THE VALIDITY AND RELIABILITY OF A SINGLE-POINT RUBRIC TO ASSESS STUDENT WRITING PERFORMANCE

David Popken
THE VALIDITY AND RELIABILITY OF A SINGLE-POINT RUBRIC
TO ASSESS STUDENT WRITING PERFORMANCE

David Popken

Master of Science, English Education, SUNY College at Oneonta, 1998
Bachelor of Science, English Education, SUNY College at Oneonta, 1995

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Education in Instructional Leadership
in the
Department of Education and Educational Psychology
at
Western Connecticut State University
2020
THE VALIDITY AND RELIABILITY OF A SINGLE-POINT RUBRIC  
TO ASSESS STUDENT WRITING PERFORMANCE

David Popken, B.S., M.S., Ed.D.

Western Connecticut State University

Abstract

Rubrics have become a widely accepted instructional and evaluation tool in contemporary classrooms. As a mechanism for formative assessment, rubrics have been found to clarify task criteria, increase motivation and self-efficacy, and provide greater efficiency and objectivity to the grading process. In this study, a convenience sample was employed to correlate student performance using a single-point rubric with student performance using a scaled analytic rubric with pre-established validity and reliability. Gender, course enrollment, grade-point average (GPA), self-reported use and student perceptions of the efficacy of single-point rubrics were also utilized as independent variables. Grade-point average and the participants’ scaled analytic rubric scores were significant predictors of participants’ single-point rubric scores, establishing validity for the single-point rubric. Student scores from single-point rubrics were also be utilized to determine inter-rater reliability and item analysis was conducted to determine inter-rater agreement. A strong, positive significant correlation between rater one and rater two essay scores suggests inter-rater reliability, abetted by a high percentage of inter-rater agreement in rater one and rater two scoring of individual items on the single-point rubric.
Copyright

By

David Popken, Ed.D.

2020
Doctor of Education Dissertation

THE VALIDITY AND RELIABILITY OF A SINGLE-POINT RUBRIC
TO ASSESS STUDENT WRITING PERFORMANCE

Presented by
David P. Popken, Ed.D.

Sally M. Dobyns, Ph.D.  Signature  April 10, 2020
Dissertation Committee Chair

Marcia A. B. Delcourt, Ph.D.  Signature  April 10, 2020
Dissertation Committee Member

Nicholas Kowgios, Ed.D.  Signature  April 10, 2020
Dissertation Committee Member

Jody S. Piro, Ed.D.  Signature  April 10, 2020
Interim Program Coordinator

2020
ACKNOWLEDGEMENTS

It would be remiss of me not to acknowledge the contributions of my professors, colleagues, and friends (all of whom fit more than one of these designations) who contributed to the completion of my coursework and research study.

My committee—consisting of Dr. Sally Dobyns, Dr. Marcy Delcourt, and Dr. Nick Kowgios—was instrumental in guiding me through the dissertation process from proposal to defense. Their encouragement and guidance, as well as the time and effort they invested in my work, were invaluable to the successful completion of my research study.

I would never have been prepared to take on the study had it not been for the preparation provided during my coursework with Dr. Jody Piro and Dr. Trish Stewart. Both provided educational experiences that extended well beyond dissertation preparation. The highest compliment I can give is that I’m a better teacher today because of my time with them.

My outside reader, Dr. Pat Cyganovich, is undoubtedly one of the most influential people in my maturation as an educator. Pat, my building principal for 17 years before her retirement, displayed the best qualities of an educational leader—her approach was visionary yet practical, global yet nuanced, resolved yet selfless. I feel lucky to have had her as a mentor and blessed to count her as a friend.

And, of course, I would like to thank (and congratulate!) Dr. Doug Coates and Dr. Julio Vazquez. Five years ago, we began this journey together as colleagues. Today, we conclude it as friends. I consider this the greatest takeaway of all.
DEDICATION

To my beautiful wife Kim—my partner-in-crime—who covered for me on the home front over the last five years. Our kids are probably better human beings for it.

To my son Jake—my inspiration—whose own educational experiences have shaped my research and practice. I’m proud of you as a student, an athlete, and—most importantly—as a young man (again, mostly mom here).

And to my daughter Kaylie—my distraction—whose most influential role in the process was undermining it at every turn. You always remind me of what’s really important in life.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>i</td>
</tr>
<tr>
<td>Copyright</td>
<td>ii</td>
</tr>
<tr>
<td>APPROVAL PAGE</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iv</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>v</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>vi</td>
</tr>
<tr>
<td>TABLE OF TABLES</td>
<td>xii</td>
</tr>
<tr>
<td>TABLE OF FIGURES</td>
<td>xiv</td>
</tr>
<tr>
<td>CHAPTER ONE: INTRODUCTION TO THE STUDY</td>
<td>1</td>
</tr>
<tr>
<td>Rationale for Selecting the Topic</td>
<td>3</td>
</tr>
<tr>
<td>Statement of the Problem</td>
<td>4</td>
</tr>
<tr>
<td>Potential Benefits of the Research</td>
<td>4</td>
</tr>
<tr>
<td>Definition of Key Terms</td>
<td>5</td>
</tr>
<tr>
<td>CHAPTER TWO: REVIEW OF THE LITERATURE</td>
<td>9</td>
</tr>
<tr>
<td>Chapter Overview</td>
<td>9</td>
</tr>
<tr>
<td>Literature Search Process Overview</td>
<td>9</td>
</tr>
<tr>
<td>Theoretical Construct</td>
<td>10</td>
</tr>
<tr>
<td>Vygotsky’s Theory of Zone of Proximal Development</td>
<td>10</td>
</tr>
<tr>
<td>Construct of Formative Assessment</td>
<td>11</td>
</tr>
<tr>
<td>Historical Context of the Use of Rubrics for Writing Assessments</td>
<td>13</td>
</tr>
<tr>
<td>Recent Studies Concerning Rubrics as Formative Assessments</td>
<td>15</td>
</tr>
<tr>
<td>Topic</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Reliability and Validity of Rubrics</td>
<td>20</td>
</tr>
<tr>
<td>Chapter Summary</td>
<td>25</td>
</tr>
<tr>
<td>CHAPTER THREE: METHODOLOGY</td>
<td>26</td>
</tr>
<tr>
<td>Researcher Biography</td>
<td>26</td>
</tr>
<tr>
<td>Research Questions and Hypotheses</td>
<td>27</td>
</tr>
<tr>
<td>Description of the Setting and the Participants</td>
<td>28</td>
</tr>
<tr>
<td>Description of Setting</td>
<td>28</td>
</tr>
<tr>
<td>Description of Participants</td>
<td>29</td>
</tr>
<tr>
<td>Program Completion Rate</td>
<td>32</td>
</tr>
<tr>
<td>Research Design</td>
<td>33</td>
</tr>
<tr>
<td>Description of Analyses</td>
<td>34</td>
</tr>
<tr>
<td>Instrumentation</td>
<td>34</td>
</tr>
<tr>
<td>Student Demographic Survey</td>
<td>35</td>
</tr>
<tr>
<td>Writing Task</td>
<td>35</td>
</tr>
<tr>
<td>Scaled Analytic Rubric</td>
<td>35</td>
</tr>
<tr>
<td>Single-Point Instructional Rubric</td>
<td>36</td>
</tr>
<tr>
<td>Procedures</td>
<td>38</td>
</tr>
<tr>
<td>Timeline for Data Collection</td>
<td>39</td>
</tr>
<tr>
<td>Statement of Ethics</td>
<td>40</td>
</tr>
<tr>
<td>Chapter Summary</td>
<td>41</td>
</tr>
<tr>
<td>CHAPTER 4: ANALYSIS OF DATA AND EXPLANATION OF FINDINGS</td>
<td>42</td>
</tr>
<tr>
<td>Overview of the Study</td>
<td>42</td>
</tr>
<tr>
<td>Research Questions and Hypotheses</td>
<td>42</td>
</tr>
</tbody>
</table>
Research Question One 42
Research Question Two 43
Data Preparation for Research Question One 44
Stepwise Multilinear Regression 44
Data Screening Process 45
Missing Data 45
Codebook 45
Multilinear Regression Assumptions 52
Initial Screening Process 52
Descriptive Statistics 54
Normality 54
Linearity 56
Homoscedasticity 57
Multivariate Outliers 58
Multicollinearity 60
Remediation of Violation of Multicollinearity 61
Suppressor Variables 61
Research Question One Analysis 62
Frequencies 63
Means, Standard Deviations, and Intercorrelations 63
Stepwise Multilinear Regression Analysis 66
Data Preparation for Research Question Two 68
Data Screening Process 68
Missing Data 68
Inter-Rater Reliability 69
  Pearson’s Correlation Coefficient 69
  Pearson’s Correlation Coefficient Assumptions 69
  Descriptive Statistics 70
  Normality 70
  Linearity and Outliers 72
  Homoscedasticity 73
Research Question Two Data Analyses of Inter-Rater Reliability 74
Inter-Rater Agreement 75
  Simple Percentage Agreement 75
  Assumptions of Simple Percentage Agreement 76
  Cohen’s Kappa 78
  Assumptions of Cohen’s Kappa 79
Research Question Two Data Analyses of Inter-Rater Agreement 79
Chapter Summary 80
CHAPTER FIVE: SUMMARY AND CONCLUSIONS 82
Synopsis of the Study 82
Results 84
  Research Question One 84
  Research Question Two 85
Discussion 86
  Research Questions One 87
Research Question Two

