematic interpretation that could have happened right there, but it didn't. (Participant O2)

Participant O2 continued with the same text and interacted with the imaginary student as the essay progressed. His statements appeared to be a running commentary on the development of the thesis as if it was a private conversation with the student. He said,

> So, I love all this because this is sort of the, this is the meaning of the book that I, that I always, that I want my students to focus on, that I want you to focus on. Um, I and I'm glad you're going here and I wish this was where you really put your focus on instead of the pain and loss element. Again, that's not wrong, but I think it would be a better, more cohesive essay if this was your focus throughout. It took a while for you to finally get here and I wish it was more consistently developed throughout. (Participant O2)

He continued with his conversation with the imaginary student,

> All right, you're, you're, you're skirting around it. You're, you're not saying it. You're doing the thing where you just like, it's important to the work as a whole, but you're doing so well. Don't, don't mess that up, right? (Participant O2)

These four quotations by Participant O2 were significant because they showed a sustained effort to interact with the text and the imaginary student. He was not just placing figurative checkmarks next to the thesis; rather, he asked questions and challenged the student regarding his claims and evidence.

Participant O4 discussed in his evaluation that the thesis had changed during the course of the essay and therefore was unclear. He stated,

> Okay, so now we're kind of losing track here. So now there's introducing an entirely new argument. So, the first part of this paper is blaming or making a connection between the death of Allie on his, um, his mental state or his inability to have relationships. And now they're starting introducing a whole new argument. Also, more generally speaking, the entire story is a result of Holden being unstable after losing his brother. So that needs to be taken out. (Participant O4)

The discussion by Participant O4 indicated that he was actively tracking the development of the student's argument over the course of the essay and was paying close attention to the claim the student made.

Lastly, Participant O5 discussed how to improve the quality of the student's thesis when she stated,

> It would be better if he ended with the point of Allie's influence rather than just restating that ally has influenced him. The point is a little bit muddled in the beginning, but it becomes stronger as the essay moves on. (Participant O5)

She continued with another significant statement when she compared an unclear thesis to the blurry photographs that she has personally taken. She said, "It's sort of muddled. It needs to be sharpened. It needs to be clarified…He doesn't, it lacks focus. It's sort of like looking at a picture. Most of my pictures aren't focused, but I can still tell what's there." In this quotation by Participant O5, she interacted with the thesis and evaluated it for its clarity and assessed the changing quality of the thesis as the essay progressed.

Table 16 displays the similarities and differences for Research Question 4.

Table 16

Research Question 4: Similarities and Differences

Similarity or Difference	Category	Description
Similarities	All participants interrupted their reading of the essays.	Participant interrupted their reading to ask questions of themselves or the imaginary student. Participant interrupted their reading to consider and comment on what they just read. Participant interrupted their reading to give preliminary evaluations of the writing.
	Expressed desire to use a rubric.	Fifteen of the 21 participants expressed a desire to use the state standardized rubric when given the choice of grading procedures.
Differences	Focused either on the conventions of standard English or on the thesis.	Interruptions were focused either on the conventions of standard English or on the thesis. Evaluative statements were focused either on the conventions of standard English or on the thesis. Overall assessment focus was either on the conventions of standard English or on the thesis.
	Binary or quality decisions regarding the thesis.	Participants made evaluative decisions regarding the thesis that were either binary in nature or assessed the quality of the thesis.

Research Question 5: Differences in Grading High and Low Essays

The fifth research question was, to what extent, if any, do teachers' decision-making processes differ when grading student writing samples of high and low performance levels in

English language arts? There was no consensus as to the difference in participants' decision-making regarding the student writing for Essay 1 regarding if the student had adequately addressed the writing prompt. Seven participants believed the writing prompt was adequately addressed, but the remaining 14 participants believed the prompt was not addressed. For Essay 2, there appeared to be consensus that the prompt was adequately addressed, with only Participant T1 dissenting. Addressing the prompt was a frequent topic among all 21 participants and the researcher asked in the cognitive laboratory interviews specifically if participants believed the prompt had been addressed. The code *Assessing-Addressing the Prompt* was created to capture these data with an occurrence of 81 instances across the 21 cognitive laboratory interviews. This researcher wrote in the field notes if the participant believed the prompt had been adequately addressed either from the participant's direct words or from the researcher's interpretation of words and actions. Table 16 presents the relevant data regarding the writing prompt.

Table 17

Participant Decisions about Addressing the Prompt

Participant	Essay 1	Essay 2
T1	No	No
T2	No	Yes
T3	Yes	Yes
T4	No	Yes
T5	Yes	Yes
T6	Yes	Yes
T7	No	Yes
T8	Yes	Yes
T9	No	Yes
T10	Yes	Yes
T11	Yes	Yes
T12	No	Yes
O1	Yes	Yes
O2	No	Yes
O3	No	Yes
O4	No	Yes
O5	No	Yes
P1	No	Yes
P3	No	Yes
P4	No	Yes
A1	No	Yes

Note. A *Yes* indicates the student addressed the prompt adequately, a *No* indicates the student did not address the prompt adequately.

In Table 17, the left column represents the participant, the next two columns indicate the decision the participant made regarding the prompt. A *Yes* indicates the student addressed the prompt adequately, and a *No* indicates the student did not address the prompt adequately. For Essay 1, seven participants felt the student had adequately addressed the prompt while 14 participants did not. For Essay 2, one participant did not feel the student had addressed the

prompt with the remaining 20 participants evaluating that the student adequately addressed the prompt.

The 2018 Massachusetts Comprehensive Assessment System writing prompt is as follows:

> Often in works of literature, a character is influenced by another person or factor. From a work of literature you have read in or out of school, select a character who is influenced by **one** of the persons or factors listed in the box below.
>
> - a friend
> - a family member
> - a spiritual belief
> - society
>
> In a well-developed composition, identify the character, describe how the character is influenced by the person or factor, and explain how the character's experience is important to the work as a whole. (Massachusetts Department of Elementary and Secondary Education, 2018)

The 2018 MCAS writing prompt asked the student to accomplish three tasks of varying difficulty and complexity. The three tasks were (a) identify a character, (b) describe how the character is influenced by a person or factor, and (c) explain how the character's experience is important to the work as a whole.

Essay 1 represented a sample of student writing about *The Giver* that contained a few errors in the conventions of standard English. "Despite some minor errors in grammar and other aspects of conventions, (e.g., "If Jonas told anyone, him and the people he told….") and other aspects of conventions, their presence does not interfere with the reader's understanding of the

composition" (Massachusetts Department of Elementary and Secondary Education, 2018). Essay 2 represented a sample of student writing that was a "fully developed composition logically organized around Holden's experience after his brother's death in *The Catcher in the Rye*" (Massachusetts Department of Elementary and Secondary Education, 2018). Of the two essays, Essay 1 was considered lower scoring because of its errors with conventions. In contrast, Essay 2 was a fully developed essay regarding the topic/idea that had few errors in the conventions of standard English.

After each cognitive laboratory interview transcript was read multiple times and the constant comparative method of coding was completed, the decision-making processes for grading high and low scoring essays were observed to be relatively identical with no noticeable differences. For both essays, the participants interrupted their reading to ask questions of themselves or the imaginary student, to consider features of the writing, or to give preliminary evaluations of the writing. The observed grading differences were related to the features of the essays the participants primarily focused upon when grading each essay.

Regarding Essay 1, the participants focused on the numerous errors in the conventions of standard English or the lack of a fully developed thesis. For Essay 2, the participants almost universally noted the fully developed thesis with few errors in conventions. These findings were to be expected considering the essays were presented as examples of those assessments.

However, the participants generally disagreed with the assessment of Essay 1 presented in the 2018 Massachusetts Comprehensive Assessment System Scoring Guide as provided by the Massachusetts Department of Elementary and Secondary Education. The participants of this study reached a near consensus in that the errors in the conventions of standard English did in

fact interfere with the reader's understanding of the composition. Although the decision-making processes for grading the lower essay were similar to the processes for the higher essay, those aforementioned processes were disrupted by the amount of convention errors in the essay. The disruption in their processes interfered with the participants' ability to comprehend the essay and, therefore, they were not reliably able to determine the student's thesis or whether the student had adequately addressed the writing prompt.

The common theme among the participants was that Essay 1 lacked sophisticated writing and that this lack of sophistication interfered with the composition. This lack of sophistication manifested itself in that they felt the essay was too short, the sentences were also too short and *choppy*, and the essay was repetitive. Of the 21 participants, 15 commented about the lack of sophistication and used that determination as a basis for their assessment decision. Participant T2 said, "I would say it's very brief, like too short. I don't know…the whole entire essay was very…weak in a sense that there was just too much repetition of Jonas, Jonas, Jonas, Jonas, Jonas, Jonas, Jonas, Jonas, Jonas." Another participant also noted the repetitive nature of the sentences when Participant O5 said, "How many sentences begin with the word Jonas? One, two, three, four, five, six, seven, eight, nine, ten. It's a repetitious sentence pattern. I missed one…it's not very sophisticated writing." The repetition in the writing of Essay 1 distracted the participants so that they could not focus on the point the student tried to make in the essay.

Other participants noted the brevity of the essay and even hinted that the length alone was an indication of a lack of sophistication. Participant O4 said, "This would be probably somewhere in the C range for me just based on the length of it. First of all [it] is very short for an essay. Um, there's a lack of a thesis [which] is really something." It is interesting to note that

Participant O4 determined the letter grade of the essay based upon just the length. Participant T12 also took the shortness into consideration, but noted the lack of a thesis as the basis for the grade as well. She said,

> I would give a 50 because it doesn't address the prompt entirely and it is so short. There's not a lot of explanations. So to me it just seems like, like I'm reading the back of the book instead of actually getting a literary analysis. So especially comparing it to the second [essay], the second response has so much explanation as to how whatever interaction that character is experiencing, it impacts the work as a whole. And the first [essay] does not. And because the sentences are so choppy. (Participant T12)

In the aforementioned quotation, Participant T12 delved even deeper into Essay 1 when she compared it to the more well-developed Essay 2.

Participant T5 concluded her assessment with, "I would give it a 50 because we would have covered this by the end of the year for sure. Especially sentence structure. There's no evidence [and] it's repetitious. It doesn't say anything. [The essay] literally doesn't say anything." Her assessment was in direct contrast to the MCAS scoring guide statement that indicated the errors in standard English conventions did not interfere with the understanding of the essay.

In contrast, seven participants determined that the student had addressed the prompt adequately for Essay 1. These participants were as follows: Participant T3, Participant T5, Participant T6, Participant T8, Participant T10, Participant T11, and Participant O1. Participant T3 began with the determination that the student addressed the prompt when she said, "He does go back to the prompt. That was good…so his factor would be society. I mean it's all about

society. He did somewhat develop, but really I feel like this essay was really lacking in support or elaboration." But then she started to contradict herself when she said,

> I feel like he did identify the character, he described how the characters influenced by society, but he didn't do a very good job, again, elaborating or explaining and he didn't really explain so much how it's important to the work as a whole. (Participant T3)

It is important to note that Participant T3 did not say the student failed to address the third part of the prompt regarding the work as a whole. Rather, she said "he didn't do a very good job." Although she appeared to contradict her original statement, this weak contradiction left this researcher with the decision to keep Essay 1 marked with a *yes*.

Participant T5 also agreed that the student addressed the prompt in Essay 1, although rather weakly. She said,

> This one [Essay 1], um, hey a little bit on, uh, to me like a very low level, you know, the scale of very simplistic type answers, the red hair. And so I do think they kind of addressed the prompt. I would give them the points for that. It would just be the high, the lack of evidence. (Participant T5)

In contrast to Participants T3 and T5, Participant T6 determined that the student addressed the prompt rather strongly.

> Okay. So now that I've read the whole thing, I would go back and look at the prompt, um, and I'll see what it's asking for. So, it's asking you in a well-developed composition identify a character. He did that. Describe the character's influence by the person or factor. Okay. It was by society. Got that. How the character's experience is important to the work as a whole. So it looks like I can check off all those boxes. (Participant T6)

Participant T6 used a checklist method of hunting the text for the instances where the student had addressed the three main parts of the prompt and felt the student had done so adequately.

Finally, Participant T11 considered whether the student addressed the prompt on three separate occasions starting with a discussion of the student needing to develop the ideas more. She said,

> I'd like to see more development on how that supports the idea of him having to lie and how he actually lies to his parents or why that is significant. I feel like it's connected, but I'd like to see the student's reasoning behind it. (Participant T11)

Next, Participant T11 reversed her trepidation and mentioned that the student addressed the prompt adequately when she observed, "I think the ideas are there. I think that they put a good effort in and I think that it adheres to the prompt." But then she contradicted herself, albeit slightly, when she said,

> Um, the conclusion, I feel like it doesn't give a lot of connection. It says after he left, everyone in the memory, in the community got their memories back in the ability to feel. I guess I'd like to ask whether or not that one connects back to the prompt or I feel like that might be a little off topic. I feel like it needs a little more closure. So probably just overall more development to show what the student's thinking. (Participant T11)

It is important to note with Participant T11, as with Participant T3, although both participants appeared to contradict themselves, neither one stated definitively that the student did not address the prompt. They both indicated the student had addressed the prompt and then expressed some doubts as to whether that was true.

Additional Analyses

Two additional themes emerged from the qualitative data that were not related to the research questions. The first theme was the relationship between the evaluation focus of the participants and the scores they assigned to Essay 1. The second theme was the relationship between the evaluation focus and the determination participants made regarding whether the student had adequately addressed the writing prompt. The additional analyses are discussed in the following sections.

Organizing the grading data into participant subgroups revealed additional findings. The participants who focused their evaluation on the thesis assigned scores with less variability than the participants who focused their evaluation on the writing formula. Table 18 presents the grading data organized by participant evaluation focus.

Table 18

Essay Scores Organized by Participant Evaluation Focus

Participant	Essay 1	Essay 2
	Evaluation Focus: Formula	
T6	90	100
T7	80	95
T2	73	92
T10	70	90
T1	70	85
T8	65	88
T4	60	95
T5	50	80
T12	50	80
O1	50	90
Range	40	20
	Evaluation Focus: Thesis	
P3	85	98
P4	75	88
O4	75	90
A1	75	75
T11	70	80
P1	70	95
O5	70	94
O3	70	90
T3	65	80
O2	65	88
T9	63	85
Range	22	23

Note. Participants scored essays out of 100 points using the grading procedures of their choice.

Table 18 is organized into two sections. The first section includes the raw scores for Essay 1 and 2 for the participants whose primary evaluation focus was on the writing formula. The second

section includes the raw scores for Essay 1 and 2 for the participants whose primary evaluation focus was on the thesis.

As reported previously, the range for Essay 1 was 40 points for all 21 participants. In contrast, the range differed when the participants were grouped by evaluation focus. The participants who focused on the writing formula represented a range of 40 points for Essay 1, but the participants who focused on the thesis represented a range of 22 points. Moreover, the range for Essay 2 for all participants was 25 points. The participants who focused on the formula represented a range of 20 points for Essay 1, but the participants who focused on the thesis represented a range of 23 points. Therefore, in Essay 1, which was the lower scoring of the two essays because of a lack of thesis and a distracting use of the conventions of standard English, the range was significantly smaller for the participants who focused on the thesis. However, this same result did not appear in the subgroup ranges for Essay 2.

The participants whose evaluation focus was on the writing formula graded Essay 1 with the highest and lowest scores. Their decision-making is highlighted in the comments they made regarding the essay. Participant T6 who evaluated Essay 1 with the highest score of 90 of 100 points said regarding the thesis,

> [H]e identifies a character. He describes how the character's influenced by the story he explained. He explains how the experience is important to the work as a whole. So he does, he crosses off all those boxes. So to me, even though I think the writing could be better, ultimately he did follow the prompt to the 'T'. (Participant T6)

In this preceding quotation by Participant T6, although his focus was on the conventions of standard English, he assessed the thesis in a binary way, using a checklist method. It is

interesting to note that although he stated directly that the essay explained how the experience was important to the work as a whole (the third requirement for the writing prompt), he was the only participant to have believed the essay did so. This may have led Participant T6 to evaluate the essay with the highest score.

Participant T6 continued his assessment of Essay 1 and discussed his grading decision-making when he said, "So my biggest issue that I would, uh, take points off for…would be, um, some of it like the grammar errors, um, the way that they wrote some of these sentences." After this quotation by Participant T6, he began to rewrite the essay for the student on the paper provided during the interview. The rewrites focused completely on sentence structure and not on the thesis which is why this researcher determined his primary evaluation focus was on the writing formula.

In contrast, Participant T5 gave Essay 1 a 50% F. Her evaluation also centered on the writing formula. She said,

> I would give it a 50 because we would have covered all this by the end of the year for sure. Especially sentence structure. It literally doesn't say anything. He just keeps repeating himself. I would even question whether he read at this point, I'd be like, did you read the whole book? Because…it's not based on any evidence. (Participant T5)

Even though Participant T5 indicated that the essay "literally [didn't] say anything", few of her other comments interacted with or discussed the thesis in any way. Her evaluation focus was strictly on the writing formula which the rest of the quotation supported.

Finally, Participant O1 also graded the essay with a 50% F. The reasons she gave were, "[O]verall the grammar structure was not what should be at ninth grade…There's no clear

thesis…It's just, I don't know, like again I think word usage and then the repetition of the stuff. Just made it not a good essay." Similar to Participant T5, she also indicated the lack of a strong thesis, yet this was one of the few comments regarding the thesis throughout her interview.

The second additional analysis necessitated organizing the decision-making data regarding the writing prompt for Essay 1 by the evaluation focus subgroups. The participants who focused their evaluations on the writing formula had the most variability in determining if the prompt had been adequately addressed. Table 19 presents the relevant data.

Table 19

Evaluation Focus and Writing Prompt Decision-Making for Essay 1

Participant	Addressed Prompt	Evaluation Focus
Evaluation Focus: Formula		
T1	No	Formula
T2	No	Formula
T4	No	Formula
T5	Yes	Formula
T6	Yes	Formula
T7	No	Formula
T8	Yes	Formula
T10	Yes	Formula
T12	No	Formula
O1	Yes	Formula
Evaluation Focus: Thesis		
T3	Yes	Thesis
T9	No	Thesis
T11	Yes	Thesis
O2	No	Thesis
O3	No	Thesis
O4	No	Thesis
O5	No	Thesis
P1	No	Thesis
P3	No	Thesis
P4	No	Thesis
A1	No	Thesis

Note. A *Yes* indicates the student addressed the prompt adequately, a *No* indicates the student did not address the prompt adequately.

Table 19 is organized in two sections. The first section contains the participants whose evaluation focus was on the writing formula with the second section containing the participants whose evaluation focus was the thesis. Of the 11 participants who focused their evaluation on

the thesis, 18% of the participants agreed that the prompt had been adequately addressed. However, of the participants who focused their evaluation on the writing formula, their decisions were divided equally with 50% of the participants agreeing the prompt had been adequately addressed and 50% of the participants agreeing that the prompt had not been adequately addressed.

The participants who determined that Essay 1 had adequately addressed the writing prompt did so in such a way that if the essay touched upon all aspects of the prompt, even poorly, credit was given for the prompt being addressed. Participant T8 said, "He does have three reasons…and he has them at the end. He just doesn't have them properly organized so that it looks like the prompt is being addressed. But it is there if you're reading it." Participant O1 indicated the prompt was addressed when she said, "Vaguely…and they barely touch upon it." Furthermore, Participant T3 indicated the prompt was not addressed thoroughly, but she still gave the essay credit for addressing the writing prompt. She said, "I feel like he did identify the character, he described how the character is influenced by society, but he didn't do a very good job, again, elaborating or explaining." Finally, T11 evaluated that the writing prompt had been adequately addressed. He said, "I think the ideas are there. I think that they put a good effort in and I think that it adheres to the prompt."

In contrast, other participants did not believe that Essay 1 had addressed the prompt. Participant A1 stated, "So this is something I didn't see too much of directly addressed was that third part. Um, so I'd say they hit two out of three." Furthermore, Participant O3 also determined the student missed the third portion of the writing prompt when he said,

[T]hey definitely identify the character even if it's in the wrong organization for it. They did tell me that they were influenced by the society. So they're addressing that part.

They didn't talk about how it is important to the work as a whole though. (Participant O3)

The commonality appears to be that each of the participants recognized that the third portion of the prompt (to explain how the character's experience is important to the work as a whole) was not adequately addressed. However, what divided these participants was whether that meant Essay 1 had addressed the prompt adequately. Some believed it did, while the others did not.

Summary

This chapter began with an introduction, which stated the purpose of the study, the research questions and a description of how the study was completed. This was followed by a description of the participants and a presentation of the demographic data. Next, the grading data were presented which showed a range of 40 points for Essay 1, representing letter grades of A-F. The grading data for Essay 2 indicated a range of 25 points representing letter grades A-C. Additionally, a description of the coding process was presented.

The analysis of the qualitative data from the 21 cognitive laboratory interviews was organized around each of the five research questions. For Research Question 1, categorization theory was used as a basis for a priori coding. A *categorizing-classical* code was created to identify each instance a participant exhibited classical categorization. The coding data revealed that participants' decision-making exhibited classical categorization in eight of the 21 participants for a total of 25 instances. For Research Question 2, categorization theory was used as a basis for a priori coding to develop the code *categorizing-prototype*. Prototype

categorization was coded 56 times and was exhibited by all 21 participants. For Research Question 3, categorization theory was used as a basis for a priori coding to develop the code *categorizing-exemplar*. Exemplar categorization was coded 25 times in 10 of the participant interviews.

For Research Question 4, open coding was used to identify similarities and differences in the decision-making processes participants used in grading samples of student writing. Two similarities were identified. The first similarity was participants interrupted their reading of the student writing samples numerous times to ask questions of themselves or the student, consider features of the writing, or to give preliminary evaluations of the writing. Interruptions were coded 557 times in 11 focused categories which were reduced to two themes. Those themes were *Interrupting-Thesis* with 191 codes and *Interrupting-Conventions* with 366 codes. An additional similarity was the participants' strong desire to use a rubric to grade the essays.

Two differences in decision-making processes emerged from the qualitative data. The first difference was the participants focused their interruptions, their evaluative statements, and their overall evaluation focus primarily on either the conventions of standard English or the student's thesis. The second difference was that the participants either made binary decisions regarding the thesis or made decisions regarding the quality of the thesis. Eight participants made binary decisions regarding the thesis; eleven participants made quality decisions regarding the thesis; and two participants did not make any comments or decisions regarding the thesis.

For Research Question 5, there was no evidence the participants exhibited any differences in grading decision-making for either of the essays. Both essays were graded using the same processes and procedures. Yet, the participants did not agree with the assessment of the

essay as stated in the 2018 Massachusetts Comprehensive Assessment System Scoring Guide provided by the Massachusetts Department of Elementary and Secondary Education. The scoring guide indicated the conventions of standard English errors did not interfere with the understanding of the essay. It was the consensus of the participants that the errors in Essay 1 were so great that they did impede the understanding of the essay.

Finally, additional analyses of the scores for Essay 1 and 2 revealed differences in the range of essay scores when grouped by participant evaluation focus. For Essay 1, the range reduced from 40 points to 22 points for the participants who focused on the thesis. However, this difference was not apparent for Essay 2. Furthermore, an analysis of the evaluation focus cross-referenced with participant decision-making regarding the writing prompt revealed a significant difference between the two subgroups.

In Chapter 5, a discussion of the findings of the study is presented. Implications for practice regarding grading student writing in English language arts are discussed, and recommendations for future research are proposed.

CHAPTER 5
SUMMARY, DISCUSSION, AND CONCLUSIONS

Introduction

In the previous chapter, qualitative data from the cognitive laboratory interviews were presented and analyzed as they pertained to the research questions. Chapter 5 includes a summary of the study, a discussion of the findings, and the identification of the emergent theory, The Theory of Disparate Purposes in Writing Assessment. This chapter concludes with implications for practice, recommendations for further research, and concluding thoughts. The purpose of Chapter 5 is to expand upon the research findings regarding the decisions teachers made while grading student writing and to present suggestions for further research.

Summary of the Study

The purpose of this study was to examine the decisions teachers made while grading samples of student writing in English language arts using a cognitive laboratory interview. To investigate the grading decisions of teachers, the following research questions guided the study.

1. In what ways and to what extent do teachers use classical categorization in their grading decision-making process when grading samples of student writing in English language arts?
2. In what ways and to what extent do teachers use prototype categorization in their grading decision-making process when grading samples of student writing in English language arts?

3. In what ways and to what extent do teachers use exemplar categorization in their grading decision-making process when grading samples of student writing in English language arts?
4. To what extent, if any, do similarities or differences exist in teachers' decision-making processes when grading samples of student writing in English language arts?
5. To what extent, if any, do teacher decision-making processes differ when grading student writing samples of high and low performance levels in English language arts?