Implications for Education

Limitations to the Study

Internal Threats

External Threats

Suggestions for Future Research

Student Perceptions of Single-Point Writing Rubrics

Teacher Attitudes Toward Single-Point Writing Rubrics

Effectiveness of Single-Point Writing Rubric in Diverse Populations

Additional Validity and Reliability Studies of Single-Point Writing Rubrics

Chapter Summary

References

Appendices

Appendix A: NYS Regents Scaled Analytic Rubric

Appendix B: Single-Point Rubric

Appendix C: Student’s Assent

Appendix D: Parent’s Consent

Appendix E: Student Demographic Survey

Appendix F: First NYS Regents Writing Task

Appendix G: Second NYS Regents Writing Task

Appendix H: Superintendent’s Consent

Appendix I: Principal’s Consent
## TABLE OF TABLES

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Student Enrollment in Each Class of the Accessible Population</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2</td>
<td>Race/Ethnicity of Accessible Population</td>
<td>31</td>
</tr>
<tr>
<td>Table 3</td>
<td>Gender with Respect Course Enrollment in Accessible Student Population</td>
<td>32</td>
</tr>
<tr>
<td>Table 4</td>
<td>Race/Ethnicity of Sample Population</td>
<td>33</td>
</tr>
<tr>
<td>Table 5</td>
<td>Gender with Respect Course Enrollment in Sample Student Population</td>
<td>33</td>
</tr>
<tr>
<td>Table 6</td>
<td>SPSS Codebook of Demographic Information and Variables</td>
<td>47</td>
</tr>
<tr>
<td>Table 7</td>
<td>Standard Deviation of Differences from Mean for SPR Scores</td>
<td>49</td>
</tr>
<tr>
<td>Table 8</td>
<td>Frequencies for Research Question One Categorical Variables</td>
<td>53</td>
</tr>
<tr>
<td>Table 9</td>
<td>Descriptive Statistics for Research Question One Continuous Variables</td>
<td>54</td>
</tr>
<tr>
<td>Table 10</td>
<td>Extreme Values Test for Research Question One</td>
<td>60</td>
</tr>
<tr>
<td>Table 11</td>
<td>Pearson Product-Moment Correlations Between Dependent Variables</td>
<td>61</td>
</tr>
<tr>
<td>Table 12</td>
<td>Stepwise Multilinear Regression Coefficients</td>
<td>62</td>
</tr>
<tr>
<td>Table 13</td>
<td>Coefficients for the Stepwise Multilinear Regression</td>
<td>65</td>
</tr>
<tr>
<td>Table 14</td>
<td>Collinearity Diagnostics for Research Question One</td>
<td>66</td>
</tr>
<tr>
<td>Table 15</td>
<td>Model Summary for Research Question One</td>
<td>67</td>
</tr>
<tr>
<td>Table 16</td>
<td>ANOVA for Research Question One</td>
<td>67</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Table 17</td>
<td>Descriptive Statistics for Research Question Two Continuous Variables</td>
<td>70</td>
</tr>
<tr>
<td>Table 18</td>
<td>Pearson’s Correlation Coefficient for Single-point Rubric Scores</td>
<td>75</td>
</tr>
<tr>
<td>Table 19</td>
<td>Simple Percentage Agreements of Single-Point Rubric Items on Student Essays</td>
<td>77</td>
</tr>
<tr>
<td>Table 20</td>
<td>Cohen’s and McHugh’s Interpretations of Cohen’s Kappa</td>
<td>79</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>----------</td>
<td>------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Figure 1</td>
<td>Normal P-P Plot of Regression Standardized Residual</td>
<td>56</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Scatterplot of Standardized Residuals</td>
<td>57</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Matrix of Scatterplots of Relationships of Continuous Variables</td>
<td>58</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Boxplot of Average Essay Scores Using Single-Point Rubric</td>
<td>59</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Normal Q-Q Plot of Single-Point Rubric Scores (1st reader)</td>
<td>71</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Normal Q-Q Plot of Single-Point Rubric Scores (2nd reader)</td>
<td>72</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Scatterplot of Rater One and Rater Two SPR Scores</td>
<td>73</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Matrix of Scatterplots of Rater One and Rater Two SPR Scores</td>
<td>74</td>
</tr>
</tbody>
</table>
CHAPTER ONE: INTRODUCTION TO THE STUDY AND IDENTIFICATION OF THE TOPIC

Rubrics have become a widely accepted instructional and evaluation tool in contemporary classrooms (Moskal, 2000). As the primary operational mechanism for pedagogy rooted in formative assessment (Black & Wiliam, 1998; Panadero & Jonsson, 2013; Schafer, Swanson, Bene, & Newberry, 2001), rubrics—when used as instructional tools—have been found to clarify task criteria (Mehrans, Popham, & Ryan, 1998), increase motivation (Brookhart, 1997), and improve self-efficacy (Andrade, Wang, Du, & Akawi, 2009) for students while providing greater efficiency and objectivity to the grading process for notoriously besieged classroom teachers (Beaglehole, 2014).

The limited empirical studies that have been conducted to verify these benefits have been primarily structured as quantitative quasi-experimental studies measuring the performance of students with and without the use of rubrics to complete writing tasks (Andrade, 1999; Bradford, Newland, Rule, & Montgomery, 2015), or the effect on achievement based on the degree to which teachers use rubrics as instructional tools during the writing process (Andrade, 1999; Panadero & Jonsson, 2013), particularly in classrooms where students engage in self-assessment (Andrade, Buff, Terry, Erano, & Paolino, 2009). Qualitative and mixed methods studies have sought a more nuanced understanding of student perceptions regarding rubrics—how they use them and to what degree rubrics impact their self-efficacy on assessments (Andrade, 1999; Andrade & Du, 2005; Andrade et al., 2009).

But a dearth of research exists regarding rubric structure. Researchers are quick to define the contextual pros and cons of the most frequently used structures (Balch, Blanck, & Balch, 2016)—holistic, analytic, and, to a lesser degree, single-point (Fluckiger, 2010)—but rarely
support these assumptions with empirical evidence. Scaled analytic rubrics (see Appendix A), matrices that outline scaled levels of performance within the domains (e.g., content; command of evidence; organization; conventions) of a given task, are one of the most frequently used structures for instructive and evaluative purposes. Yet a gap in the literature exists regarding the widely accepted belief that the scaled analytic rubric, most often modeled after the holistic rubrics utilized on standard assessments or the institutional standards that inspired them, is the most effective rubric structure, with proponents conflating teachers’ ability to differentiate degrees of proficiency for evaluation with students’ ability to practically utilize scaled rubrics as instructional tools to improve their performance.

Further, single-point rubrics, which provide only one level of performance (i.e., competency) and are much more linear and concrete to the structure of the task and with more prescribed domain descriptors (see Appendix B), are often glossed over in literature reviews as adequate for measuring competency but powerless to serve as comprehensive evaluation tools. Indeed, these assumptions may be true regarding the evaluation of writing for summative assessments, but in classrooms that utilize formative assessments as instructional tools, single-point rubrics could potentially offer a learner-centered alternative to the teacher-centered matrices that have become commonly accepted as the most effective instructional rubric structure. This study promises to add to the literature on the impact of rubrics on student performance on written assessments by correlating student performance using a single-point rubric with student performance using a scaled analytic rubric with pre-established validity and reliability. Gender, course enrollment, GPA, self-reported use and student perceptions of its efficacy as an instructional tool were also considered as potential predictor variables of student performance using the single-point rubric. Lastly, the overall inter-rater reliability of the single-
point rubric, as well as inter-rater agreement among specific single-point rubric items, were measured.

**Rationale for Selecting the Topic**

The use of rubrics as instructive and evaluative tools in the writing classroom are generally accepted practices from elementary school through higher education. While rubrics “are not a replacement for good instruction” (Andrade, 2005, p. 29), the benefits—objectivity and alignment for teachers, clarity and specificity for students (Allen & Tanner, 2006)—are conspicuous and logical. Yet few recent studies have examined the effect of rubrics on student growth (Bradford et al., 2015), and even fewer compare the effectiveness of rubric design. Writ large, all rubrics do not equally serve as learner-centered instructional tools. The usefulness of rubrics is contingent on the quality of the rubrics (Jonsson & Svingby, 2007). This includes “setting clear and specific product goals” (Beaglehole, 2014, p. 13), with language and structure that is accessible to the students, as not all rubrics are “self-explanatory” (Andrade, 2005, p. 29). Scaled analytic rubrics, for example, commonly use language such as “substantially developed, mostly developed, developed, and underdeveloped” (Allen & Tanner, 2006, p. 200) to distinguish gradations of support representing a scale of performance. But considering teachers spend hours collaboratively training to establish inter and intra-rater reliability when interpreting rubrics and anchor papers to evaluate essays for standardized tests, how can we expect the students themselves to distill the vagueness between “substantially” and “mostly”? Hence, single-point rubrics could potentially offer a more learner-centered alternative, representing the presence or absence of a criterion or standard.

This is the crux of Brookhart’s (2003) argument for developing a measurement theory for classroom assessment purposes—that classroom assessment instruments should not be developed
by “borrowing theory” (p. 5) from the domain of large-scale assessments, and that assessments should be integrated with instruction to benefit all students. The assumption that scaled analytic rubrics have this power has been challenged (Andrade & Du, 2005; Orsmond & Merry, 1996), but little research exists comparing the impact of rubric structures on student performance. Noting the need for how students actually use rubrics, Andrade and Du (2005) encourage research that can “inform our understanding of whether and how assessments can serve the purpose of learning” (p. 7). This researcher seeks to contribute to the literature by establishing validity for a single-point rubric by correlating student performance using a single-point rubric with student performance using a scaled analytic rubric with pre-established validity. This researcher also seeks to contribute to the literature by establishing inter-rater reliability among overall scores and inter-rater agreement for each item on the single-point rubric.

**Statement of the Problem**

While some quantitative evidence exists that measures the effect of rubrics on student performance (Andrade & Du, 2005; Bradford et al., 2015; Francis, 2018), little research has been conducted measuring the effectiveness of specific rubric design on instructional outcomes.

**Potential Benefits of the Research**

This research may add to the literature on the impact of rubrics on student performance by correlating student performance using single-point rubrics with student performance using scaled analytic rubrics, as well as gender, course enrollment, GPA, self-reported use and student perceptions of the efficacy of single-point rubrics. Additionally, this study measured the inter-rater reliability of overall scores using a single-point rubric and was used to examine the inter-rater agreement of specific items on the single-point rubric. While much previous peer-reviewed literature has been anecdotal, instructional, and promotional (Allen & Tanner, 2006; Baker,
Cooperman, & Storandt, 2013; Beaglehole, 2014; Fluckiger, 2010; Martin, 2013; Wolf & Stevens, 2007), few studies (Andrade & Du, 2005; Andrade et al., 2009; Bradford et al., 2015; Dogan & Uluman, 2017; Francis, 2018; Jonsson & Svingby, 2007) have quantitatively or qualitatively measured the efficacy of rubrics or rubric design on improving instruction and outcomes.

The general belief that well designed instructional rubrics can be valid and reliable (Jonsson & Svingby, 2007), support learning (Andrade, 2000), increase motivation (Brookhart, 1997), and improve evaluation efficiency for the “busy classroom teacher” (Beaglehole, 2014, p. 13) has led to the omnipresence of rubrics at all levels of education (Jonsson & Svingby, 2007; Popham, 1997; Wolf & Stevens, 2007). However, additional research specific to rubric design could help improve efficacy of rubrics—in this case, single-point rubrics—as instructional tools for students to better and more consistently meet achievement goals, what Schafer et al. (2001) consider rubrics’ primary role, as well as the primary need of further investigation.

**Definition of Key Terms**

1. **Analytic rubrics** differ from holistic rubrics by allowing the rater to evaluate the student’s performance exclusive to each domain (Balch et al., 2016). For example, a student essay evaluated using an analytic rubric with scaled gradations from 1-4 might receive a score of 3 for support and a score of 2 for language use, with an overall score determined based on the average, sum, or some other coalescing of the scores for each domain.

2. **Formative assessments** are designed to not only measure student achievement but also to provide feedback to teachers and students to adjust instruction and performance in real time; a formative assessment “embeds assessment in instruction”
by providing the rubric (and potentially task-specific lessons using the rubric) prior to an assessment task, allowing students to “plan their assessment task, clarify targets, determine and focus effort where needed, identify issues related to the task, regulate the process in the effort to produce high-quality work” (Ragupathi & Lee, 2020, p. 79).

3. **Holistic rubrics** provide descriptions for varying levels of student performance in all domains of a task (e.g., in a persuasive essay rubric, the domains of support, language use, conventions), but only a single score is given in evaluation of the student’s overall performance (Dogan & Uluman, 2017).