Research Questions 1 through 3 utilized categorization theory as the theoretical framework. Additionally, grounded theory provided the theoretical framework for Research Questions 4 and 5.

Qualitative data were gathered from 21 cognitive laboratory interviews in which the participants graded two samples of student writing in response to the following writing prompt.

> Often in works of literature, a character is influenced by another person or factor. From a work of literature you have read in or out of school, select a character who is influenced by **one** of the persons or factors listed in the box below.
>
> - a friend
> - a family member
> - a spiritual belief
> - society
>
> In a well-developed composition, identify the character, describe how the character is influenced by the person or factor, and explain how the character's experience is important to the work as a whole. (Massachusetts Department of Elementary and Secondary Education, 2018)

Participants were classroom teachers during the 2019–2020 school year in a large urban public school district in the Southeast United States. A purposive, criterion sampling technique was used to select the participants based on the criteria that they taught at least one section of 9th or 10th grade English language arts.

The cognitive laboratory interviews were recorded, professionally transcribed, and analyzed using the ATLAS.ti quantitative analysis software tool. The interviews were coded using a constant comparative method until major themes emerged from the data.

Discussion of the Findings

The goal of this study was to examine the decisions teachers made while grading two samples of student writing during a cognitive laboratory interview. This section contains a discussion of the findings of the grading data, the five research questions, and the additional analysis.

Grading Data

The grading data obtained from the 21 cognitive laboratory interviews indicated a large variation in the scores and letter grades assigned for Essays 1 and 2. For Essay 1, the scores varied from a low score of 50 points to a high score of 90 points with a range of 40 points. The letter grades assigned to Essay 1 included all five letter grades A through F. For Essay 2, the scores varied from a low score of 75 points to a high score of 100 points with a range of 25 points. The letter grades assigned to Essay 2 included grades A through C.

The findings from the grading data revealed a lack of inter-rater agreement. Inter-rater agreement is defined as "the degree to which a rater assigns scores to a particular set of examinee responses that are consistent with scores assigned to those responses by other raters" (Wolfe et al., 2016, p. 2). More specifically, the lack of inter-rater agreement for Essay 1 was so great that some participants assessed the student writing as exceeding expectations and assigned an A, yet other participants assessed the student writing as a failure to meet minimum expectations and assigned the writing an F.

The data from this research study were consistent with the findings of over a century of inter-rater agreement studies. Edgeworth (1888) through his research first identified a lack of inter-rater agreement and determined the differences in grades teachers assigned to student work fell into three primary areas. These three areas were, (a) chance, (b) differences among the raters, and (c) the rater's error in judgement becomes the estimation of the examinee's actual proficiency (Edgeworth, 1888). In further studies, Starch and Elliot (1912) found a range of 34 and 49 points for student writing, and Ashbaugh (1924) found a variation of 51 points in grading mathematics. Nearly 100 years later, Brimi (2011) replicated Starch and Elliot's study with similar findings resulting in a variation in scores of 47 points for student writing. The findings regarding the grading data for the present study were consistent with the previously mentioned studies. This suggests that a lack of inter-rater agreement is a universal reality of student assessment. This is also consistent with Speck's (1998) argument that teacher professional judgement is central to assessment and when something as complex as student writing is assessed, professional judgement can be unreliable.

The third portion of Edgeworth's (1888) statement regarding the rater's error becoming the estimate of the student's proficiency speaks to the central issue of this study. Depending upon which participant graded Essay 1, students could either move forward, thinking their writing was exceeding expectations or be held back thinking their writing was failing. The feedback the participant could potentially give students could change the trajectory of their writing and overall academics for years to come. That feedback could be "excellent job," or it could be "your essay contains too much summary, lacks a clear thesis, and needs revising." More importantly, the negative feedback accompanying a failing grade could cause the student to be turned off to writing in the future. Because grades are significant indicators of future post-secondary success (Sawyer, 2013) and they are also a better predictor of college success than standardized test scores (Atkinson & Geiser, 2009), the grade and feedback students receive carries great significance for them.

Research Question 1: Classical Categorization

The first research question was, in what ways and to what extent do teachers use classical categorization in their grading decision-making process when grading samples of student writing in English language arts? The results of the study indicated the majority of the participants did not use classical categorization while grading samples of student writing. Twenty-five occurrences of classical categorization were identified within the 21 cognitive laboratory interviews. Of the 21 participants, only eight participants exhibited any classical categorization leaving the remaining 13 participants who did not exhibit classical categorization. This finding was consistent with previous research (Haswell, 1998; Lynne, 2004), which indicated that

although rubrics are classical in nature, most teachers use prototype and exemplar categorization to grade writing.

The participants who used classical categorization to grade the essays searched for an essential list of features necessary for an essay to be considered of high quality. Those items included an introduction, thesis, transitions, evidence/elaboration, and a conclusion, to name a few. These checklist items were assessed in a binary manner meaning that either the student included or did not include the feature in the essay. There was little, if any, assessment as to the quality of the transition or to the quality of the thesis. The mere fact that the essay included a thesis was satisfactory to those participants.

Participant T6 exemplified classical categorization in that he specifically mentioned checking off boxes in his assessment of the essay. He said, "He did [identify a character] … Okay. It was [influenced] by society. Got that. How the character's experience is important to the work as a whole. So, it looks like I can check off all those boxes." In this statement by Participant T6, there was a noticeable lack of assessment for the quality of the features he was searching for in the essay. In contrast, this participant gave the student credit for including those text features, of any quality level, in the essay.

Research Question 2: Prototype Categorization

The second research question was, in what ways and to what extent do teachers use prototype categorization in their grading decision-making process when grading samples of student writing in English language arts? Prototype categorization is a type of decision in which a teacher compares the student's writing to the ideal version, or prototype, of the essay. The

hallmark of a prototype categorization is the judgement of how close or how far the writing sample is from the ideal version to which it is compared. Through a priori coding, prototype categorization was identified 56 times throughout the 21 cognitive laboratory interviews. This type of categorization was present with each of the 21 participants, which was in direct contrast to the data for classical categorization which was present in only eight of the participant interviews.

Participants used different types of prototypes to compare the student writing. Participant T8 used the ideal version of each letter grade as a prototype when she said, "I would say it's like a high B, high B, low A, I mean it's not like an A yet. So, I'm going to go high B because I can't give it an A." Furthermore, Participant O5 used a grade level as a prototype when she stated, "This is clearly a pretty strong essay for a 10th grader." The significance of prototype categorization is that even though the majority of the participants used the state standardized writing rubric in some way (which is a form of classical categorization), they also used prototypes to compare the student work. There appeared to be a tension between using classical and prototype categorization while grading student writing which aligned with Armstrong et al. (1983) in which odd numbers (classical categorization) were judged in a prototypical way when some odd numbers were considered to be "more odd" than others.

These findings suggest that participants used prototypes quite frequently as one of their tools for assessment. Regardless of the evaluation tool (rubric or other method), their use of prototypes to categorize the writing was an important and useful way to determine the grade students should receive for their writing.

Research Question 3: Exemplar Categorization

The third research question was, in what ways and to what extent do teachers use exemplar categorization in their grading decision-making process when grading samples of student writing in English language arts? Exemplar categorization is the third and final type of decision in categorization theory. It is characterized by the teacher comparing the student writing to a set of exemplars teachers have gathered in their collective memory as teachers. Exemplar categorization contrasts with prototype categorization in that teachers have numerous exemplars in their mind in which to compare student writing.

The findings in the present study indicated 12 of the 21 participants used exemplar categorization in grading student writing. Additionally, the findings indicated that participants of all experience levels used this type of categorization. Implicit in the definition of exemplar categorization is enough teaching experience in order to have amassed the exemplars to use during grading (Haswell, 1998). Yet, the findings in the present study did not support that the most experienced participants were more likely to use exemplars. The 12 participants who used this form of categorization had experience that ranged from one to 24 years of experience. Conversely, those participants who did not exhibit any exemplar categorization also had experience from three to 19 years of teaching experience. Therefore, the participants who used exemplars to grade the student writing were gathering the exemplars from various sources outside their classroom or were using very recent examples from inside their classroom if they lacked extensive teaching experience.

Participants used exemplar categorization in different ways. Some used examples of previous student writing, examples of previous students, their experience in teaching a specific

letter grade, or even examples outside their classroom, including important authors or website summaries of novels. It was clear that the majority of participants used their teaching experience in knowing what a good essay was, what a bad essay was, or what to expect from a student in that grade level. It is interesting, however, to note that not all participants called upon their experience when grading the essays.

Examining all three categorization types together, the findings also support that each type did not have clear and distinct boundaries. Five participants used all three types of categorization, and 11 participants used two types of categorization, with the remaining five participants using only prototype categorization. These findings further aligned with the literature in that although rubrics, by definition, are classical categorization in nature (Haswell, 1998; Lynne, 2004), the majority of participants used either prototype categorization (all 21 participants) or exemplar categorization (12 participants) when grading samples of student writing. This was likely due to the unique qualities of each essay the participants graded and is consistent with the findings of Broad (2000) who determined "weird cases" are difficult to compare to a limited number of exemplars.

Research Question 4: Similarities and Differences in Grading

The fourth research question was, to what extent, if any, do similarities or differences exist in teachers' decision-making processes when grading samples of student writing in English language arts? To answer Research Question 4, this researcher applied open and focused coding until themes began to emerge from the data. The emergent themes are discussed in two sections: (a) similarities and (b) differences.

Similarities

Similarities were that all participants interrupted their reading of the student essays to consider features of the essay, asked questions of the imaginary student or themselves, and made general comments about the essay. An additional similarity was the expressed desire of several participants to use the state standardized rubric to grade the essays even though they were given instructions to use whatever grading method they would normally use in their classroom.

Interruptions

Participants interrupted their reading of the student writing 557 times during the 21 interviews. This is not surprising and perhaps expected that a participant would stop at various points in the essay to notice a grammatical error, to take note of the thesis, to suggest a better transition, or to be dismayed at the lack of evidence and elaboration. That would appear to be a natural aspect to grading student writing. This finding, however, is not consistent with that of Huot (1993) who found that inexperienced raters tended to start and stop during the rating process and that experienced raters were less likely to do so. In this study, each of the 21 raters interrupted their reading of both essays with similar recurrence. It is noteworthy that none of the participants read either essay without interruption, giving the entire essay their full consideration when determining the quality and assigning a grade. Rather, each participant stopped at numerous points and made comments and mini-assessments of the student's work.

The interruptions participants made included general remarks regarding the writing, questions they asked themselves, and, most interesting, participants would interact with the imaginary student as if the student were there with them in the room. This finding reflects the

findings of Kalthoff (2013) and Vaughn (1991) who found that raters interacted with the imaginary student as if they were in the room at the time of grading. Illustrating this concept, Participant P1 stated, "This needs more development so that you understand what society's explanation is. Our society's influence is on the line. Why does he have to lie? Um, this just says that he does it doesn't say why he has to." In this quotation, Participant P1 addressed the student directly as if he was there next to the participant to hear his feedback about the essay, which needed more development. Participant P1 further asked the student directly "Why does he have to lie?" These statements by Participant P1 illustrated the process of the participant interacting with the text and the student. The participant did not merely read the essay and assess it. On the contrary, he interacted with the essay and the student and became directly involved with the text. This suggests that participants were not necessarily grading the writing, but may have been grading the student, consistent with Wyatt-Smith and Castleton's (2005) similar results in their Australian study.

Another type of interaction was a direct interaction with the text. This occurred when Participant O4 discussed the lack of clarity regarding the direction of the student's essay. The following statement indicated the interruption took place to assist the participant to understand and make sense of what was written. O4 said, "This doesn't make any sense to me. I'm not understanding what this has to do with *The Giver*, the man himself. So, this seems out of place. Not, well…If it's relevant, it's not very well explained." After Participant O4 finished reading the essay, he confirmed his fears expressed in his previous statement regarding the thesis. He said,

> Yeah. Um, this could be better, a lot better in my opinion. Um, there's not really a like a clear thesis. Uh, the prompt asks for the character to be developed by one person and he talks about the giver, but then he goes on to talk about everyone else. So there's not a very clear, it's not a clear thesis, which is very important to me. (O4)

It is important to note the connection between these two statements by Participant O4. In the middle of the essay, he was concerned about the lack of a clear thesis, interrupted his reading to consider that concern, and then confirmed the concern at the conclusion when he gave his final evaluation. This indicated there was a definite purpose to the interruptions and that they in fact assisted the participant in processing and analyzing the student writing so that a more thorough evaluation could be given.

Need a Rubric

The second similarity that emerged from the data was the desire of many of the participants to use the state standardized rubric to evaluate student writing. During the interviews, participants were given instructions to use any grading process they typically used in their own classroom. The majority of the participants brought no materials with them, yet referenced the state standardized rubric from memory. Several participants echoed Participant T3 when she indicated she did not feel comfortable without a rubric. She said, "I wouldn't really feel comfortable just giving him a grade. I don't think that that's good teaching. I would definitely want a rubric and…I would teach the students, uh, how to use the rubric." Her statement went beyond the need for a tool to aid in evaluation. She did not believe that good teaching could occur without a rubric.

The findings in this study were consistent with those of Hunter and Docherty (2011) who found that raters used rubrics inconsistently when assessing the same sample of student writing. Even when the participants in this study used the state standardized rubric from memory, the scores that resulted for Essay 1 had a range of 50 points and a range of 25 points for Essay 2. The unreliability of professional judgement (Speck, 1998) and the inconsistent application of rubrics (Hunter & Docherty, 2011) together created an assessment environment in which inter-rater agreement was nearly impossible.

Differences

For the differences in grading decision-making, the participants focused their interruptions, evaluative statements, and overall evaluation focus on either the conventions of standard English or the thesis. Furthermore, of the evaluative decisions regarding the thesis, the participants either acknowledged the existence of a thesis in a binary fashion or focused on the quality and completeness of the thesis. An additional difference was the participants' lack of agreement regarding if the students had adequately addressed the writing prompt for Essay 1. The differences in grading decision-making are discussed in the following three sections.

Focus: Conventions or Thesis

As the participants read the essays, interrupted their reading, and then evaluated the essay after the reading was concluded, a distinct theme emerged. The interruptions, evaluative statements, and overall evaluation focus of each participant fell into one of two categories. Either the participant focused on the conventions of standard English or they focused on the

thesis. This did not mean that the participant made only statements regarding the thesis or made only statements regarding conventions. Almost all 21 participants verbalized statements about both. It is noteworthy that Participants T1, T2, and T7 did not interrupt their reading once to consider the thesis. Moreover, Participant T2 did not make any evaluative statements regarding the thesis. Participant T2 did not verbally discuss the thesis and focused primarily on the conventions when evaluating both essays.

This finding is consistent with previous research by Stern and Solomon (2006) who found that raters tended to focus on the conventions of standard English because those features were easy to identify and score. The participants in this study did not need to use a rubric, but many chose to use the state standardized rubric as their method of assessment. Rubrics have been found to allow raters to focus more on the argument and substance of an essay (Rezaei & Lovorn, 2010). The findings in this study were not consistent with those of Stern and Lovorn because the participants who used the rubric focused more closely on the writing formula rather than the thesis.

As previously discussed, participants often interacted with the text and imaginary students as though they were in the room (Kalthoff, 2013; Vaughn, 1991). It became clear in the 21 interviews that participants either interacted with the text regarding the conventions or they interacted with the text regarding the thesis. These findings indicated a distinct difference in the way in which participants graded the essays and a fundamental difference in expectations of the participants. Participants either expected the student to demonstrate conventions of standard English, or to demonstrate a well-developed thesis that was supported by evidence and

elaboration. As Participant O5 stated succinctly, "There has to be a clear point and there has to be some level of clean writing and sophisticated writing and sometimes that's easily fixed."

Unlike Vaughn (1991) who found a third type of rater who had a dual focus on mechanics and content, the findings in this study indicated only the singluar focus expressed by Stern and Soloman (2006). Although the participants took both aspects into consideration, these findings were consistent with those of Vaughn (1991), suggesting that there was a primary focus on the conventions or the thesis, but not both.

Binary or Quality

When participants discussed the thesis, they did so in either a binary fashion as though the thesis was another necessary text feature to include in the essay or they discussed the thesis regarding its quality and completeness. A more focused thesis with support and elaboration was considered high quality, and an unfocused thesis that lacked clarity was considered a low quality thesis. There is a connection between three of the emergent themes that is deeply related and important to understand. That connection is between (a) classical categorization, (b) the participant's focus on conventions, and (c) the evaluation of the thesis as binary decision. All three point toward assessment decision-making as a checklist of items that the student needs to include in the essay. This checklist includes items such as a thesis, transitions, grammar, topic sentences, etc., which were previously mentioned as the formula (Participant O2). Participant O2 expressed this concept succinctly when he stated, "…they've got the formula down, they have a beginning, middle and end." The important word in that statement is *formula*. The

formula refers to how to use the conventions of standard English to create an essay so that it will convey sophisticated writing.

The findings in this study are consistent with previous research that teachers often over-focus on the conventions of standard English because they are easily corrected (Stern & Solomon, 2006). This was clear from the numerous interruptions by the participants who interrupted their reading 366 times during 21 interviews to notice and comment on the mistakes students made with conventions. Moreover, Brimi (2011) found that some teachers were perhaps incapable of grading beyond the five-paragraph essay. This is perhaps true for the 10 participants who focused on the formula and even more true for Participants T1, T2, and T7 who did not interrupt themselves once to consider the thesis.

However, it may be difficult to judge participants' decision-making processes solely on their verbalizations during the cognitive laboratory interview. This is due to a limitation of this study in that participants may have revealed only what they were willing to reveal during the interview (Alshenqeeti, 2014). Although the participants readily verbalized their thoughts during the interview, it is quite possible that there was much more that was not verbalized or that was subconscious and not revealed during the interview. Similarly, it is difficult to infer from the quantitative analysis of the codes that participants' decision-making was focused on the conventions of standard English just because they verbalized more often about the conventions. This is consistent with the limitation of the study noted by Diederich et al. (1961). The quantitative analysis of the comments written on the essays in that study did not necessarily indicate which aspects were more important in the grading of the essays. This is due to the

greater opportunities in commenting, in verbal or written form, on the conventions of standard English than the argument the writer is attempting to make.

Research Question 5: Differences in Grading High and Low Essays

The fifth research question was, to what extent, if any, do teachers' decision-making processes differ when grading student writing samples of high and low performance levels in English language arts? Although the decision-making processes were relatively identical for the high and low scoring essays, there was a difference that manifested itself in the grading of Essay 1. Of the 21 participants, 14 did not believe the prompt had been adequately addressed as did the remaining seven participants who felt the writing prompt had been adequately addressed.

The 2018 Massachusetts Comprehensive Assessment System Scoring Guide provided a rationale for Essay 1 that indicated in part, "Despite some minor errors in grammar (e.g., "If Jonas told anyone, him and the people he told…") and other aspects of conventions, their presence does not interfere with the reader's understanding of the composition" (Massachusetts Department of Elementary and Secondary Education, 2018). The findings of this study were not consistent with the scoring guide from the Massachusetts DESE. The evidence of this discrepancy is the lack of consensus as to the writing prompt being adequately addressed. Due to the number of participant interruptions regarding the conventions of standard English and the lack of consensus about the writing prompt, it does appear that the errors in grammar interfered with the comprehension of the essay. Because the errors in conventions were so great as to impede the understanding of the essay, it is likely a contributing factor to the discrepant assessment of the adequate answer to the writing prompt.

It is a possibility that the findings in this study suggest a reason why some participants focus on the conventions of standard English when they evaluate student writing. Without at least a baseline of adequate conventions, it is difficult for the reader to understand the content of the essay. In contrast, the evidence from this study did not indicate that the raters changed their focus from one essay to the other. Specifically, their focus was consistently on either the conventions or the thesis for both essays. Although the decision-making processes were similar, the number of errors in the conventions of standard English appear to have disrupted the grading processes of the raters, interfering with their ability to create an accurate assessment of the student's thesis in Essay 1.

Additional Analyses

The additional analyses revealed differences in the scores participants gave the essays based upon their evaluation focus. The qualitative data revealed that the participants assessed each essay with a primary focus either on the thesis or on the conventions of standard English, also described as the writing formula by Participant O2. These differences in scores led to less variability for the participants who focused on the thesis. The variability was 22 points compared to 40 points for the variability of all participants. There are several potential explanations for this difference. However, given the qualitative research design using 21 participants in the present study, it was not possible to generalize beyond the subjects being studied (Fraenkel et al., 2015).

A potential explanation is that participants whose evaluation focus was on the thesis exhibited more inter-rater agreement than the participants who focused their evaluation on the

writing formula. Therefore, it might be concluded that an evaluation focus on the thesis could be a more reliable way to grade essays. However, this difference in scoring did not appear in the scoring data for Essay 2; thus, that explanation was not supported in the data. Alternatively, considering Essay 1 was the lower performing essay that included many errors in the conventions of standard English, it is possible the participants who focused on the writing formula were distracted by the numerous errors. Therefore, they may have had difficulty comprehending the content of Essay 1, which resulted in three participants (T5, T12, and O1) who scored the essay with a 50% F. Of the participants who focused their evaluations on the thesis, the lowest score any participant gave was a 63% D. Given the scoring and qualitative data as reported by evaluation focus subgroup, it is just not possible to provide a reliable explanation. This suggests this topic is an area in need of further exploration.

The second theme to emerge in the additional analyses was the variation by evaluation focus subgroup regarding the determination whether Essay 1 had adequately addressed the writing prompt. The participants whose evaluation focus was on the thesis reached near consensus that Essay 1 did not address the writing prompt adequately. In contrast, the participants whose evaluation focus was on the writing conventions were evenly split between the writing prompt being adequately answered or not adequately answered. This also suggests this is an area for further research as the data did not support a definitive conclusion.

Emergent Theory: The Theory of Disparate Purposes of Writing Assessment

A distinct interrelationship exists between several of the findings that point towards a difference in expectations teachers have regarding the purpose of student writing. Classical

categorization suggests that teachers analyze student writing as a set of specific attributes that need to be present in order to be graded with a high score. Furthermore, the evaluation focus on the conventions of standard English and the evaluation of the thesis in a binary manner comingle to form a formula grading model that resembles a checklist of items. These items on the virtual checklist are not evaluated for their quality. Rather, they are catalogued for their mere presence. If the essay contained a thesis, even if it was weak, the student was given credit for stating a thesis. Furthermore, if the essay had topic sentences, transitions, evidence, and a conclusion, then the student completed a well-written essay according to the teachers who follow the formula grading model.

The formula grading model is in stark contrast to the teachers who interacted with the thesis, evaluated it for its completeness, strength, and evidence, and focused on whether the student had something important to express in their writing. It is clear that these two sets of teachers have vastly different expectations for the purpose of the writing task. For the formula grading model teachers, the purpose of writing is to demonstrate the conventions of standard English. They see the purpose is for the student to demonstrate the writing of a five-paragraph essay and show that they can follow the proper writing formula. For the thesis grading model teachers, the purpose of the writing task is to make a claim and support the claim with evidence. Furthermore, to these teachers the claim needs to be strong, important, debatable, and supported with evidence that effectively strengthens the claim.

The theory of disparate purposes of writing assessment reaches to the root of the discrepant scores. The possible reason that scores have been discrepant for more than 100 years is perhaps because educators have been attempting to mitigate the symptom, i.e., the scores, with

rubrics, training, grading reform and the like. This theory speaks to a root cause that can be addressed which may in turn lessen the lack of inter-rater agreement that has plagued education for so long.

Implications for Practice

Potential implications for practice emerged at the conclusion of the analysis of the data. These implications are discussed for three populations: high school teachers, school district leaders, and teacher preparation programs.

High School Teachers

Based upon the findings of this study, high school teachers should reflect upon their grading practices and begin discussing grading of student writing in their professional learning communities (PLC). During PLCs, teachers should include common grading time to their common planning time and discussions regarding data. This common grading time could be to discuss specific student samples of writing, the grades that teachers have assigned, and the rationales for those grades. Furthermore, teachers could include their own cognitive laboratory interview session in which a teacher would read aloud a student essay including their thoughts regarding the features of the student work. The other teachers within the PLC could observe the cognitive laboratory interview and think of their own assessment of the work. At the conclusion, the teachers in the PLC could discuss the grading and rationale to develop their own assessment skills further. Additionally, teachers could discuss their overall evaluation focus and their view of the purpose of writing task to understand who believes the purpose is to demonstrate the five

paragraph essay, i.e., the formula and who believes the purpose is to create a claim and say something important through their writing.