4. **Inter-rater reliability** and **inter-rater agreement** apply when there is more than one judge, or in the case of a classroom setting, teacher, who is assessing student performance using a new instrument. The objective of these values is to gauge the statistical measurement of the variability of the scores when there are different raters using the same instrument. Ideally, raters would assign the same scores to the same performances using the same instrument (Muijs, 2011). Inter-rater agreement is “whether diagnoses, scores, or judgments are identical or similar or the degree to which they differ” when using multiple scorers (Kottner & Streiner, 2011, p. 701), which is often measured using simple percentage agreement (Stemler, 2004). Inter-rater reliability is “the ratio of variability between scores of the same subjects (e.g., by different raters or at different times) to the total variability of all scores in the sample” utilizing coefficients (Kottner & Streiner, 2011, p. 701).
5. Rubrics come in many forms (e.g., holistic, analytic, single-point) but are generally understood as matrices that outline scaled levels of performance within the domains of a given task, the purpose of which is to outline standards for assessment and levels of achievement based on the criteria of those standards (Allen & Tanner, 2006). A rubric essentially “lists the criteria for a piece of work, or ‘what counts’ (for example, purpose, organization, details, voice, and mechanics are often what count in a piece of writing); it also [often] articulates gradations of quality for each criterion, from excellent to poor” (Andrade, 2014, p. 1). Rubrics can be utilized to assess many types of performances (e.g., processes like playing an instrument or making an oral presentation; products like a written essay or wooden bench) in different academic disciplines (Brookhart, 2013). For summative assessments, rubrics can provide “fairness and consistency” while for formative assessments rubrics “can act as scaffolds for assessment as learning” (Ragupathi & Lee, 2020, pp. 77-79).

6. Single-point rubrics differ from holistic or analytic rubrics in that they provide “only one level of performance, the proficient level” of each area to be measured in a task (Fluckiger, 2010, p. 19). No additional levels above or below proficiency are described or employed.

7. Summative assessments are used to judge the merits of performance at the conclusion of a task. The primary objective of summative assessments is evaluation of performance in relation to the criteria of the task (Scriven, 1967). Examples of summative assessments in school settings include teacher-generated unit tests to evaluate student learning and state and national assessments like New York State’s subject specific Regents exams (New York State Education Department, 2019) or the
College Board’s SAT test (College Board, 2020) and ACT test (ACT, 2020) to gauge the college readiness of students at the conclusion of high school.

8. The *validity* of an instrument, in this case a rubric, is “the extent to which it measures what it is intended to measure” (Rossi, Lipsey, & Freeman, 2004, p. 220). Demonstrations of the content validity of an instrument can be based on correlations with a similar instrument (i.e., one with alternative measures of the same outcomes) that has pre-established validity (Jonsson & Svingby, 2007).
CHAPTER TWO: REVIEW OF THE LITERATURE

Chapter Overview

First, a discussion of Vygotsky’s (1978) zone of proximal development as it relates to the use of rubrics as instruments of formative assessments is discussed in theory and empirical research. Second, a brief history of the use of rubrics in writing assessments to contextualize their current role in the evaluation process is provided. Third, recent studies concerning the use of rubrics in formative settings is discussed. The data from these studies generally supports the efficacy of rubrics as formative instruments that can concomitantly be used for instruction and assessment. Fourth, shifting theory regarding the validity and reliability of rubric design and application is explored. Recent studies question the conventional wisdom that rubrics alone can increase reliability and validity to the assessment process. The literature search process details the selection of all articles used in this review of literature.

Literature Search Process Overview

A search of the literature was conducted to ground this study in theory and research. Internet and article database searches were conducted using, among others, the following search terms: “rubrics,” “rubric history,” “writing rubrics,” “rubrics and assessment,” “Vygotsky and rubrics,” “formative assessments and rubrics,” “Vygotsky and formative assessments,” “Vygotsky and formative assessments and rubrics,” “rubrics and validity and reliability,” and “single-point rubrics.” The application of these terms resulted in over 10,000 articles that were potentially useful in theory and research.

Theoretical resources regarding zone of proximal development (Vygotsky, 1978) and formative assessments provide the construct for the use of rubrics as instruments that can potentially influence both instruction and assessment. Sources chronicling the use of rubrics in
the assessment of writing were selected for review because they contextualize the historical
objectives of rubrics in educational settings. Recent studies regarding the use of rubrics in
formative settings inform that discussion with empirical evidence. Articles and studies
exploring, questioning, and testing the validity and reliability of rubrics, even exclusively as
assessment tools, inform the research design of this study. Excluded were a plethora of
methodological articles—how-to articles—that focused on design and implementation without
theoretical or empirical support. Given the relatively short history of rubric usage, most
empirical studies found were conducted within the last 20 years and, therefore, included in this
review.

The 31 sources included for this review of literature were ultimately chosen because of
theoretic discussions, empirical evidence, research design and methodology that paralleled,
contradicted, informed, or influenced the theory, hypotheses, design, methodology, and
conclusions drawn in this study.

**Theoretical Construct**

**Zone of Proximal Development (ZPD)**

The underlying theoretical foundation for rubrics as instructional tools is rooted in
Vygotsky’s (1978) zone of proximal development:

It is the distance between the actual developmental level as determined by independent
problem solving and the level of potential development as determined through problem
solving under adult guidance or in collaboration with more capable peers. (p. 86)

Vygotsky’s theory is best illustrated through his example of two 10-year-old children who test at
an eight-year-old child’s developmental level. Vygotsky considers this the first level of
development—“the actual developmental level” (p. 85). The second level is determined when
the teacher offers assistance to an assessment task through suggestions, explanations, and demonstrations to establish the child’s proximal development. One child, with assistance, reveals the mental capabilities of a 12-year-old, the other of a nine-year-old. Vygotsky posited that “what a child can do with assistance today she will be able to do by herself tomorrow” (p. 87), thus the learning process fosters the developmental process, revealing “how external knowledge and abilities in children become internalized” (p. 91). With an emphasis on the learning process rather than the product, Vygotsky asked learner-centered questions such as “what did they do under varying task conditions?” and “how did they seek to meet task demands?” (Driscoll, 1994, p. 242). His experiments led to his mediational view of development, which Cole and Scribner (1978) defined as “the individual actively modifies the stimulus situation as a part of the process of responding to it” (pp. 13-14).

**Construct of Formative Assessment**

A pedagogy utilizing rubrics as formative assessments—that is “supporting learning through the instructional scaffolding, including feedback, and the active involvement of students in the assessment/learning process” (Heritage, 2010, p. 15)—is reflective of the basic tenets of the zone of proximal development and Vygotsky’s mediational view. But rubrics “are not a replacement for good instruction” (Andrade, 2005, p. 29). To wit, rubrics alone cannot act as a proxy for the conceptualized teacher assistance that could move students from an “embryonic state”—the “buds” or “flowers” of development—to the “fruits,” where students broker the boundary of developing capabilities and undeveloped capabilities (Vygotsky, 1978, p. 86). Such assistance is indicative of a learner-centered ideology (Schiro, 2008), at the heart of which is the use of formative assessments to facilitate in improving student performance.
Indeed, the use of rubrics by classroom teachers as instructional tools with formative assessments can counteract familiar laments of the traditional summative assessment process—that it fails to offer guidance, that it promotes superficial learning, that it is subjective, that it promotes a deficit model (Black & Wiliam, 1998). The zone of proximal development conceptualizes Vygotsky’s belief that assessments should be used as instructional tools to develop higher cognitive processes (Edens & Shields, 2015).

Ash and Levitt (2003) conducted two case studies of the zone of proximal development as the “core feature” (p. 23) of formative assessment processes. The researchers took the view that formative assessments create a reciprocal transformative learning opportunity for both students and teachers, with the “cyclic determination of differences between teacher and learner understanding […] interpreted as the ongoing diagnosis of the distance between the learner’s current and potential levels of ability in the zpd” (p. 28).

The first case study followed two teachers in California working with a teacher research group focused on using formative assessments to guide students in science inquiry. The second case study focused on the relationship between a teacher candidate and her mentor, specifically within the context of science inquiry instructional practices. The researchers acted as participant observers in both settings, collecting ethnographic field notes, conducting in-class observations, and interviewing participating teachers. A proposed “trajectory of teacher change” (p. 7) was developed to interpret the researchers Vygotskian view of formative assessments:

1. Teachers examine student work closely, using a prescribed scaffolding tool, i.e., a rubric, as a guideline. Often, this is done in collaboration with other teachers.
2. Teachers begin to see a mismatch (using the scaffolding tool) between their expectation and students’ level of performance, either in competence, proficiency, ability or overall grades.

3. Teachers self-reflect, observe their own practice, and begin to adjust pedagogy, either by changing the requirements of the task, by providing specific student guidance, or by reevaluating the goals of the task.

4. Teachers continue to self-reflect and adjust or constrain tasks and/or expectations according to the student’s ability. This helps students move towards desired conceptual goals, but also moves the teachers toward increasing sophistication in diagnosing and understanding their own pedagogy. (p. 31)

The researchers concluded that transformative learning for both teachers and students was the outcome of a pedagogy rooted in formative assessments. Teachers and students engaged in an “ever-changing zone of mutual understanding” (p. 41) revealing the expanding upper limits of the zpd. Particularly germane to this study was Ash and Levitt’s (2003) observation that the teachers’ “ever-increasing fine-tuning of scaffolding tool use”—that is, their revision of standard rubric use—within the context of formative assessments “led the learner to move towards more independent understanding” (p. 42). The researchers explicitly recommended research in the role of tools—such as rubrics—in formative assessments.

**Historical Context for the Use of Rubrics for Writing Assessments**

Historical factors such as industrialization, immigration, and exponential population growth in cities in the first half of the 19th century provided the early impetus for contemporary writing rubrics (Wilson, 2006). With increased access to high school education in population centers, colleges sought to use writing assessments in the admissions process to distinguish
prospective students’ abilities. But with the growing number of applicants, the need to standardize the assessment process to foster both fairness and efficiency led to early incarnations of rubrics like Hill’s “English Composition Card” (p. 16), which focused exclusively on traditional grammatical competencies, and the use of Hillegas’s “Composition Scales” (p. 19), in which raters would compare student writing to a group of pre-ranked papers to determine a score.

By the middle of the 20th century, with the growth in popularity of standardized testing, the College Board commissioned the Educational Testing Services to examine the reliability of then-current grading methods of “General Impressions Marking” (Cooper & Odell, 1977, p. 5). Similar to the earlier use of “Composition Scales” (Wilson, 2006, p. 16), raters would initially meet as a group and discuss grading standards and rate sample papers until an acceptable degree of consensus was reached (Diederich, 1974, p. 55). Diederich, French, and Carlton (1961) rationalized in their seminal study what many educators and researchers still grapple with today: “Grading essays for writing ability, even under the most carefully controlled conditions, is extremely unreliable” (p. 1). The researchers sought to understand why. Using factor analysis of the 53 raters’ comments on 300 student papers, the researchers established groups of adjectives—or “rubrics”—that were logically comparable to one another (i.e., ideas, form, mechanics) from which the domains of today’s standardized writing rubrics emerged (Diederich, 1974; Diederich et al., 1961).

By the late 1970s, raters of the Advanced Placement English exams and the New York State Regents Writing exam utilized rubrics in the grading process (Cooper & Odell, 1977; Dirlam, 2019), establishing a standardized method of assessing student writing that has remained relatively unchanged since. However, the increased emphasis on curriculum and instruction in
recent years, particularly the popularity of backwards planning (McTighe & Wiggins, 2005), has expanded the role of rubrics from merely assessment instruments to instructional tools to be used by teachers and students in the classroom.

**Recent Studies Concerning Rubrics as Formative Assessments**

In a narrative content analysis, Panadero and Jonsson (2013) focused on if and how rubrics have an impact on student learning in formative contexts. The researchers sought to determine in what ways the formative use of rubrics mediated improved student performance, and which factors may moderate—either positively or negatively—the effects of using rubrics. The 21 studies that these authors culled from the initial literature search met specific criteria: They included empirical data on the use of rubrics for formative purposes and they were published in peer-reviewed journals. A qualitative approach to data analysis (in lieu of a meta-analysis) was a consequence of a dearth of studies on the topic, research questions that are focused on identifying effects rather than the relative strength of those effects, and the variety of approaches in the studies.

Data analysis involved parsing pertinent information from close readings of the studies in a data matrix and then reviewing the findings for patterns. Most of the studies investigated the effects of rubrics for self-assessment purposes. The primary educational benefits derived from the content analysis include increased transparency, reduced anxiety, improved feedback, and student self-efficacy. While Panadero and Jonsson (2013) concluded that the formative use of rubrics can “mediate” (p. 140)—a Vygotskian word—improved student performance, they found no studies on the effect of rubric design (and thus recommended such).

One such study from Panadero and Johnson’s (2013) content analysis was conducted by Andrade, Du, and Wang (2008), who investigated the effects of using writing rubrics in a
formative context, specifically as a self-assessment tool. Guided by research questions that focused on the effect of teachers providing models, collaboratively generating criteria assessment criteria with students, and having students engage in rubric-referenced self-assessment, the researchers measured the effects of such interventions prior to and during the composition of a writing assessment.

The study, a quasi-experimental design with a control group, included 166 third and fourth grade students with four classrooms in the treatment and three classrooms in the control group. The treatment group received a written rubric to self-assess their first drafts, using colored pencils to annotate parts of their writing that aligned with rubric criteria. A general linear model (GLM) two-way analysis of variance (ANOVA) was used to analyze the effect of treatment and gender. The analysis determined a significant effect for the use of rubrics on student scores. The treatment group’s writing scores ($M = 28.5$, $SD = 4.9$) were higher than the comparison group’s scores ($M = 24.3$, $SD = 4.7$), $F(1, 111) = 18.9$, $p = .000$, partial $n^2 = .15$.

Girls tended to have somewhat higher essay scores ($M = 27.7$, $SD = 5.1$) than boys ($M = 25.8$, $SD = 5.2$) but not to the degree of statistical significance, $F(1, 111) = 1.9$, $p = .17$.

An interesting finding was that the treatment had a significant effect on all the criteria in the rubric (i.e., ideas, organization, paragraphs, voice, words) except sentences and conventions, the only criterion not self-assessed by students. The researchers attributed this outcome to the effect of formal assessment—that is, in this case, students self-assessing drafts of their writing using a rubric that was reviewed collaboratively in class. The practical implications for Andrade et al. (2008) are clear: Engaging collaborative analysis of writing evaluation criteria and self-assessing works in progress using that criteria have a positive effect on student performance.
The efficacy for the use of rubrics in instruction and assessment, particularly with struggling students, is reinforced in a study of first and second graders writing opinion essays with and without rubrics during the writing process (Bradford et al., 2015). The researchers sought to determine how the use of rubrics influences the overall quality of early elementary school student writing and to what degree rubric use improved students’ attitudes towards writing. This counterbalanced, repeated measures design included 32 students from a low socio-economic, racially and ethnically diverse elementary setting designated a School in Need of Assistance. After a pretest to evaluate baseline writing scores, the initial treatment group used the assessment rubric during writing lessons about the target criteria of focus and idea generation, organization, voice, word choice sentence fluency, and conventions. The control group covered the same criteria in lessons without the rubric as an instructional instrument. A midpoint test was then administered before the groups switched conditions. The researchers wanted to see how much the initial control group’s writing performance would improve after also receiving the treatment. At the conclusion of the study, the researchers, in an attempt to gauge student attitudes towards rubric use, had both groups respond to the following question: “In your opinion, is it best to use a rubric when writing or not? Why?” (Bradford et al., 2015, p. 466).

Pretest results verified that students in both groups scored very similarly with mean scores of 9.8 and 9.6 points out of 24 possible points, indicating both groups were at the same skill level. Results of the midpoint test before the treatment and control group switched conditions revealed a significant difference ($p < 0.002$) on student scores for those who had been exposed to a rubric during instruction. The treatment group had a mean score of 18.5 points while the control group—who had not been exposed to the rubric—had a mean score of 14.9 out
of a possible 24 points. With a large effect size (Cohen’s $d = 0.92$) on student scores when providing student access to and instruction with the rubric prior to the final assessment, Bradford et al. (2015) concluded that “using a rubric, affected student writing skills in an easily noticeable and useful way” (p. 468). Posttest scores of the group who had received the treatment after the midpoint test reinforce this conclusion. Their mean score rose from 14.9 to 18, and effect size for both groups from pretest to posttest (Cohen’s $d = 2.08$) revealed rubrics used during instruction had a large impact on student performance.

While student attitude ratings showed no statistically significant difference between student attitudes based on rubric use, student responses supporting their ratings provide a more nuanced perspective. The general tenor of the most frequent comments from the group who did not initially use the rubric was “relief in being able to complete a task they viewed as arduous” and pride in the effort put forth (Bradford et al., 2015, p. 470). Conversely, students who initially used the rubric felt more positive about their writing skills and the capacity for the rubric to help them perform the task successfully. The researchers suggest future studies extend their work to other educational contexts—particularly different subject areas and populations.

Acknowledging the increasing use of rubrics as both instructional and evaluative tools, Andrade and Du (2005), citing Brookhart’s (2003) seminal essay on the importance of formative assessments to student learning, sought a more nuanced understanding of how students use rubrics. The researchers, as part of a broader study of undergraduates’ responses to criterion-referenced self-assessment, reported on how assessments can promote achievement and provoke learning behaviors associated with academic success. This qualitative study included 14 undergraduate teacher-education students divided into four focus groups segregated by gender. Students utilized a rubric they collaboratively created in an educational psychology course to
complete assessments and discussed their perceptions about rubrics and the role rubrics play in their learning.

Consensual Qualitative Research (CQR) methodology was used to analyze interview data. The CQR protocol was initiated by coding domains followed by identifying core ideas, then culling confirmatory evidence from the transcripts, charting the results in a data matrix, and finally composing a narrative summary. Students cited clarity of expectations, specificity of feedback, and perceived veracity of grades as the primary benefits. Andrade and Du (2005) suggest future research should be conducted on students’ actual use of rubrics rather than their reported use, as well as their conceptions and misconceptions of rubrics. Interestingly, and germane to this study, is that while students reported using the rubrics during the planning process, some admitted they did not read all the levels of quality. This result is consistent with Ormond and Merry’s (1996) conclusion that students might lack the knowledge and skills to precisely measure their performance against the criteria of a rubric. This study seeks to explore if the use of single-point rubrics could compensate for this limitation.

To measure the effectiveness of single-point rubrics, Fluckiger (2010) engaged in a collective case study approach, identifying recurring themes in 10 action research reports dealing with the practical application of single-point rubrics in the classroom. Citing the theoretical constructs of Vygotsky’s (1978) zone of proximal development and self-regulated learning (SRL), Fluckiger (2010) identified the following recurring themes:

1. The quality of work by students using a single-point rubric was equal to or better than work completed using a scaled rubric.
2. Students perceived that they learned more using single-point rubrics.
3. Students revealed higher engagement.
4. Students acknowledged less student anxiety.

5. Students revealed a better understanding of the traits of good writing.


Fluckiger (2010) concluded that single-point rubrics are an effective tool for qualitative feedback in a formative setting where students utilize the rubric for self-assessment throughout the writing process.

**Reliability and Validity of Rubrics**

Arguing that most teacher-generated classroom assessments fail to meet traditional definitions of validity and reliability, Brookhart (2003) promoted the use of formative assessments as the integration of evaluation and instruction, recommending that teachers make “the test part of the environment itself and part of the fabric of students’ learning” (p. 7), thus establishing construct-relevant validity and reliability (Gronlund, 1998) that allow instructional decisions based on student performance to be made in real time while preserving instructional goals. Indeed, Balch et al. (2016) propose that evaluation criteria—even rubrics for rubrics—be established to address validity and reliability.

Brookhart’s (2003) assertion is that psychometric methodologies for testing validity and reliability, primarily established in large-scale assessment contexts, do not meet the needs in the classroom. Validity concepts in large-scale assessments focus on the performance of students as “subjects” in a standardized context (“construct irrelevant”), and reliability is the consistency of scores over irrelevant factors (p. 9). In contrast, Brookhart (2003) advocates for a new set of validity and reliability concepts in classroom assessments that apply a formative approach. Rather than the students as “subjects,” they are “observers jointly with teachers,” and the validity goal of the assessment is “an understanding of how students’ work compares to the ‘ideal’ work
(as defined in the learning objectives) and/or effective use of that information for further learning,” and the sufficiency of such information determines the assessment’s reliability (p. 9).

Concerns regarding the validity and reliability of traditional rubrics extend to whether rubrics can accurately assess performance at all. Labeling rubrics “the tool of submission,” Farenga, Ness, and Sawyer (2015, p. 13) argue that rubrics “often fail to operationalize a set of standards or objectives” due to poor design. Mabry (1999) perceives the same shortcomings as Brookhart (2003), asserting traditional psychometric concepts and practices—like a rubric—undermine new ideas and techniques in assessment. More alarming is Mabry’s (1999) claim that the purpose of writing rubrics—to promote assessment reliability by standardizing scoring—also promotes, as an unintended consequence, the standardization of writing. Consequently, construct validity is undermined rather than reinforced by reliability, as the pre-specified criteria of a rubric (standardized domains) work against self-expression—that such rubrics are “testing not the construct of writing achievement but the construct of compliance to the rubric” (p. 678).

Jonsson and Svingby (2007) conducted a comprehensive article review study to investigate whether empirical evidence exists of increased validity and reliability when using rubrics as assessment instruments. The rationale for the study is the acknowledged shift from summative to formative assessments in the classroom operationalized through more authentic performance tasks. The researchers assert that classroom assessments do not require the same high degree of reliability, but that validity and some degree of scoring consistency—even if it does not meet the traditional 70% agreement threshold (Stemler, 2004)—is necessary. Seventy-five papers were culled from peer reviewed journals and dissertation databases that presented empirical evidence on validity and reliability measures of rubrics to determine if rubrics
“enhance the reliability of scoring” and “facilitate valid judgment of performance assessments” (Jonsson & Svingby, 2007, p. 132).

Since intra-rater reliability—that is, the consistency of scores from a single rater—was only measured in seven of the studies with Cronbach’s alpha values consistently above .70 for the rubrics examined, the researchers concluded that intra-rater reliability is not a “major concern” (Jonsson & Svingby, 2007, p. 134). Inter-rater agreement, the consistency of scores among multiple raters, was examined in more than half the studies, with most relying on consensus agreement—the percentage of agreed upon scores between raters—as the most frequently applied statistical method. Few of the studies achieved the .70 threshold (Stemler, 2004) for exact agreement, only reaching good levels of consistency when adjacent agreement, within one score point on a traditional four- or six-point scale, was considered. Cohen’s kappa, which considers how consensus agreements vary from chance agreements when using rubrics with few or dichotomous scoring levels, was applied in some studies with most only reaching fair agreement levels above .20 and below .70. When correlations between raters’ scores were considered, Pearson’s correlation coefficient values often failed to meet .70 thresholds (Gall, Gall, & Borg, 2003). Despite the marginal findings, Jonsson and Svingby (2007) concluded that reliability is “probably” improved when scoring with a rubric (p. 136), yet they question the applicability of such psychometric measurements to classroom assessments.

The researchers do, however, acknowledge the importance of content, construct, and criterion validity and deny, in the absence of empirical evidence, that the use of a rubric “probably” improves these multiple dimensions of validity concomitantly (p. 137). The studies examined by Jonsson and Svingby (2007) often relied on expert opinion to measure internal validity and the correlation of scores to established instruments, such as existing rubrics, to
report external validity, with most reporting “modest correlations from .4 to .6” (p.137).