School District Leaders

School district leaders may want to consider focusing professional development endeavors on decision-making processes teachers use while grading student writing. Grading is often a solitary activity that is rarely discussed among educators. The findings in this study regarding the lack of agreement as to whether or not a student adequately addressed the writing prompt, could be useful in designing professional development. These trainings could include not only how to use the state standardized rubric for student writing but could also include training and discussion on what it means to adequately address a prompt and how to use a rubric when the prompt has not been addressed. Additionally, professional development could be designed regarding a balanced approach between the conventions of standard English and the thesis, unlike the findings in this study whereby participants held disparate purposes for writing assessment. Finally, professional development could be developed to help teachers avoid binary decisions regarding the thesis and the conventions of standard English so that they consider the quality of the student writing and not reduce grading to just a mere formula to follow.

Research over the last 100 years has shown a lack of inter-rater agreement that has remained unchanged, including the findings in this study. However, the efforts to improve inter-rater agreement have largely been unsuccessful. School district administrators could instead focus on grading decision-making processes, a balanced approach between the conventions and

thesis, and quality. It is possible that a focus on decision-making processes rather than on inter-rater agreement as an end result, could result in more reliable scores.

Teacher Preparation Programs

Teacher preparation programs may benefit from this research by including more robust discussions regarding grading decision-making of student writing. Prospective teachers need to understand how to balance the conventions of standard English with the thesis, and to also focus on the quality of the aforementioned rather than using the rubric as a mere checklist of text features the student included in the essay. Furthermore, prospective teachers could benefit from collaborative grading training along with the collaborative planning that is already present in many preparation programs. If teachers are to arrive in modern classrooms ready to manage, teach, and assess, they need to be adequately trained in how to assess student writing for quality.

Recommendations for Further Research

Following are suggestions for future research regarding the grading decisions teachers use in grading student work.

1. Future research could be conducted to determine the similarities and differences in teacher grading decisions while using the state standardized rubric for student writing.
2. Future research could be conducted to determine the similarities and differences in teacher grading decisions of contextualized writing samples vs. decontextualized writing samples.

3. Future research could be conducted to determine the relationship, if any, between decision-making patterns, teacher demographics, and school demographics.
4. Future research could be conducted to investigate the grading decisions teachers make regarding assignments other than student writing, such as tests, quizzes, and homework, in English language arts.
5. Future research could be conducted to study the use of cognitive laboratory interviews as a professional development tool.
6. Further research could be conducted to determine the similarities and differences in evaluation focus and scoring data.
7. Future research could be conducted to determine the similarities and differences in teacher grading decisions in other content areas such as mathematics, science, or social studies.
8. Future research could be conducted to determine the relationship, if any, between the grading of the correct answer in mathematics vs. grading the mathematical process and student achievement in mathematics.

The cognitive laboratory interview proved to be an invaluable tool in this research study. Although it is time-consuming and generates copious amounts of data necessitating extensive analysis, this method lends itself to teacher decision-making studies regarding grading and other educational contexts. Future research in education should include this method.

Conclusions

Inter-rater agreement in grading student work has been a concern for researchers for over 130 years. Starch and Elliot (1912) first found that teachers graded student work at such a high variability, that grades A-F were often given for the same student sample. Research since then has yielded similar results (Brimi, 2011). Despite all the advancements in educational practices, however, the inter-rater agreement has remained unchanged.

The purpose of this study was to investigate the decisions teachers made while grading samples of decontextualized student writing in English language arts. The findings expanded upon the work of previous researchers and indicated there are multiple similarities and differences in the decisions teachers make while grading student writing. This study revealed that participants interrupted their reading of student work to consider the conventions of standard English or the thesis, to ask themselves questions about the writing, or to ask an imaginary student about the writing as if he or she were in the room. The differences were that participants exhibited an overall evaluation focus on the conventions of standard English or on the thesis but not both. Participants were either concerned with how the student wrote the essay or what the student was trying to express through writing.

Furthermore, this study revealed new insights that participants either made binary decisions regarding the thesis and conventions, or made quality decisions about the thesis and conventions. The participants who made binary decisions used a virtual checklist in which items were to be checked off as the participant read the essay. In contrast, those participants who focused on the quality of the thesis challenged the student to make the thesis clear and to make sure the evidence and elaboration, along with the conventions, supported the thesis effectively.

Additional analyses revealed that participants whose evaluation focus was on the thesis demonstrated more inter-rater agreement and also reached consensus regarding the writing prompt for Essay 1. These analyses indicated that when the conventions of standard English interfered with the understanding of the work as a whole, there was less inter-rater agreement.

The theory of disparate purposes of writing assessment emerged from the interrelationships between classical categorization, teachers' evaluation focus on the conventions of standard English, and the binary decision-making regarding the thesis. Teachers approach the grading of student writing from two perspectives. One perspective is that the purpose of student writing is to demonstrate the five-paragraph essay, i.e., the formula, as well as the conventions of standard English. The second perspective is that the purpose of student writing is to express something important by making a claim and supporting it with evidence. The theory of disparate purposes of writing assessment points towards the root cause of discrepant scores, rather than trying to treat the symptom of the problem which is the scores themselves.

The essential question in the study was to determine the decision-making processes that led to the varying scores for samples of student writing in English language arts. After analyzing the qualitative data and answering the research questions, the findings suggest many factors that lead to such varied scores for the same essay. Each participant used multiple grading models to assess the sample of student writing. They used a rubric; they compared the writing to prototypes and exemplars; they graded for the formula or the thesis; and they used a 100-point grading scale with five letter grades. Each of these models has value and each of the participants used a mixture of some or all of the models in their grading decision-making. Each participant had differing experience, values, and foci in relation to writing assessment. Vaughn (1991)

found similar results in that raters focused on different aspects of writing, regardless of the training received. Given the various tools at a teacher's disposal and the varying ways in which they can be used and combined, it is no wonder that the grades are just as varied.

The findings of Starch (1913) and Brimi (2011) were consistent with the findings in this study. Starch argued that variability existed because of four major factors, two of which are the value different teachers placed on certain features of an essay and the inability of teachers to distinguish "between closely allied degrees of merit" (Starch, 1913, p. 630). The participants in this study were unable to agree on the merits of the essay when either using their own method of grading or using a rubric. Furthermore, they were unable to agree if the writing prompt had been adequately addressed in Essay 1. It is possible that essays with significant errors in the conventions of standard English were more difficult to reliably grade.

This conclusion is consistent with Huot's (1996) argument that a theoretical basis for writing assessment is lacking because the focus has been on developing procedures that ensure inter-rater agreement. The procedures present in this study were to use the state standardized writing rubric even when participants were given the opportunity to use their own grading methods. However, research over the last 130 years has indicated that the various processes and procedures have not increased inter-rater agreement (Ashbaugh, 1924; Brimi, 2011; Eells, 1930; Starch & Elliot, 1912). Huot further stated that for a writing assessment to be valid, it must include a theoretical foundation. The findings of this study align with Huot's premise in that without a theoretical foundation, the varying mixture of grading processes the participants used are just a collection of procedures based upon the whim of the participant.

Experienced and future educators could benefit from the findings of this study by thoroughly examining their own grading practices to ensure student work is given a balanced assessment. The conclusions of Starch and Elliot in 1912 were as true today as they were then. "The promotion or retardation of a pupil depends to a considerable extent upon the subjective estimate of his teacher" (Starch & Elliot, 1912, p. 454). Indeed, which teacher is grading the student work has a significant effect upon the success of the student.

APPENDIX A
MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM 2018
GRADE 10 COMPOSITION RUBRIC

2018 Grade 10 English language arts
Composition Rubric

Student compositions that do not address the writing prompt will be deemed non-scorable (NS), earning them a 0 out of a possible 12 points for topic development and 0 out of a possible 8 points for standard English conventions.

Scoring Guide for Topic/Idea Development

Score	Description
6	• Rich topic/idea development • Careful and/or subtle organization • Effective/rich use of language
5	• Full topic/idea development • Logical organization • Strong details • Appropriate use of language
4	• Moderate topic/idea development and organization • Adequate, relevant details • Some variety in language
3	• Rudimentary topic/idea development and/or organization • Basic supporting details • Simplistic language
2	• Limited or weak topic/idea development, organization, and/or details • Limited awareness of audience and/or task
1	• Little topic/idea development, organization, and/or details • Little or no awareness of audience and/or task

Scoring Guide for Standard English Conventions

Score	Description
4	• Control of sentence structure, grammar and usage, and mechanics (length and complexity of essay provide opportunity for student to show control of standard English conventions)
3	• Errors do not interfere with communication and/or • Few errors relative to length of essay or complexity of sentence structure, grammar and usage, and mechanics
2	• Errors interfere somewhat with communication and/or • Too many errors relative to the length of the essay or complexity of sentence structure, grammar and usage, and mechanics
1	• Errors seriously interfere with communication AND • Little control of sentence structure, grammar and usage, and mechanics

APPENDIX B
2018 MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM
STUDENT WRITING SAMPLE – GRADE 10 ENGLISH LANGUAGE ARTS
STANDARD ENGLISH CONVENTIONS – SAMPLE 1 – SCORE 3

Student Response 1

In the novella, The Giver, Jonas, the protagonist, was a young boy. Jonas lived in a small utopia. He had everything going for him. Jonas had two best friends, John and Mary. They were always curious about what was past their community. Little did they know it was a whole different world. Jonas community chose the best job for everyone by monitoring them since they were babies. Jonas got a job as the memory holder of the community. There Jonas met an old wise man everyone called the Giver. Jonas receives memories of how life really is, as how to feel, see color, war, hate, pain and much more. Jonas is influenced by society by lying, not being able to love and living a fake life.

Jonas is influenced by society by having to lie. While Jonas explores more about the memories, he is forced to keep it all a secret. Jonas is forced to lie to his family and best friends about his job. Jonas' parents notice something is bothering him, yet Jonas can't say anything.

Jonas' society influenced him to not be able to love. Jonas and Mary are best friends, nothing could seperate them. Jonas gets a memory of the color red. Mary's hair is red. Every time Jonas sees Mary he can see her true self. He then grows to love Mary. Jonas' community doesn't allow them to be in love. He is then forced to lose his feelings for her.

Finally, Jonas is influenced by society by living a fake life. His community is a utopia, so they shut out all the emotion, evil, fun out of everyones life. Therefor the Giver and Jonas are the only two people who can really feel emotions. Jonas has to live a fake life pretending to be happy when he isn't. If Jonas told anyone, him and the people he told were to be put to sleep.

In conclusion, Jonas couldnt keep it all a secret, so he left. Jonas ran away into the great behond. Once Jonas reached the great behind and left the community, everyone got the memories back and ability to feel. Jonas was forced to lie, not feel love, and live a fake life all because of the influence his society had on him.

APPENDIX C
2018 MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM – SCORING GUIDE FOR STANDARD ENGLISH CONVENTIONS – SAMPLE 1 – SCORE 3

Scoring Guide for Standard English Conventions – Sample 1 – Score 3

- Errors do not interfere with communication and/or
- Few errors relative to the length of the essay or complexity of sentence structure, grammar and usage, and mechanics

The composition is repetitive in structure, with many sentences beginning with "Jonas is/Jonas was." Complexity is attempted but introduces some awkward phrasing and minor errors in punctuation: "Jonas' society influenced him to not be able to love. Jonas and Mary are best friends, nothing could separate them." Spelling is accurate but not particularly complex (e.g., "monitoring," "community," "emotion"). Despite some minor errors in grammar (e.g., "If Jonas told anyone, him and the people he told…") and other aspects of conventions, their presence does not interfere with the reader's understanding of the composition.