Nevertheless, just providing a rubric does not, according to the findings, improve “content representativeness, fidelity of the scoring structure to the construct domain or generalizability” (p.137). Jonsson and Svingby (2007) qualify this conclusion by positing that if rubrics positively impact instruction, consequential validity—only explored in two of the studies—could improve.

More recent studies have examined the application of validity and reliability measures to instructor-designed rubrics. Angell (2015) sought to establish reliability and validity for a citation rubric to be used when teaching and assessing research skills. The researcher determined through Likert scale survey scores of 42 subject-matter experts and a Cronbach’s alpha score of .514 (measuring the internal consistency of rubric item mean variances to overall mean scores) that content validity and internal consistency results warranted further improvements to the rubric. With similar purposes, Van Helvoort, Brand-Gruwel, Huysmans, and Sjoer (2016) utilized percentage agreement and intra-class correlation to measure inter-rater reliability and factor analysis, reliability analysis, and Pearson correlations to assess validity of an instructor-designed information literacy rubric. The results of 80 student papers graded by two qualified raters suggested the scoring rubric was a valid and reliable instrument. Inter-rater reliability was good for most of the criteria examined using intra-class correlation, except one domain was deemed vulnerable to rater subjectivity. High correlations between overall scores with an existing assessment rubric supported validity conclusions, along with the premise of careful deliberate design by instructional experts during the construction of the rubric.

Positing that the evaluation process is highly subjective and that analytic rubrics have not been adequately studied for validity and reliability, Rezaei and Lorvorn (2010) conducted an experimental study to investigate the extent to which writing rubrics help raters focus on
reasoning, content knowledge and logic instead of command of writing mechanics such as grammar, punctuation, and style. The researchers’ rationale for the study questions the oft-accepted premise that the use of rubrics provides objectivity to a scoring process that has been found to be influenced by factors such as gender, handwriting, and even physical attractiveness of the student.

A standard writing rubric with frequently utilized domains (i.e., understanding, support, organization) was provided to 326 college graduate students divided in four groups, two of which consisted of education majors and two of which consisted of business majors. The analytic writing rubric was weighted to value a logical, in-depth response to the writing prompt more than writing mechanics. Copies of two sample essays were provided to the different groups of raters to evaluate individually—one that deserved a high score based on the rubric despite poor mechanics and penmanship because it answered the question logically, specifically, and comprehensively, and one that deserved a low score despite its command over style and conventions because it did not fully address any part of the task. One group of education majors and one group of business majors would grade each essay using the rubric while the other two groups would rate the essay without using a rubric. The hypotheses of the study, based on a review of the literature according to the researchers, were that the rubric would increase the reliability and validity of the assessment process through lower variability in the scores, and that education students would be less influenced than business students by the poor mechanics in the more exemplary response due to their experience with rubric-based assessments.

The results of the study did not support the hypothesis that the rubric would lower the variability of scores and increase reliability and validity. The participants were strongly influenced by the poor mechanics and penmanship in the otherwise exemplary written response.
Particularly surprising was that both groups of education students, who had greater familiarity with the use of rubrics in the evaluation process, rated the essay that did not answer the question with a higher mean score than the essay that contained only superficial flaws.

Rezaei and Lorvorn (2010) concluded that raters are ultimately influenced by their intuitive impressions of student writing, even when provided with a rubric, and that general familiarity with a rubric is not enough. Special training is needed for using rubrics with specific assignments to not only increase the reliability but also construct and criterion-based validity. Germane to this study is the researchers’ belief that a limiting factor in the study was the broadness of the rubric used. They recommend further research utilizing a rubric with “more detailed descriptors” (Rezaei & Lorvorn, 2010, p. 28) like, perhaps, the single-point rubric used in this study.

**Chapter Summary**

This review of literature has grounded this study in the theoretical constructs of Vygotsky’s (1978) zone of proximal development and formative assessments. Seminal works and research support the use of the ZPD and formative assessments within the present study. Recent studies offering empirical evidence on the use of rubrics to support these constructs were used to ground this study. Theoretical discussions and empirical research regarding the validity and reliability of rubrics as assessment instruments was used to inform the research design, methodology, and conclusions.
CHAPTER THREE: METHODOLOGY

This chapter provides a thorough and detailed description of the methodology used to establish this study’s validity and generalizability. Sections are included to describe the processes and procedures followed during this study. The researcher’s biography provides background information to establish credibility of the researcher and contextualize the genesis of this study. The research questions and hypotheses guiding this study are then provided, followed by the explanations of the setting, participants, and sampling procedures. The research design, including instrumentation, the procedures and timeline, data collection, and a description and justification of the analyses, are all provided. This chapter concludes with a discussion of the limitations of the study. All of these sections were included for the benefit of future research of rubric structure.

Researcher Biography

David Popken teaches high school English in Westchester County, New York. He holds a Bachelor’s degree and a Master’s degree in English Education from the State University of New York at Oneonta, and this research study is the culmination of his pursuit of a doctoral degree in Instructional Leadership from Western Connecticut State University. Much of Dave’s work as a practitioner and teacher-leader has focused on the teaching and learning of writing, formative assessments, and rubrics.

The increased popularity of Advanced Placement (AP) courses has led to an increased heterogeneity of the student population in the AP English courses Dave teaches. Thus, the challenge he has faced is how to differentiate instruction to help all students meet the curricular objectives of the course, particularly as writers. Single-point rubrics have emerged as a potential instructional and evaluative instrument to meet these needs. The research skills fostered through
formal doctoral coursework focused on research design, implementation, and data analysis, along with twenty-five years as a high school English teacher, provided the foundation for Dave to conduct this research study.

Dave’s other interests include the challenge of reconciling progressive educational concepts (e.g., critical and creative thinking, collaboration, problem solving), often labeled 21st Century Skills (National Education Association, 2016), in a traditional public educational setting with predictable and seemingly contradictory external standards and expectations.

**Research Questions and Hypotheses**

This quantitative study was used to establish criteria for sampling, design, and statistical analyses primarily to establish validity and reliability of a teacher-generated single-point writing rubric. Student-performance using the single-point writing rubric during the writing process was correlated with student-performance using a scaled analytic rubric that has pre-established validity and reliability. Course enrollment, gender, overall GPA, self-reported single-point rubric use, and self-reported student perceptions of single-point rubric efficacy were also included as predictors in the correlation. Ratings of student performance from a single-point rubric were utilized to determine intra and inter-rater reliability of the criteria from the single-point rubric. This study addressed the following questions.

1. To what degree and in what manner do scaled analytic rubric scores, course level, gender, GPA, self-reported single-point rubric use, and self-reported student perceptions of single-point rubric efficacy predict writing scores for students who were provided a single-point writing rubric during the writing process?

Non-directional hypothesis: Scaled analytic rubric scores, course enrollment, gender, overall GPA, self-reported single-point rubric use, and self-reported student
perceptions of single-point rubric efficacy will significantly predict writing scores for students who were provided a single-point writing rubric during the writing process.

2. What is the inter-rater reliability of a single-point writing rubric?

Non-directional hypothesis: There will be a significant degree of inter-rater reliability of a single-point rubric.

**Description of Setting and Participants**

The following sections describe the setting, subjects, and sampling procedures utilized for this study.

**Description of the Setting**

This convenience sample was derived from a small, suburban, relatively affluent (about 10% qualify for free or reduced-price lunch) public middle/high school in New York. The overall population of the school at the time of the study was about 650 students in grades 6-12 (about 50% male and 80% White). Graduation rates at the setting hover near 100% annually, and 90% of students attend four-year colleges immediately after high school. A convenience sampling method was utilized for several established reasons in educational research (Gall et al., 2003), including researcher access, familiarity, and administrative approval.

The setting chosen is both purposive and convenient—the geographic location, socioeconomic makeup of the school, the characteristics of student groupings, and the students’ educational experiences with rubrics are reflective of the study’s design, and the researcher was granted access to the participants in this school.

Noteworthy is the setting’s English department policy for enrollment in an Advanced Placement English course that has led to high percentages of the population—almost 70% in the sample population—taking Advanced Placement English. Based on a point system, the policy
purports that students are required to earn a minimum of two out of a possible four points: 1-point for a 90% or above average in English the prior academic year; 1-point for scoring a 90% or above on the New York State Common Core English Regents Examination; 1-point for receiving a teacher recommendation from the student’s previous English teacher; 1-point for “testing in” by earning a predetermined minimum score on sample sections from an Advanced Placement sample examination. Nonetheless, the high percentage of students taking AP English can be attributed to several global and local factors: the increasing percentage of students taking Advanced Placement courses nationally (College Board, 2017); the English department’s almost exclusive use of formative assessments and subsequent revisionary and reflective addendums to those assessments (e.g., test corrections; essay revisions) that result in higher student grade point averages in English; an academic culture that promotes high achievement coupled with an inclusive approach.

**Description of Participants**

There were 86 students in four twelfth grade English classes and all students were potential participants in the study. This convenience sample was chosen because they met the basic criteria of the study (i.e., they were enrolled in a senior English Language Arts course). The 86 students in the accessible population were members of four classes, two of which were Advanced Placement English courses, and two of which were general education English courses (“English 12”), as noted in Table 1.
Table 1

*Student Enrollment in Each Class of the Accessible Population*

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>English 12, Period 1</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>AP English, Period 4</td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>AP English, Period 6</td>
<td>29</td>
<td>34</td>
</tr>
<tr>
<td>English 12, Period 7</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>Total (All Groups)</td>
<td>86</td>
<td>100</td>
</tr>
</tbody>
</table>

As reported earlier, the accessible student population was not racially/ethnically diverse. The gender demographics in the school building (grades 6-12) was approximately equivalent between males and females. The race/ethnicity of the accessible student population (grade 12) is presented in Table 2. These data were used to describe the sample and were not used for analysis purposes.
Twelfth graders were chosen as the population of interest based on specific criteria and the data they can provide. First, all of these twelfth graders were likely to have been exposed to the myriad rubric structures and instructional approaches with rubrics, and their experiences were inherently representative of the varied and sometimes piecemeal assessment experiences most students have moving through K-12 education in United States public schools. Second, the make-up of the groups (i.e., Advanced Placement and general education English classes) was conducive to analyzing the impact of the rubric structure on student writing performance with distinct student populations.

As a consequence of aforementioned educational trends, the percentage of students enrolled in AP English each year in this setting was relatively high, sometimes approaching 70% of an entire grade level. While teachers and administrators celebrated this culture of high achievement, there was significant concern for the homogenous nature of general education senior English courses. These general education courses were composed of primarily low
performing students, students with low motivation, at-risk students, and students who receive special education accommodations. Also noteworthy was the high percentage of male students who populated these classes. Gender with respect to course enrollment is presented in Table 3. Through the analysis of the relationship of course enrollment, GPA, and student survey feedback with student performance using a single-point rubric, this researcher hoped to exploit this disparity in achievement levels for the benefit of all students.

Table 3

*Gender with Respect to Course Enrollment in the Accessible Student Population*

<table>
<thead>
<tr>
<th>Course</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Percentage</td>
<td>Frequency</td>
</tr>
<tr>
<td>English 12</td>
<td>20</td>
<td>23</td>
<td>7</td>
</tr>
<tr>
<td>AP English</td>
<td>25</td>
<td>29</td>
<td>34</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td>52</td>
<td>41</td>
</tr>
</tbody>
</table>

*Program Completion Rate*

Of the 86 students in the accessible population, 66 students opted to participate in the study, providing assent (Appendix C) and consent documentation (Appendix D). Of the 66 participants, complete data were collected from 64. Two participants were absent during data collection and were omitted from the study. Tables 4 and 5 report participant demographics for the cleansed sample.
Table 4

*Race/Ethnicity of Sample Population*

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Asian</td>
<td>1</td>
<td>1.60</td>
</tr>
<tr>
<td>Hispanic</td>
<td>7</td>
<td>10.90</td>
</tr>
<tr>
<td>White</td>
<td>55</td>
<td>85.90</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>1.60</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>64</strong></td>
<td><strong>100.00</strong></td>
</tr>
</tbody>
</table>

Table 5

*Gender with Respect to Course Enrollment in Sample Student Population*

<table>
<thead>
<tr>
<th>Course</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Frequency</td>
<td>Percentage</td>
<td>Frequency</td>
</tr>
<tr>
<td>English 12</td>
<td>8</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>AP English</td>
<td>22</td>
<td>34</td>
<td>29</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>30</td>
<td>47</td>
<td>34</td>
</tr>
</tbody>
</table>

**Research Design**

This quantitative correlational study utilizing a convenience sample sought to predict student performance using single-point rubrics (criterion variable, interval data) based on the following predictor variables: student performance using scaled analytic rubrics (interval data),
Description of Analyses

A stepwise multilinear regression (Meyers, Gamst, & Guarino, 2006) was utilized to determine if the non-directional hypothesis for research question one was supported by the data. SPSS 25 was used to determine if scaled analytic writing rubric scores, course level, gender, GPA, self-reported single-point writing rubric use, and self-reported student perceptions of single-point writing rubric efficacy predict writing scores for students who were provided a single-point writing rubric during the writing process.

To determine if the non-directional hypothesis for research question two was supported by the data, Pearson’s correlation coefficient (Gall et al., 2003) was conducted to establish inter-rater reliability of overall writing scores using a single-point writing rubric. Additionally, item analysis of each item on the single-point writing rubric (19 in total) was measured to determine inter-rater scoring agreement using traditional inter-rater agreement percentages (Fraenkel & Wallen, 2003) as well as the non-parametric Cohen’s kappa (Stemler, 2004) to account for the small sample size (Pallant, 2016).

Instrumentation

The following sections discuss the instruments administered for quantitative data collection.
**Student demographic survey.** Student participants completed a demographic survey (Appendix E) identifying racial/ethnic status, gender, and course enrollment (AP English or general education English) for descriptive statistics. The five-minute survey was administered by the classroom teacher at the conclusion of the administration of the writing task, outside of instructional time.

**Writing task.** Students completed two separate New York State Regents Examination in English Language Arts writing tasks (Appendices F and G), used with permission (New York State Education Department, 2019), two weeks apart. Students used a scaled analytic writing rubric as a writing guide for the first task (Appendix A) in the pre-writing, writing, and editing phases of the writing process. Students then used a single-point writing rubric for the second task (Appendix B) in the pre-writing, writing, and editing phases of the writing process. Each text-analysis response required students to identify a central idea in a prose text and analyze how the author’s use of one writing strategy develops the central idea. The New York State Education Department (2018) test development process includes reviewing standards during item development, field testing, and item analysis for reliability and generalizability.

**Scaled analytic rubric.** The scaled analytic rubric (Appendix A) provided to participants during the pre-writing, writing, and editing phase of the first task is a New York State Regents Examination in English Language Arts rubric. The rubric measures four criteria (Content and Analysis; Command of Evidence; Coherence, Organization, and Style; Control of Conventions) across four gradations with a score of four as the highest possible score for each criterion. Scoring reliability of a 2014 field test was measured using Pearson’s correlation coefficients. Results for all writing tasks on that exam “ranged from 0.79 to 0.94, indicating a fair to high degree of reliability” (New York State Education Department, 2014, p. 5). Inter-rater
agreement was also measured, determining 100% of first and second scores (each essay is read and scored by two different readers) were aligned (e.g., a score of 3 by the first reader and a score of 3 by the second reader) or were within one point of one another (e.g., a score of 4 by the first reader and a score of 3 by the second reader).

While this rubric was initially developed as a holistic rubric for this summative assessment, it is also reflective of the broader skills and techniques necessary to be successful on a text-analysis writing assessment, regardless of the topic or scope of the task. The assumption and experience of the researcher are that in practice many teachers use this rubric, or one similar in structure and criteria, as an analytic rubric to prepare their students for standardized assessments.

**Single-point instructional rubric.** The single-point rubric (Appendix B) used by participants during the pre-writing, writing, and editing phase of the second task is a researcher-generated writing rubric that is not only task specific but also, like the Regents exam scaled analytic rubric, reflective of the broader skills and techniques necessary to be successful on a text-analysis writing assessment, regardless of the topic or scope of the task. Since this task requires that students identify a central idea and writing strategies used to develop the central idea, the rubric reflects the criterion of a successful literary analysis essay but is generic enough that the same rubric could be applied to a similar task with a different text, or a similar task with a different scope (e.g., rhetorical analysis). In this sense, the rubric, similar to traditional scaled analytic rubrics, is an instructional tool to promote transcendent skills rather than solely to evaluate performance on an isolated task.

The single-point writing rubric utilized in this study was designed by classroom teachers in the study’s setting who were attempting to increase clarity, objectivity, and efficiency in
writing instruction and evaluation for increasingly heterogeneous (i.e., academically) student populations in Advanced Placement English courses identified in the research study sample description. Like traditional single-point rubrics, the single-point rubric in this study measures one level of performance—proficiency. Unlike most descriptions of single-point rubrics, this rubric offers no additional space for teacher comments. Instead, criteria are listed and described in a linear format (unlike the traditional matrices of scaled analytic and holistic rubrics) and teachers reserve comments for the margins of papers. Domains are structural (e.g., introduction; body; conclusion) rather than conceptual (like in those listed above in traditional scaled rubrics). The criteria listed in the domains (e.g., transitions; quotation context; synthesis of sources) are intended to be concomitantly instructive and evaluative for teachers and students. While the single-point rubric had the option of 0 points, it also weighs some of the criteria as 1, 2, or 3. For instance, a criterion with a value of 3 indicates that this is the only option other than 0. A range of ratings from 0 to 3 for each criterion are not to be employed in this type of rubric.

Like most teacher-generated instruments and assessments, the single-point rubric utilized in this study has not been tested for validity and reliability. Teachers who designed, refined, and used the single-point rubric loosely correlated student achievement on standardized tests deemed valid and reliable by psychometric testing with the use of the single-point rubric as a teaching and assessment instrument. While this is consistent, relatively speaking, with the determination of construct and content validity of rubrics, most often reliant on expert opinion and correlations to other measures (Jonsson & Svingby, 2007), the primary objective of this study is greater empirical evidence of validity and reliability for this instrument.
Procedures

The first phase of data collection took place over three 42-minute class periods with the same teacher in each class period. During the first class period, the teacher introduced and discussed the task and the scaled analytic rubric with each group. Students were provided time to complete any pre-writing activities (e.g., annotating, outlining, brainstorming). The second class period was designated an in-class writing period under the supervision of the teacher. Students were reminded that they should use the scaled analytic rubric as a guide to successfully complete the task. During the third and final class period for phase one of data collection, the teacher had students complete, review, and edit their drafts using the scaled analytic rubric as a guide. The classroom teacher also administered a survey after students completed the task. A detailed outline of the procedure was provided to the teacher to avoid experimenter bias (Gall et al., 2003).

In phase two, blind scoring of all four classes of essays was conducted by two members of the setting’s English department other than the researcher or the classroom teacher administering the instruments. Those teachers, experienced participants in Regents grading protocols, nonetheless completed the prescribed Regents task evaluation training to establish inter-rater reliability (Jonsson & Svingby, 2007), which includes reviewing the task and rubric, scaled anchor papers, and practice papers. The New York State English Language Arts Common Core Regents Exam rubric (Appendix A), which has pre-established validity and reliability, was used to evaluate student performance. All essays were read by two different scorers and evaluated on a scale of 1-4 for each criterion domain (Content and Analysis; Command of Evidence; Coherence, Organization, and Style; Control of Conventions). In the event the two readers awarded different scores, the scores were averaged (e.g., a 4 and a 5 would be entered as
a 4.5). In the event that scores were more than 1-point apart, a third reader would adjudicate the score, consistent with NYS Regents scoring protocol. It was never necessary to recruit a third reader, consistent with NYS inter-rater agreement results for writing. Scores for each of the four criteria were then added together, the sum was divided by the total possible points (i.e., 16), and after the product was multiplied by 100, a score was determined. This scoring conversion is consistent with common classroom practices since rubric scores often need to be converted to numerical percentages for grade reporting procedures (Depka, 2019; Mertler, 2000; Quinlan, 2012).

During phases three and four—approximately two weeks after the administration of the first task—a second task (Appendix G), identical in form and structure (a NYS Regents text-analysis response) but utilizing a different prompt (i.e., text) was administered to the same student participants by the same teacher. During this phase, students used a single-point writing rubric (Appendix B) rather than the scaled analytic rubric during the writing process. Teachers evaluating student essays also utilized the single-point writing rubric during the grading process. The entire protocol for phases three and four was identical to those of phases one and two except that students were provided a similar but not identical task, students and teachers utilized a single-point writing rubric rather than a scaled-analytic writing rubric, and the survey at the conclusion of the task only asked two questions about rubric use; demographic questions were not repeated (Appendix E).

Timeline for data collection. The following procedures were conducted according to the proposed timeline.

1. Submit application to Institutional Review Board (IRB), December 3, 2018
2. Obtain approval from Institutional Review Board (IRB), meeting date December 12, 2018
3. Obtain Superintendent and Principal consent (see Appendices H and I), January 2019
4. Distribute and collect teacher consent forms (see Appendix J), January 2019
5. Distribute and collect student assent and parent consent forms (see Appendices C-D), January 2019
6. Training for classroom teacher administering writing tasks, February 2019
7. Participants complete writing tasks and surveys, February-March 2019
8. Training for inter-rater reliability for teacher-participants evaluating student writing; evaluation of student writing, March 2019
9. Data collection of student GPAs, April 2019
10. Analyze data and complete report

**Statement of Ethics**

Permission to participate in this research was sought from the district’s superintendent (Appendix H), school principal (Appendix I), and all participating teachers (Appendix J). Consent from parents (Appendix D) and assent from students (Appendix C) was acquired in writing prior to the study. Participation was totally voluntary and teacher-participants and student-participants could have withdrawn at any time without penalty. To ensure confidentiality, each participant was assigned a coded identification number. The data were locked in a filing cabinet by the researcher and will be maintained there for at least seven years after the findings have been published. A report of the study was made available to those who requested it.
Chapter Summary

The methodology provided for this study serves to explain the processes and procedures followed during its administration. The researcher biography established the researcher’s credibility by explaining his knowledge and experience with the setting, population, and subject matter. The two research questions, their hypotheses, and the data analysis for each question are provided for greater transferability. Setting, sampling, instrumentation, and procedures and timeline were described to provide context, scope, and sequence of data collection and analysis.
CHAPTER FOUR: ANALYSIS OF DATA AND EXPLANATION OF FINDINGS

Two research questions were addressed and analyzed in this chapter. A quantitative analysis has been conducted for research questions one and two to determine if the data have supported or refuted the non-directional hypotheses.