APPENDIX D
2018 MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM
STUDENT WRITING SAMPLE – GRADE 10 ENGLISH LANGUAGE ARTS
TOPIC/IDEA DEVELOPMENT – SAMPLE 2 – SCORE 5

Student Response 2

Family has the connotation of love, safety and trust. These overwhelming feelings of familiarity and unselfish loyalty make it easy to become attached to these important people and think of them as one of the greatest aspects to life however this makes it even more devastating when loss of a family member occurs. Holden Caulfield, the protagonist in J.D. Sallinger's The Catcher in the Rye deals with these emotions of love and loss. He often mentions that Allie is the best person he has ever met. When Allie dies of leukemia, Holden feels angry, sad and lonely, all of which are normal emotions during the grieving process. However, Holden is incapable of handling these emotions properly and faces difficulty when establishing new relationships, growing up and moving on. The influence of Allie's death causes Holden to be incapable of forming any long-term relationships or connections with people and places. Right from the start of the novel this is made clear when Holden is kicked out of Pency Prep. This is not his first time dealing with being expelled, proving that he is unable to attach himself to any one school. Similar to not being able to attatch to one place, Holden can not attatch to other people either. Though it is clear Holden does want to establish relationships because he always initiates the interaction, he can never fully commit himself. For example, when Holden leaves Pencey and is staying in Manhattan he gives Sally, a girl he used to date, a call. By calling her it shows his attempt at reaching out to someone, but Holden still finds a way to shut her out. After skating, Holden throws out the idea that they should run away together and when Sally denies it, he rashly calls her a vulgar name, spoiling their entire day together. This shows that whenever Holden begins to grow close to someone he creates a reason for the connection to break. Furthermore, on night, when Holden is feeling especially lonesome he orders for a prostitute to come to his hotel room. This act shows his desperation for love and affection. However, as soon as she arrives he immediately changes his mind, proving that as soon as a bond (even if it is superficial), is attainable he runs away from it. After losing Allie it is clear that Holden fears getting close to people again. Allie's death has great influence on Holden, forcing him to isolate himself because he wants to avoid the feeling of pain and loss again.

Allie also influences Holden because his death causes Holden to hate change and therefore resist growing up. Allie's death came as a traumatic change for Holden and because of this he can not find comfort in any type of change. Infact, one of Holden's favorite places is the Natural Museum of History because no matter when he goes there, nothing ever changes. Similar to this, Holden is very intrigued by the question "Where do the ducks go when the pond freezes over?". It seems that he likes this inquiry because it is one of the few things he can think of that is only a temporary change. Holden likes the idea of this because unlike Allie's death, this change is not permanent. Along the same lines of hating change, Holden also resists the idea of having to grow up. Allie, being just a kid, symbolized innocence to Holden. Meanwhile, whenever Holden looks at an adult all he can see is a phony. He looks at all adults as hypocrites and liars, pretending to be people that they are not. This "phoniness" associated with adulthood makes it extremely undesirable for Holden to want to grow up. The influence of Allie's innocence makes it so that Holden does not want to change and grow up to be a "phony adult."

Allie's influence on Holden is extremely important to the novel as a whole. The relevance of the novel's title is learned when Phoebe, Holden's little sister, asks him what he wants to do with his life. Holden replies that he wishes he could be the "Catcher in the Rye." This job entails catching the little kids before they fall off the cliff that the field of rye is on. This represents Holden's desire to help kids keep their innocence and "catch" them before they "fall off the cliff" into adulthood. This is all in response to Allie's death because Allie showed Holden the beauty off a child's innocence. Holden feels responsible and blames himself for not being able to "catch" Allie even though there was nothing he could do to save him from cancer. Also, more generally speaking, the entire story is the result of Holden being unstable after losing his brother. Holden's actions are very clearly him acting out solely because he does not know how to cope with such a big loss. Holden's experience learning to live without Allie is what the entire novel is about, which clearly is important to the work as a whole.

Highly influenced by Allie's death, Holden is left with a series of mental battles that he must fight. This loss influences almost every aspect of his life, from self-sabotage during relationships to avoid pain, all the way to hating every moment of having to grow up because it means the loss of innocence. Holden is lucky to have had such a great bond with his brother. Holden's broken heart shows how much he's been loved as well as the fact that despite his current state, he is in fact capable of loving someone else. Even though Holden faces many struggles, the pain he feels shows the power of Allie's influence.

APPENDIX E
2018 MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM – SCORING GUIDE FOR TOPIC/IDEA DEVELOPMENT – SAMPLE 2 – SCORE 5

Scoring Guide for Topic/Idea Development – Sample 2 – Score 5

- Full topic/idea development
- Logical organization
- Strong details
- Appropriate use of language

This fully developed composition is logically organized around Holden's experience after his brother's death in *The Catcher in the Rye*. A variety of strong details are included to analyze how the "death of his little brother Allie has a great influence on…his [Holden's] everyday behavior." Holden's expulsion from school supports the writer's assertion that Holden resists "connections with people and places," leading to Holden's subsequent rejection of relationships with Sally and others because he "fears getting close to people again" and desires the safety being alone can offer. Beyond disrupting his ability to engage in the more functional aspects of relationships, Allie's death is identified as that which is preventing Holden from adapting to change. Because "in Holden's eyes, Allie was perfect," Holden fears the change of growing into an adult and sees adults as phony. Holden's relationship with Allie enables him to see "the beauty of a child's innocence," but he feels a great deal of guilt and "blames himself for not being able to 'catch' Allie[,] even though there was nothing he could do to save him from cancer." There is an appropriate, rather than rich, use of language about Holden's feelings ("a series of mental battles that he must fight" and "the pain he feels shows the power of Allie's influence"), and the details support a full analysis of how Allie's death affected Holden.

APPENDIX F
UNIVERSITY OF CENTRAL FLORIDA
INSTITUTIONAL REVIEW BOARD APPROVAL

UNIVERSITY OF CENTRAL FLORIDA

Institutional Review Board
FWA00000351
IRB00001138
Office of Research
12201 Research Parkway
Orlando, FL 32826-3246

EXEMPTION DETERMINATION

June 18, 2019

Dear Guy Swenson:

On 6/18/2019, the IRB determined the following submission to be human subjects research that is exempt from regulation:

Type of Review:	Initial Study, Category 3
Title:	An Analysis of Teacher Decision-Making in Grading 10th Grade Student Writing
Investigator:	Guy Swenson
IRB ID:	STUDY00000627
Funding:	None
Grant ID:	None

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made, and there are questions about whether these changes affect the exempt status of the human research, please contact the IRB. When you have completed your research, please submit a Study Closure request so that IRB records will be accurate.

If you have any questions, please contact the UCF IRB at 407-823-2901 or irb@ucf.edu. Please include your project title and IRB number in all correspondence with this office.

Sincerely,

Racine Jacques, Ph.D.
Designated Reviewer

APPENDIX G
SCHOOL DISTRICT NOTICE OF APPROVAL TO CONDUCT RESEARCH

Research and Evaluation
Orange County Public Schools

OCPS Application to Conduct Research
Research Notice of Approval

Approval Date: 7/30/2019 Study ID Number: 241
Expiration Date: 7/29/2020
Project Title: *An Analysis of Teacher Decision-Making in Grading 10th Grade Student Writing in English Language Arts*

Requester: Guy Swenson
Sponsoring Agency/Organization/Institutional Affiliation: Orange County Public Schools

Thank you for your request to conduct research in Orange County Public Schools. We have reviewed and approved your application. This *Research Notice of Approval (R-NOA)* expires one year after issue date, 7/29/2020.

Additionally, we have received principal approvals from the following schools to participate in your study:

- Apopka High, Principal Lyle Heinz, lyle.heinz@ocps.net
- Olympia High, Principal Guy Swenson, guy.swenson@ocps.net
- Timber Creek High, Principal Kelly Paduano, kelly.paduano@ocps.net

If you are interacting with OCPS staff or students, you may email the school-based or district-based administrators who have indicated interest in participating, including this notice as an attachment. After initial contact with applicable administrators, you may email any necessary staff included in your application. This approval notice does not obligate administrators, teachers, students, or families of students to participate in your research study/project; participation is entirely voluntary.

OCPS badges are required to enter any OCPS campus or building. Additionally, you are required to bring a copy of the R-NOA with you during research activities.

You are responsible for submitting a Change/Renewal Request Form to this department prior to implementing any changes to the currently approved protocol. If any problems or unexpected adverse reactions occur as a result of this study, you must notify this department immediately. Allow 45 days prior to the expiration date, if you intend to submit a Change/Renewal Request Form to extend your R-NOA date. Otherwise, submit the Executive Summary (along with the provided Cover Page) to conclude your research with OCPS and within 45 calendar days of the

R-NOA expiration. Email the form/summary to research@ocps.net. All forms may be found at this link.

Should you have questions, need assistance or wish to report an adverse event, please contact us at research@ocps.net or by phone at 407.317.3370.

Best wishes for your continued success,

Xiaogeng Sun, Ph.D.
Director of Research and Evaluation
xiaogeng.sun@ocps.net

APPENDIX H
INFORMED CONSENT

UNIVERSITY OF CENTRAL FLORIDA

EXPLANATION OF RESEARCH

Title of Project: An Analysis of Teacher Decision-Making in Grading 10th Grade Student Writing in English Language Arts

Principal Investigator: Guy Swenson

Other Investigators: None

Faculty Supervisor: RoSuan Bartee

You are being invited to take part in a research study because you teach at least one 9th or 10th grade English Language Arts class for the 2019-2020 school year. Whether you take part is up to you.

The purpose of this research is to study the decision-making process teachers use while grading 10th grade student writing in English language arts. More specifically, this research aims to study what information teachers use and what decisions teachers make when determining a grade for a sample of student writing.

At your work location, you will meet with the researcher and be provided two samples of 10th grade student writing. You will be asked to read and grade the samples while verbalizing your thoughts in real time.

The time needed to complete the research will be approximately 20 minutes.

You will be audio recorded during this study. If you do not want to be recorded, you will not be able to participate in the study. Discuss this with the researcher or a research team member. If you are recorded, the recording will be kept in a locked, safe place. The recording will be erased or destroyed when the study has been completed.

No identifiable information will be collected or stored. Your participation will be completely confidential.

You must be 18 years of age or older to take part in this research study.

Study contact for questions about the study or to report a problem: If you have questions, concerns, or complaints contact Guy Swenson, graduate student, Department of Educational Leadership and Higher Education, College of Community Innovation and Education by email at guy.swenson@Knights.ucf.edu or Dr. Jerry Johnson, Faculty Supervisor, Department of Educational Leadership and Higher Education at jerry.johnson@ucf.edu.

IRB contact about your rights in this study or to report a complaint: If you have questions about your rights as a research participant, or have concerns about the conduct of this study, please contact Institutional Review Board (IRB), University of Central Florida, Office of Research, 12201 Research Parkway, Suite 501, Orlando, FL 32826-3246 or by telephone at (407) 823-2901, or email irb@ucf.edu.

APPENDIX I
COGNITIVE LABORATORY INTERVIEW PROTOCOL

Cognitive Laboratory Interview Protocol

My name is Guy Swenson and I am a doctoral candidate at the University of Central Florida. I am conducting a study that explores teachers' decision-making processes while they are grading samples of student writing in English language arts. The purpose of this study is to determine both the commonalities and differences in teachers' decision-making processes while grading student writing. This interview is confidential and not an evaluation of your work. Rather it is designed to understand how teachers, such as yourself, think while grading student essays.

Before we begin, I have a few questions to ask you.

A) What is your teaching experience, including this year?

B) What is your current class schedule?

[Record participant responses on demographic information form along with other data elements]

Thank you. What I would like for you to do is grade two essays written by students for the same writing prompt. Grade the essays as if you gave the assignment yourself in the 4th marking period using whatever grading policies and procedures you use in your classroom with two exceptions. First, score the essay out of 100 points. Second, I would like for you to read the essay aloud while also verbalizing your thoughts as you are grading the essay. I want you to do this so that I can understand how teachers think as they score essays and what they think about while scoring essays. When I say tell me everything, I really mean every single thought you have. There is no need to plan what you are saying or try to edit your thoughts. You may also annotate the essay, but please verbalize your annotations as you make them. I may interrupt while you are reading the essay aloud, verbalizing your thoughts, or annotating to ask you describe your thoughts or your annotations. Please make your verbalizations loud enough so that the digital recording device can record your voice. Before we take an opportunity to practice, do you understand what we will be doing today?

[Answer any questions]

Let's begin a practice session.

[Give participant the practice writing prompt with student essay]

I will audio record your work while you grade the essays. The audio recording will be transcribed into written form and then kept in a secured location. No one but me and the professional transcription service will hear your recording. Your name will never be included with the recording or the transcript. When I press the record button, I will speak a few identifying words and then you may begin.

[Press record]

[End recording when finished]

Let's move onto the primary essays. You have two student essays written for the same prompt. Please take a moment to read through the prompt.

[Give prompt to participant]

Do you have any questions?

[Answer any questions]
[Start audio recorder]

Today is [current date], this is Guy Swenson and I am with interviewee number [number]. You may begin.

[Interview concludes]
Thank you for participating in this study. Your time and efforts are much appreciated. Do you have any questions?

[Answer any questions]

Thank you and have a nice day.