Overview of Study

This section provides a complete and thorough analysis of all data collected for this quantitative, correlational study to establish validity and inter-rater reliability of the teacher-generated single-point writing rubric. Student-performance using the single-point writing rubric (SPR) during the writing process was correlated with student-performance using a scaled analytic rubric (SAR) that has pre-established validity and reliability. Course level (general education or advanced placement), gender, overall grade point average (GPA), self-reported single-point rubric use (Survey Question #1), and self-reported student perceptions of single-point rubric efficacy (Survey Question #2) were also correlated with student performance using a single-point writing rubric. Rater one and rater two essay scores based on the single-point rubrics and item analysis of each rubric component for each of the two raters were also correlated to determine inter-rater reliability and inter-rater agreement of the single-point writing rubric.

Research Questions and Hypotheses

Research Question One

To what degree and in what manner do scaled analytic rubric scores, course level (general education and advanced placement), gender (male, female, transgender, other, prefer not to answer), GPA, self-reported student use of a single-point rubric, and self-reported student perceptions of single-point rubric efficacy predict writing scores for students who were provided
a single-point writing rubric during the writing process? The following non-directional hypothesis was developed: Scaled analytic rubric scores, course enrollment (general education and advanced placement), gender (male, female, transgender, other, prefer not to answer), overall GPA, self-reported student use of a single-point rubric, and self-reported student perceptions of single-point rubric efficacy will significantly predict writing scores for students who were provided a single-point writing rubric during the writing process. This quantitative research question was addressed using the stepwise method of multilinear regression (Meyers et al., 2006) to predict writing scores for students who were provided a single-point writing rubric during the writing process with the primary goal of establishing validity for the single-point writing rubric.

Students completed two similar writing tasks, one using a scaled analytic rubric with pre-established validity and reliability as a guide during the writing process, and the other using a teacher-designed single-point writing rubric as a guide. The essays were blindly evaluated by two teacher raters who have years of experience evaluating student writing with and without rubrics. Demographic data and data regarding whether or not students used the single-point writing rubric as a guide, and to what degree they found the rubric helpful in the writing process were collected from a survey administered at the conclusion of the study. GPA data were retrieved, with student, parent, and administrator approval, from school records.

**Research Question Two**

What is the inter-rater reliability of a single-point writing rubric? The following non-directional hypothesis was developed: There will be a significant degree of inter-rater reliability of a single-point rubric. This quantitative research question was initially addressed using Pearson’s correlation coefficient (Gall et al., 2003) to establish inter-rater reliability of overall writing scores of rater one and rater two using a single-point writing rubric. Outcomes using
Pearson’s tests revealed significant correlations between rater one and rater two overall essay scores using the single-point rubric, supporting instrument reliability.

However, after visual inspection of rater one and rater two scores of student essays using the single-point rubric revealed some widely divergent scores, item analysis to determine inter-rater scoring agreement (Fraenkel & Wallen, 2003) on each component of the single-point rubric (19 in total) was performed using traditional inter-rater agreement percentages as well as the non-parametric Cohen’s kappa (Stemler, 2004) to account for the small sample size (Pallant, 2016) and inherently dichotomous scoring system of the single-point rubric. The literature (Angell, 2015; Rezaei & Lorvorn, 2016; Van Helvoort et al., 2016) supports the analysis of inter-rater agreement on individual items of a rubric to determine specific weaknesses in design that affect reliability.

Data Preparation for Research Question One

To answer research question one, a multilinear regression (Meyers et al., 2006) was utilized to determine if scaled analytic rubric scores, course level (general education or advanced placement), gender (male, female, transgender, other, prefer not to answer), GPA, self-reported student use of the single-point rubric, and self-reported student perceptions of single-point rubric efficacy predict writing scores for students who were provided a single-point writing rubric during the writing process.

Stepwise Multilinear Regression

A stepwise multilinear regression was selected for the analysis of research question one (Meyers et al., 2006). SPSS 25 was used to perform stepwise multiple regression in which each predictor variable—scaled analytic writing rubric scores, gender, course level (general education or advanced placement), GPA, whether or not students reported using the single-point rubric, and
to what degree they found it helpful—was added one at a time to determine the order and importance of those predictors (Meyers et al., 2006) for determining the students’ scores using a single-point rubric. The purpose of analyzing the predictive power of these variables, particularly the scaled analytic rubric with pre-established validity and reliability, is to establish construct validity of the single-point writing rubric.

Data screening process. Demographic and variable data were recorded in a spreadsheet and then imported into SPSS 25 for coding and analysis.

Missing data. Data screening was initially conducted to cleanse the data for accurate data analysis. Of the 66 participants, 64 completed both essays. Two students, code numbers 19 and 40, did not complete the second essay using the single-point writing rubric as a guide. Since the score of this essay was the criterion variable for analysis using multilinear regression in research question one, as well as the basis of inter-rater reliability analysis for research question two, all data associated with these two students were removed from the study.

Codebook. Participants and variables were coded in a spreadsheet before being uploaded to SPSS 25. Each participant was assigned a number, each variable a name and label, and all alphanumeric data were converted numerically. Code and value cleaning were executed to check “the appropriateness of numerical codes for the values of each variable under study” (Meyers et al., 2006, p. 44). A draft execution of multilinear regression revealed negative correlations on nominal data for variables with no implied value (i.e., gender, course level). Additionally, data for survey question one, where value was intended, had an inverse correlation, with data for students who answered “yes” to the question “did you use the single-point rubric as a guide?” applied a lower code value (1) than those who answered “no” (2). Data for students who answered “no” to the question “did you use the single-point rubric as a guide?” should have
received a code value less than those who said “yes” since “a lack or absence of the characteristic of interest” (Pallant, 2016, p. 171) should be coded lower. Thus, these variables were re-coded into new variables. Additionally, a kurtosis value of 2.53 ($SE = .63$) for survey question number two revealed a violation of conventional standards of normality. Students who responded to the question regarding the helpfulness of the single-point rubric during the writing process with a Likert rating of 1-7 were recoded 0, and students who responded with a Likert score of 8-10 were recoded 1. Dichotomization of continuous variables in response to skewed data is a common justification noted in MacCallum, Zhang, Preacher, and Rucker’s (2002) analysis of 110 articles that contained 159 instances of dichotomization. The updated codebook is reported in Table 6.
Table 6

*SPSS Codebook of Demographic Information and Variables*

<table>
<thead>
<tr>
<th>Label</th>
<th>Code Name</th>
<th>SPSS Field</th>
<th>Assigned Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant #</td>
<td>Id</td>
<td>Numeric</td>
<td>1-56</td>
</tr>
<tr>
<td>Sex</td>
<td>GenderNew</td>
<td>Numeric</td>
<td>1 = female</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 = male</td>
</tr>
<tr>
<td>Course</td>
<td>EnglishCourseLevelNew</td>
<td>Numeric</td>
<td>1 = English 12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 = AP English</td>
</tr>
<tr>
<td>Overall Weighted GPA</td>
<td>GPA</td>
<td>Numeric</td>
<td>0-105</td>
</tr>
<tr>
<td>Essay 1-SAR Essay Score</td>
<td>essay1a</td>
<td>Numeric</td>
<td>0-100</td>
</tr>
<tr>
<td></td>
<td>essay2a</td>
<td>Numeric</td>
<td>0-100</td>
</tr>
<tr>
<td></td>
<td>essay2b</td>
<td>Numeric</td>
<td>0-100</td>
</tr>
<tr>
<td></td>
<td>Essay2a-essay2b</td>
<td>Numeric</td>
<td>0-100</td>
</tr>
<tr>
<td></td>
<td>essay2avg</td>
<td>Numeric</td>
<td>0-100</td>
</tr>
<tr>
<td>SQ#1-Use SRP?</td>
<td>useSRP</td>
<td>Numeric</td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>SQ#2-Was SRP helpful?</td>
<td>Helpfursrpnew</td>
<td>Numeric</td>
<td>0 = 1-7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1 = 8-10</td>
</tr>
</tbody>
</table>

(continued)
Table 6

SPSS Codebook of Demographic Information and Variables

<table>
<thead>
<tr>
<th>Label</th>
<th>Code Name</th>
<th>SPSS Field</th>
<th>Assigned Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race/ethnicity</td>
<td>Ethnicity</td>
<td>Numeric</td>
<td>1 = Afr. Amer.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 = Asian</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3 = Hispanic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4 = White</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5 = Other</td>
</tr>
</tbody>
</table>

Visual inspection of the individual raters’ scores of student essays using the single-point writing rubric as a guide, the average of which is used as the criterion variable, revealed that some essays elicited widely divergent scores from the two raters. The standard deviation of mean differences of each essay’s score from rater one and rater two were calculated. Eight participants (11, 18, 23, 32, 41, 45, 58, 61) whose essays had score differences 2 standard deviations from the mean, the “typical definition of an outlier” (Hellerstein, 2008, p. 8), were eliminated from the data. The resulting sample to be used for multilinear regression was 56 participants. Standard deviations of differences from mean single-point writing rubric essay scores of rater one and two are recorded in Table 7.
Table 7

Standard Deviation of Differences from Mean for SPR Scores

<table>
<thead>
<tr>
<th>Student</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Difference</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>96</td>
<td>96</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>2</td>
<td>96</td>
<td>96</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>76</td>
<td>4</td>
<td>.73024</td>
</tr>
<tr>
<td>5</td>
<td>96</td>
<td>100</td>
<td>-4</td>
<td>-.20916</td>
</tr>
<tr>
<td>6</td>
<td>96</td>
<td>96</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>7</td>
<td>88</td>
<td>92</td>
<td>-4</td>
<td>-.20916</td>
</tr>
<tr>
<td>8</td>
<td>84</td>
<td>84</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>9</td>
<td>84</td>
<td>80</td>
<td>4</td>
<td>.73024</td>
</tr>
<tr>
<td>10</td>
<td>86</td>
<td>80</td>
<td>6</td>
<td>.96509</td>
</tr>
<tr>
<td>11*</td>
<td>80</td>
<td>100</td>
<td>-20</td>
<td>-2.08797</td>
</tr>
<tr>
<td>12</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>13</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>14</td>
<td>68</td>
<td>68</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>15</td>
<td>88</td>
<td>80</td>
<td>8</td>
<td>1.19994</td>
</tr>
<tr>
<td>16</td>
<td>96</td>
<td>96</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>17</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>18*</td>
<td>72</td>
<td>92</td>
<td>-20</td>
<td>-2.08797</td>
</tr>
</tbody>
</table>