APPENDIX J
DEMOGRAPHIC DATA COLLECTION FORM

Demographic Data Collection Form

Participant ID Number

School ID Number

Gender:

Age:

Teaching experience including this year:

Current class schedule:

Score given on Essay 1

Score given on Essay 2

APPENDIX K
PRACTICE ESSAY

Writing Prompt

People often say “Don’t judge a book by its cover.” Describe a time when you misjudged someone based on his or her appearance or when someone misjudged you.

Practice Essay Student Response

They say you shouldn’t judge a book by it’s cover, but people often do. I learned my lesson about this in high school when I met Maria Mariella. I didn’t think she was worth getting to know but I was very wrong. She turned out to be a great friend, but by the time I realized it she was gone. Maria Mariella came to our school from Italy, she stayed with a friend of mine, Joanne. I saw Maria Mariella a lot at school and parties but I never really talked to her. Just from how she looked and dressed (like a gypsy), I didn’t think I’d like her. Then one night Joanne asked me to take Maria Mariella home because I was leaving early and she wanted to leave early too. So I did, and I found out she loved the 10,000 Maniacs as much as I did, not even my best friend liked the same music. After that we started talking and hanging out, and we kept finding that we had all kinds of things in common. The more we talked, the more we liked each other. Its a sad thing that our friendship was so short. Maria Mariella had to go back to Italy a few weeks later because her mother got sick. At her goodbye party, we were playing “Truth or Dare.” It was our favorite game. When it was Maria Mariella’s turn she said “truth.” Denise asked her to tell the truth about something she regrets. Maria Mariella said, “I wish I’d gotten to know you sooner, I didn’t think you were worth my time.” I said, me too, and that’s something we both regret.

REFERENCES

Alshenqeeti, H. (2014). Interviewing as a data collection method: A critical review. *English Linguistics Research, 3*(1). https://doi.org/10.5430/elr.v3n1p39

American Federation of Teachers, National Council on Measurement in Education, and National Education Association (1990). *Standards for teacher competence in educational assessment of students*. Retrieved from https://buros.org/standards-teacher-competence-educational-assessment-students

Anfara, V., Mertz, N. (2015). *Theoretical frameworks in qualitative research* (2nd ed.). Thousand Oaks, CA: Sage.

Armstrong, S., Gleitman, L., & Gleitman, H. (1983). What some concepts might not be. *Cognition: International Journal of Cognitive Science, 13*(3), 263–308. https://doi.org/10.1016/0010-0277(83)90012-4

Ashbaugh, E. (1924). Reducing the variability of teachers' marks. *Journal of Educational Research, 9*, 185–198. https://doi.org/10.1080/00220671.1924.10879447

Ashby, F., & Alfonso-Reese, L. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39*(2), 216–233. https://doi.org/10.1006/jmps.1995.1021

Atkinson, R., & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher, 38*, 665–676. https://doi.org/10.3102/0013189x09351981

Behizadeh, N., & Engelhard, J. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing, 16*(3), 189–211. https://doi.org/10.1016/j.asw.2011.03.001

Bowers, A. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research & Evaluation, 17*(3), 141–159. https://doi.org/10.1080/13803611.2011.597112

Branthwaite, A., Trueman, M., & Berrisford, T. (1981). Unreliability of marking: Further evidence and a possible explanation. *Educational Review, 33*(1), 41–46. https://doi.org/10.1080/0013191810330105

Brimi, H. (2011). Reliability of grading high school work in English. *Practical Assessment, Research & Evaluation, 16*(17). 1–12. Retrieved from https://doaj.org/article/9526e269c63b4065915183dc9ccc1d79

Broad, B. (2000). Pulling your hair out: "Crises of standardization in communal writing assessment." *Research in the Teaching of English, 35*(2), 213–260. Retrieved from https://www2.ncte.org/resources/journals/research-in-the-teaching-of-english/

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Logan, UT: Utah State University Press.

Brookhart, S. (1991). Grading practices and validity. *Educational Measurement: Issues and Practice*, *10*(1), 35–36. https://doi.org/10.1111/j.1745-3992.1991.tb00182.x

Brookhart, S. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues & Practice, 22*(4), 5–12. https://doi.org/10.1111/j.1745-3992.2003.tb00139.x

Brookhart, S., Guskey, T., Bowers, A., McMillan, J., Smith, J., Smith, L., Welsh, M., & Stevens, M. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803–848. https://doi.org/10.3102/0034654316672069

Campbell, A. (1921). Keeping the score. *The School Review, 29*(7), 510–519. https://doi.org/10.1086/437427

Casale, M., Roeder, J., & Ashby, G. (2012). Analogical transfer in perceptual categorization. *Memory & Cognition, 40*(3), 434–449. https://doi.org/10.3758/s13421-011-0154-4

Cascio, W. (1982). *Applied psychology in personnel management*. Reston, VA: Reston.

Charmaz, K. (2014). *Constructing grounded theory* (2nd ed.). Los Angeles, CA: Sage.

Cizek, G. (1996). Grades: the final frontier in assessment reform. *NASSP Bulletin, 80*(584), 103–110. https://doi.org/10.1177/019263659608058416

Cizek, G. (2010). An introduction to formative assessment: History, characteristics and challenges. In G. Cizek, & H. Andrade (Eds.). *Handbook of formative assessment* (pp. 3–17). New York, NY: Routledge.

Cizek, G., Fitzgerald, S., & Rachor, R. (1995). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment, 3*(2), 159–80. https://doi.org/10.1207/s15326977ea0302_3

Cooksey, R., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation: An International Journal on Theory and Practice, 13*(5), 401–434. https://doi.org/10.1080/13803610701728311

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Los Angeles, CA: Sage.

Creswell, J. (2007). *Qualitative inquiry & research design: Choosing among five approaches*. Thousand Oaks, CA. Sage.

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin, 52*(3), 177–193. https://doi.org/10.1037/h0044919

Cronbach, L. (1990). *Essentials of psychological testing* (5th ed.). New York, NY: Harper & Row.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing, 28*, 43–56. https://doi.org/10.1016/j.asw.2016.03.001

Culham, R. (1995). *6 + 1 traits of writing*. Retrieved from https://www.scholastic.com/teachers/articles/teaching-content/61-traits-writing/

Dempsey, M., PytlikZillig, L., & Bruning, R. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a web-based environment. *Assessing Writing, 14*(1), 38–61. https://doi.org/10.1016/j.asw.2008.12.003

Diederich, P. (1974). *Measuring growth in English*. Urbana, IL: National Council of Teachers of English.

Diederich, P., French, J. & Carlton, S. (1961). Factors in judgments of writing ability. *Research Bulletin,* 61–15. Princeton, NJ: Educational Testing Service (ERIC Document Reproduction Service ED 002 172). https://doi.org/10.1002/j.2333-8504.1961.tb00286.x

Edgeworth, F. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51*(3), 599–635. https://doi.org/10.2307/2341746

Eells, W. (1930). Reliability of repeated grading of essay type examinations. *Journal of Educational Psychology, 21*(1), 48–52. https://doi.org/10.1037/h0071103

Ericsson, K. (2003). Valid and non-reactive verbalization of thoughts during performance of tasks: Towards a solution to the central problems of introspection as a source of scientific data. *Journal of Consciousness Studies, 10*(9–10), 1–18. Retrieved from https://www.imprint.co.uk/product/jcs/

Farr, B. (2000). Grading practices: An overview of the issues. In E. Trumbull & B. Farr (Eds.), *Grading and reporting student progress in an age of standards* (pp. 1–22). Norwood, MA: Christopher-Gordon.

Ferriter, B. (2015, September 18). *If grades don't advance learning, why do we give them?* [Blog post]. Retrieved from https://blog.williamferriter.com/2015/09/18/if-grades-dont-advance-learning-why-do-we-give-them/

Finkelstein, I. (1913). *The marking system in theory and practice*. Baltimore, MD: Warwick & York.

Fraenkel, J., & Wallen, N., Hyun, H. (2015). *How to design and evaluate research in education.* Boston, MA; McGraw-Hill Higher Education.

Garwood, S. (1976). First-name stereotypes as a factor in self-concept and school achievement. *Journal of Educational Psychology, 68*(4), 482–487. https://doi.org/10.1037/0022-0663.68.4.482

Glaser, B. (1992). *Emergence vs forcing: Basics of grounded theory analysis*. Mill Valley, CA: Sociology Press.

Grainger, P., & Adie, L. (2014). How Do Preservice Teacher Education Students Move from Novice to Expert Assessors? *Australian Journal of Teacher Education, 39*(7). https://doi.org/10.14221/ajte.2014v39n7.9

Guilford, J. (1954). *Psychometric methods*. New York, NY: McGraw-Hill.

Guskey, T. (2004). Are zeros your ultimate weapon? *Education Digest: Essential Readings Condensed for Quick Review, 70*(3), 31–35. Retrieved from http://www.eddigest.com/index.php

Guskey, T. (2009). *Bound by tradition: Teachers' views of crucial grading and reporting issues.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA. Retrieved from https://eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED509342

Guskey, T. (2011). Five obstacles to grading reform. *Educational Leadership, 69*(3), 16–21. Retrieved from http://www.ascd.org/publications/educational-leadership/nov11/vol69/num03/Five-Obstacles-to-Grading-Reform.aspx

Guskey, T., & Bailey, J. (2001). *Developing grading and reporting systems for student learning*. Thousand Oaks, CA: Corwin Press.

Guskey, T., & Jung, L. (2016). Grading: Why you should trust your judgment. *Educational Leadership, 73*(7), 50–54. Retrieved from http://www.ascd.org

Hammersley, M. (1992). *What's wrong with ethnography?* London, UK: Routledge.

Harris, W. (1977). Teacher response to student writing: A study of the response patterns of high school English teachers to determine the basis for teacher judgment of student

writing. *Research in the Teaching of English, 11*(2), 175–185. Retrieved from https://www2.ncte.org/resources/journals/research-in-the-teaching-of-english/

Hartog, P. & Rhodes, E. (1936). *An examination of examinations*. London, UK: Macmillan.

Haswell, R. (1998). Rubrics, prototypes, and exemplars: Categorization theory and systems of writing placement. *Assessing Writing, 5*(2), 231–268. https://doi.org/10.1016/S1075-2935(99)80014-2

Haswell, R. (2001). *Beyond outcomes: Assessment and instruction within a university writing program*. Westport, CT: Ablex.

Healy, K. (1935). A study of the factors involved in the rating of pupils' compositions. *The Journal of Experimental Education, 4*(1), 50–53. https://doi.org/10.1080/00220973.1935.11009995

Hennink M. (2008). Language and communication in cross-cultural qualitative research. In P. Liamputtong (Ed.), *Doing Cross-Cultural Research: Ethical and Methodological Perspectives*. Social Indicators Research Series (pp. 21– 61). Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-8567-3_2

Hill, C., O'Grady, K., & Price, P. (1988). A method for investigating sources of rater bias. *Journal of Counseling Psychology, 35*(3), 346–350. https://doi.org/10.1037/0022-0167.35.3.346

Hodges, T., Wright, K., Wind, S., Matthews, S., Zimmer, W., & McTigue, E. (2019). Developing and examining validity evidence for the writing rubric to inform teacher educators (WRITE). *Assessing Writing, 40*, 1–13. https://doi.org/10.1016/j.asw.2019.03.001

Hook, C., & Rosenshine, B. (1979). Accuracy of teacher reports of their classroom behavior. *Review of Educational Research, 49*(1), 1–11. https://doi.org/10.2307/1169924

Hulten, C. (1925). The personal element in teachers' marks. *The Journal of Educational Research, 12*(1), 49–55. https://doi.org/10.1080/00220671.1925.10879575

Hunter, M. (1984). Knowing, teaching, and supervising. In P. Hosford (Ed.) *Using what we know about teaching* (pp. 169–192). Alexandria, VA: Association for Supervision and Curriculum Development.

Hunter, K., & Docherty, P. (2011). Reducing variation in the assessment of student writing. *Assessment & Evaluation in Higher Education, 36*(1), 109–124. https://doi.org/10.1080/02602930903215842

Huot, B. (1988). The influence of holistic scoring procedures on reading and rating student essays. In M. Williamson, & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206–236). Cresskill, NJ: Hampton Press.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. Williamson, & B. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 206–236). Cresskill, NJ: Hampton Press.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*(4), 549–566. https://doi.org/10.2307/358601

Huot, B. (2002). *(Re)articulating writing assessment for teaching and learning*. Logan, UT: Utah State University Press.

Jaccard, J., Jacoby, J. (2010). *Theory Construction and Model-Building Skills: A Practical Guide for Social Scientists*. New York: Guilford Press.

Johnson, R., Penny, J., Gordon, B., Shumate, S., & Fisher, S. (2005). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores? *Language Assessment Quarterly, 2*(2), 117–146. https://doi.org/10.1207/s15434311laq0202_2

Kalthoff, H. (2013). Practices of grading: An ethnographic study of educational assessment. *Ethnography & Education, 8*(1), 89–104. https://doi.org/10.1080/17457823.2013.766436

Krawczyk, M. (2018). Do gender and physical attractiveness affect college grades? *Assessment & Evaluation in Higher Education, 43*(1), 151–161. https://doi.org/10.1080/02602938.2017.1307320

Lazarus, S., Thurlow, M., Rieke, R., Halpin, D., & Dillon, T. (2012). Using cognitive labs to evaluate student experiences with the read aloud accommodation in math (Technical Report 67). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from https://nceo.umn.edu/docs/OnlinePubs/Tech67/TechnicalReport67.pdf

Leighton, J. (2017). *Using think-aloud interviews and cognitive labs in educational research*. New York, NY: Oxford University Press.

Lincoln, Y., & Guba, E. (1985). *Naturalistic Inquiry*. Beverly Hills, CA: Sage.

Long, T., & Johnson, M. (2000). Rigour, reliability, and validity in qualitative research. *Clinical Effectiveness in Nursing, 4*(1), 30–37. https://doi.org/10.1054/cein.2000.0106

Lynne, P. (2004). *Coming to terms: Theorizing writing assessment in composition studies*. Logan, UT: Utah State University Press.

Massachusetts Department of Elementary and Secondary Education (n.d.). Massachusetts Comprehensive Assessment System 2018 MCAS Sample Student Work Scoring Guides. Retrieved from http://www.doe.mass.edu/mcas/student/2018/question.aspx?GradeID=10&SubjectCode=ela&QuestionTypeCode=wp

McMillan, J. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, *20*(1), 20–32. https://doi.org/10.1111/j.1745-3992.2001.tb00055.x

McMillan, J. (2003). Understanding and Improving Teachers' Classroom Assessment Decision Making: Implications for Theory and Practice. *Educational Measurement: Issues and Practice, 22*(4), 34–43. https://doi.org/10.1111/j.1745-3992.2003.tb00142.x

Miles, M., & Huberman, A. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Myers, J., Scales, R., Grisham, D., Wolsey, T., Dismuke, S., Smetana, L., Yoder, K., Ikpeze, C., Ganske, K., & Martin, S. (2016). What about writing? A national exploratory study of writing instruction in teacher preparation programs. *Literacy Research and Instruction, 55*(4), 309–330. https://doi.org/10.1080/19388071.2016.1198442

Myford, C., & Wolfe, E. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422. Retrieved from http://jampress.org/

Nagin, C. (2003). *Because writing matters: Improving student writing in our schools*. San Franscisco, CA: Jossey-Bass.

National Assessment of Educational Progress (n.d.). *Data tools: State profiles*. Retrieved from https://www.nationsreportcard.gov/profiles/stateprofile/overview/MA?cti=PgTab_ScoreComparisons&chort=1&sub=MAT&sj=MA&fs=Grade&st=MN&year=2017R3&sg=Gender%3A+Male+vs.+Female&sgv=Difference&ts=Single+Year&tss=2002R3-2017R3&sfj=NP

Newton, P. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170. https://doi.org/10.1080/09695940701478321

Orasanu, J., & Connolly, T. (1993). The reinvention of decision making. In G. Klein, J. Orasanu, R. Calderwood, & C. Zsambok (Eds.), *Decision making in action: Models and methods*, (pp. 3–20). Norwood, NJ: Ablex.

Patton, M. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research 34*(5), 1189–1208. Retrieved from http://www.hsr.org/

Pezalla, A., Pettigrew, J., & Miller-Day, M. (2012). Researching the researcher-as-instrument: An exercise in interviewer self-reflexivity. *Qualitative Research*, *12*(2), 165–185. https://doi.org/10.1177/1487941111422107

Phipps, S., & Borg, S. (2009). Exploring tensions between teachers' grammar teaching beliefs and practices. *System, 37*(3), 380–390. https://doi.org/10.1016/j.system.2009.03.002

Pula, J., & Huot, B. (1993). A model of background influences on holistic raters. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237–265). Cresskill, NJ: Hampton Press.

Purpura, J. (2016). Second and foreign language assessment. *The Modern Language Journal*, 100 (Supplement 2016), 190–208. https://doi.org/10.1111/modl.12308

Rashidi N., Moghadam M. (2014). The discrepancy between teachers' belief and practice. *Sociocultural Perspective: Studies in English Language Teaching, 3*(3), 252–274. https://doi.org/10.22158/selt.v3n3p252

Rezaei, A., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18–39. https://doi.org/10.1016/j.asw.2010.01.003

Rink, J. (1993). Teacher education: A focus on action. *Quest*, 45, 308–320. Retrieved from http://www-tandfonline-com.ezproxy.net.ucf.edu/

Rosen, Y., Ferrara, S., & Mosharraf, M. (2016). *Handbook of research on technology tools for real-world skill development*. https://doi.org/10.4018/978-1-4666-9441-5

Rugg, H. (1918). Teachers' marks and the reconstruction of the marking system. *The Elementary School Journal, 18*, 701–719. https://doi.org/10.1086/454643

Saal, F., Downey, R., & Lahey, M. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413–428. https://doi.org/10.1037/0033-2909.88.2.413

Sakyi, A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate ESL compositions. In: A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 129–152). Cambridge, UK: Cambridge University Press.

Saldaña, J. (2014). *Thinking qualitatively: Methods of mind*. Los Angeles, CA: Sage.

Sawyer, R. (2013). Beyond correlations: Usefulness of high school GPA and test scores in making college admissions decisions. *Applied Measurement in Education, 26,* 89–112. https://doi.org/:10.1080/08957347.2013.765433

Schneider, C., & Bodensohn, R. (2017). Student teachers' appraisal of the importance of assessment in teacher education and self-reports on the development of assessment competence. *Assessment in Education: Principles, Policy & Practice, 24*(2), 127–146. https://doi.org/10.1080/0969594x.2017.1293002

Scullen, S., Mount, M., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*(6), 956–970. https://doi.org/10.1037/0021-9010.85.6.956

Seger, C. (2009). The involvement of corticostriatal loops in learning across tasks, species, and methodologies. In H. Groenewegen, P. Voorn, H. Berendse, A. Mulder, & A. Cools (Eds.). *The basal ganglia IX*. [electronic resource]. Dordrecht, NY: Springer. Retrieved from https://link.springer.com/book/10.1007/978-1-4419-0340-2

Seger, C., & Miller, E. (2010). Category learning in the brain. *Annual Review of Neuroscience, 33*, 203–219. https://doi.org/10.1146/annurev.neuro.051508.135546

Seger, C., & Peterson, E. (2013). Categorization = decision making + generalization. *Neuroscience and Biobehavioral Reviews, 37*(7), 1187–1200. https://doi.org/10.1016/j.neubiorev.2013.03.015

Shavelson, R. (1973). What is the basic teaching skill? *Journal of Teacher Education, 24*(2), 144–151. https://doi.org/10.1177/002248717302400213

Shavelson, R., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research, 51*(4), 455–498. https://doi.org/10.2307/1170362

Smith, W. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206–236). Cresskill, NJ: Hampton Press.

Smith, C., & Dunstan, A. (1998). Grade the learning, not the writing. In F. Zak, & C. Weaver (Eds), *The theory and practice of grading writing: Problems and possibilities* (pp. 163–170). Albany, NY: State University of New York Press.

Speck, B. (1998). Unveiling some of the mystery of professional judgment in classroom assessment. *New Directions for Teaching and Learning, 74*(Summer 1998), 17–31. https://doi.org/10.1002/tl.7402

Speck, B., & Jones, T. (1998). Direction in the grading of writing? In F. Zak, & C. Weaver (Eds.), *The theory and practice of grading writing: Problems and possibilities* (pp. 17–30). Albany, NY: State University of New York Press.

Spool, M. (1978). Training programs for observers of behavior: A review. *Personnel Psychology, 31*(4), 853–888. https://doi.org/10.1111/j.1744-6570.1978.tb02128.x

Starch, D. (1913). Reliability and distribution of grades. *Science, 38*(983), 630–636. https://doi.org/10.1126/science.38.983.630

Starch, D. (1915). Can the variability of marks be reduced? *School & Society, 2*(33), 242–243. Retrieved from http://hdl.handle.net/2027/coo.31924112891472

Starch, D., & Elliott, E. (1912). Reliability of the grading of high school work in English. *School Review, 20*(7), 442–457. https://doi.org/10.1086/435971

Stern, L., & Solomon, A. (2006). Effective faculty feedback: The road less traveled. *Assessing Writing, 11*(1), 22–41. https://doi.org/10.1016/j.asw.2005.12.001

Stiggins, R. (1991). Assessment literacy. *The Phi Delta Kappan, 72*(7), 534–539. Retrieved from https://journals.sagepub.com/home/pdk

Straub, R. (1997). Students' reactions to teacher comments: An exploratory study. *Research in the Teaching of English, 31*(1), 91–119. Retrieved from https://www2.ncte.org/resources/journals/research-in-the-teaching-of-english/

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Thousand Oaks, CA: Sage.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21–36. https://doi.org/10.1017/S0267190509090035

Tierney, R., Simon, M., & Charland, J. (2011). Being fair: Teachers' interpretations of principles for standards-based grading. *Educational Forum, 75*(3), 210–227. https://doi.org/10.1080/00131725.2011.577669

van Teijlingen, E., & Hundley, V. (2001). The importance of pilot studies. *Social Research Update,* (35), 1–4. Retrieved from http://sru.soc.surrey.ac.uk/SRU35.html

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–125). Norwood, NJ: Ablex.

Wade, B. (1978). Responses to written work: The possibilities of utilizing pupils' perceptions. *Educational Review, 30*(2), 149–158. https://doi.org/10.1080/0013191780300208

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–87. https://doi.org/10.1177/026553229801500205

Weigle, S. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*(3), 194–209. https://doi.org/10.1016/j.jslw.2007.07.004

Wells, F. (1907). A statistical study of literary merit (Columbia university contributions to philosophy and psychology, vol. XVI, no. 3). *Archives of Psychology*, 7, 1–20. Retrieved from https://brocku.ca/MeadProject/Wells/Wells_1907.html

White, E. (2009). Are you assessment literate? Some fundamental questions regarding effective classroom-based assessment. *OnCUE Journal, 3*(1), 3–25. Retrieved from https://jaltcue.org/content/about-oncue-journal

Wilen, W., Ishler, M., Hutchinson, J., & Kindsvatter, R. (2000). *Dynamics of effective teaching* (4th edition). New York, NY: Longman.

Willis, G. (2015). *Analysis of the cognitive interview in questionnaire design: Understanding qualitative research*. New York, NY: Oxford University Press.

Wiseman, C. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*(3), 150–173. https://doi.org/10.1016/j.asw.2011.12.001

Whittemore, R., Chase, S., & Mandle, C. (2001). Validity in Qualitative Research. *Qualitative Health Research*, *11*(4), 522–538. https://doi.org/10.1177/104973201129119299

Wolcott, H. (1994). *Transforming qualitative data: Description, analysis, and interpretation*. Thousand Oaks, CA: Sage

Wolfe, E., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing, 27*(2016), 1–10. https://doi.org/10.1016/j.asw.2015.06.002

Wyatt-Smith, C. (1999). Reading for assessment: How teachers ascribe meaning and value to student writing. *Assessment in Education: Principles, Policy and Practice, 6*(2), 195–223. https://doi.org/10.1080/09695949992874

Wyatt-Smith, C. & Castleton, G. (2005). Examining how teachers judge student writing: An Australian case study. *Journal of Curriculum Studies. 37*(2), 131–154. https://doi.org/10.1080/0022027032000242887

Wylie, E., & Lyon, C. (2015). The fidelity of formative assessment implementation: issues of breadth and quality. *Assessment in Education: Principles, Policy & Practice, 22*(1), 140–160. https://doi.org/10.1080/0969594x.2014.990416

Yamauchi, T., Love, B., & Markman, A. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 585–593. https://doi.org/10.1037/0278-7393.28.3.585