Note: * = Removed from multilinear regression model due to SD > 2 from mean

(continued)
Table 7

*Standard Deviation of Differences from Mean for SPR Scores*

<table>
<thead>
<tr>
<th>Student</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Difference</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>88</td>
<td>92</td>
<td>-4</td>
<td>-0.20916</td>
</tr>
<tr>
<td>21</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0.26054</td>
</tr>
<tr>
<td>22</td>
<td>84</td>
<td>88</td>
<td>-4</td>
<td>-0.20916</td>
</tr>
<tr>
<td>23*</td>
<td>96</td>
<td>76</td>
<td>20</td>
<td>2.60905</td>
</tr>
<tr>
<td>24</td>
<td>60</td>
<td>76</td>
<td>-16</td>
<td>-1.61827</td>
</tr>
<tr>
<td>25</td>
<td>88</td>
<td>80</td>
<td>8</td>
<td>1.19994</td>
</tr>
<tr>
<td>26</td>
<td>92</td>
<td>92</td>
<td>0</td>
<td>0.26054</td>
</tr>
<tr>
<td>27</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0.26054</td>
</tr>
<tr>
<td>28</td>
<td>72</td>
<td>68</td>
<td>4</td>
<td>0.73024</td>
</tr>
<tr>
<td>29</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0.26054</td>
</tr>
<tr>
<td>30</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0.26054</td>
</tr>
<tr>
<td>31</td>
<td>88</td>
<td>100</td>
<td>-12</td>
<td>-1.14857</td>
</tr>
<tr>
<td>32*</td>
<td>92</td>
<td>72</td>
<td>20</td>
<td>2.60905</td>
</tr>
<tr>
<td>33</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0.26054</td>
</tr>
<tr>
<td>34</td>
<td>96</td>
<td>100</td>
<td>-4</td>
<td>-0.20916</td>
</tr>
<tr>
<td>35</td>
<td>92</td>
<td>96</td>
<td>-4</td>
<td>-0.20916</td>
</tr>
<tr>
<td>36</td>
<td>96</td>
<td>100</td>
<td>-4</td>
<td>-0.20916</td>
</tr>
<tr>
<td>37</td>
<td>96</td>
<td>100</td>
<td>-4</td>
<td>-0.20916</td>
</tr>
</tbody>
</table>

*Note:* * = Removed from multilinear regression model due to SD > 2 from mean

(continued)
Table 7

*Standard Deviation of Differences from Mean for SPR Scores*

<table>
<thead>
<tr>
<th>Student</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Difference</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>84</td>
<td>88</td>
<td>-4</td>
<td>-.20916</td>
</tr>
<tr>
<td>39</td>
<td>96</td>
<td>100</td>
<td>-4</td>
<td>-.20916</td>
</tr>
<tr>
<td>41*</td>
<td>80</td>
<td>100</td>
<td>-20</td>
<td>-2.08797</td>
</tr>
<tr>
<td>42</td>
<td>96</td>
<td>100</td>
<td>-4</td>
<td>-.20916</td>
</tr>
<tr>
<td>43</td>
<td>80</td>
<td>80</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>44</td>
<td>96</td>
<td>100</td>
<td>-4</td>
<td>-.20916</td>
</tr>
<tr>
<td>45*</td>
<td>96</td>
<td>80</td>
<td>16</td>
<td>2.13935</td>
</tr>
<tr>
<td>46</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>47</td>
<td>96</td>
<td>96</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>48</td>
<td>84</td>
<td>84</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>49</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>50</td>
<td>84</td>
<td>88</td>
<td>-4</td>
<td>-.20916</td>
</tr>
<tr>
<td>51</td>
<td>64</td>
<td>72</td>
<td>-8</td>
<td>-.67887</td>
</tr>
<tr>
<td>52</td>
<td>72</td>
<td>68</td>
<td>4</td>
<td>.73024</td>
</tr>
<tr>
<td>53</td>
<td>80</td>
<td>80</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>54</td>
<td>96</td>
<td>100</td>
<td>-4</td>
<td>-.20916</td>
</tr>
<tr>
<td>55</td>
<td>92</td>
<td>100</td>
<td>-8</td>
<td>-.67887</td>
</tr>
<tr>
<td>56</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>.26054</td>
</tr>
</tbody>
</table>

*Note:* * = Removed from multilinear regression model due to SD > 2 from mean

(continued)
Table 7

*Standard Deviation of Differences from Mean for SPR Scores*

<table>
<thead>
<tr>
<th>Student</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Difference</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>96</td>
<td>96</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>58*</td>
<td>72</td>
<td>100</td>
<td>-28</td>
<td>-3.02738</td>
</tr>
<tr>
<td>59</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>60</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>61*</td>
<td>48</td>
<td>72</td>
<td>-24</td>
<td>-2.55768</td>
</tr>
<tr>
<td>62</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>63</td>
<td>96</td>
<td>96</td>
<td>0</td>
<td>.26054</td>
</tr>
<tr>
<td>64</td>
<td>88</td>
<td>100</td>
<td>-12</td>
<td>-1.14857</td>
</tr>
<tr>
<td>65</td>
<td>64</td>
<td>76</td>
<td>-12</td>
<td>-1.14857</td>
</tr>
<tr>
<td>66</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>.26054</td>
</tr>
</tbody>
</table>

*Note:* * = Removed from multilinear regression model due to $SD > 2$ from mean

**Multilinear regression assumptions.** Data analysis to identify violations of assumptions of multilinear regression was performed. Violations of normality, linearity, homoscedasticity, multivariate outliers, multicollinearity, and the presence of suppressor variables could potentially distort multivariate statistical test results (Meyers et al., 2006). Each assumption was addressed.

**Initial screening process.** Frequencies are the most effective analysis for categorical variables as means and standard deviations are relatively meaningless for variables such as gender and course level (Pallant, 2016). Because the kurtosis value of survey question number two violated conventional standards of normality, a common occurrence of data collected
through Likert scales (Wu, 2007), the variable was transformed into a dichotomous variable as well (MacCallum et al., 2002; Pallant, 2016). Because the median Likert rating was an 8, the students who responded to the question regarding the helpfulness of the single-point rubric during the writing process with a Likert rating of 1-7 were recoded 0, and students who responded with a Likert score of 8-10 were recoded 1. Frequencies for initial predictor variables of gender, course level, and responses to survey questions one and two are recorded in Table 8.

Table 8

*Frequencies for Research Question One Categorical Variables*

<table>
<thead>
<tr>
<th>Predictor Variable</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>27</td>
<td>48.20</td>
</tr>
<tr>
<td>Male</td>
<td>29</td>
<td>51.80</td>
</tr>
<tr>
<td>Total</td>
<td>56</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>Course Level</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eng 12</td>
<td>11</td>
<td>19.60</td>
</tr>
<tr>
<td>AP</td>
<td>45</td>
<td>80.40</td>
</tr>
<tr>
<td>Total</td>
<td>56</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>SQ #1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1</td>
<td>1.80</td>
</tr>
<tr>
<td>Yes</td>
<td>55</td>
<td>98.20</td>
</tr>
<tr>
<td>Total</td>
<td>56</td>
<td>100.00</td>
</tr>
<tr>
<td><strong>SQ #2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-7</td>
<td>8</td>
<td>14.30</td>
</tr>
<tr>
<td>8-10</td>
<td>48</td>
<td>85.70</td>
</tr>
<tr>
<td>Total</td>
<td>56</td>
<td>100.00</td>
</tr>
</tbody>
</table>

*Note.* $N = 56$. SQ #1 = survey question one: Did you use the single-point rubric as a guide? SQ #2 = survey question two: To what degree did you find the single-point rubric helpful? (1-7 = Least; 8-10 = Most).
Frequency results revealed that 55 of 56 students replied identically to survey question one (Did you use the single-point rubric as a guide?), making the variable logically meaningless to a correlation study and statistically inappropriate to include in parametric analysis (Pallant, 2016). Thus, survey question one was removed as a predictor variable but maintained as descriptive information.

Descriptive statistics. Descriptive statistics were utilized to analyze continuous variables to provide “basic summary statistics such as mean, median, and standard deviation” (Pallant, 2016, p. 56). Descriptive statistics for predictor variables of scaled analytic rubric writing score, and GPA, as well as for the criterion variable of average score (based on two independent raters) using the single-point writing rubric, are recorded in Table 9.

Table 9
Descriptive Statistics for Research Question One Continuous Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min.</th>
<th>Max.</th>
<th>Mean</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPR</td>
<td>68</td>
<td>100</td>
<td>91.38</td>
<td>10.13</td>
<td>-1.142</td>
<td>.142</td>
</tr>
<tr>
<td>SAR</td>
<td>50</td>
<td>100</td>
<td>84.79</td>
<td>15.19</td>
<td>-0.674</td>
<td>-.627</td>
</tr>
<tr>
<td>GPA</td>
<td>77</td>
<td>101</td>
<td>93.82</td>
<td>5.98</td>
<td>-1.047</td>
<td>.682</td>
</tr>
</tbody>
</table>

Note. N = 56; Single-Point Rubric, SPR = Average score (two raters) using single-point rubric; SAR = score using scaled analytic rubric; Grade Point Average, GPA = overall weighted grade point average.

Normality. The tests for normality required skewness and kurtosis to be determined for each of the predictor variables with continuous data—scaled analytic rubric writing scores, GPA, and data from survey question two (to what degree students found the single-point rubric helpful)—and the criterion variable of students’ writing score using a single-point writing rubric as a guide during the writing process. The skewness and kurtosis values for all variables were
recorded in Table 9. While skewness for student scores using the scaled analytic rubric fell within the conventionally acceptable range of -1 to +1 (Meyers et al., 2006), skewness for single-point writing rubric scores and student GPAs fell just outside the conventional range. However, the results fall well within more liberal interpretations of acceptable skewness of -2 to +2, (George & Mallery, 2010), with the acknowledgement that “many scales and measures in the social sciences have scores that are skewed […] revealing] the underlying nature of the construct being measured” (Pallant, 2016, p. 64). Kurtosis values for single-point rubrics, scaled analytic rubrics, and GPA all fall within the conventionally acceptable level of -1 to +1.

Additionally, no major deviations from normality were observed in the Normal Probability (P-P) Plot of the Regression Standardized Residual in Figure 1.
**Figure 1.** Normal P-P Plot of Regression Standardized Residual

**Linearity.** A scatterplot of standardized residuals (Figure 2) had a “roughly rectangular distribution, with most of the scores concentrated in the center (along the 0 point)” (Pallant, 2016, p. 160), revealing no violations to linearity between the criterion variable and the predictor variables.
**Figure 2.** Scatterplot of Standardized Residuals

**Homoscedasticity.** A matrix of bivariate scatterplots was run to determine the presence of homoscedasticity among the continuous independent variable of average essay scores using the single-point writing rubric, continuous independent variables of scaled analytic rubric scores, and GPA. The absence of a curvilinear relationship in the scatterplots in Figure 3 revealed no homoscedasticity was present.
**Figure 3.** Matrix of Scatterplots of Relationships of Continuous Variables

**Multivariate outliers.** The scatterplot of standardized residuals (Figure 2) did not reveal any outliers with standardized residuals of more than 3.3 or less than -3.3 (Pallant, 2016).

Additionally, a boxplot (Figure 4) of average essay scores using a single-point rubric revealed no outliers.
A Mahalanobis distance test was conducted to determine each case’s Mahalanobis distance for the detection of multivariate outliers (Meyers et al., 2006). The critical value of chi-square distribution (18.47) with an alpha level of $p < .001$ for four degrees of freedom ($df = 4$) was used to evaluate each case (Meyers et al., 2006). No multivariate outliers were revealed because none of the Mahalanobis distance values equaled or exceeded the established chi-square criterion as seen in Table 10.

Figure 4. Boxplot of Average Essay Scores Using Single-Point Rubric
Table 10

*Extreme Values Test for Research Question One (n = 64)*

<table>
<thead>
<tr>
<th>Mahalanobis Distance</th>
<th>Case Number</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Highest</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Lowest</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

**Multicollinearity.** Multicollinearity is the strong correlation of two predictor variables that “can distort the interpretation of multiple regression results” (Meyers et al., 2006, p. 180). Pearson correlations above the middle .7s between predictor variables reveal multicollinearity (Meyers et al., 2006). If multicollinearity is determined between two predictor variables, one of the variables should be removed (Muijs, 2011). Multicollinearity results are reported in Table 11.
Table 11

*Pearson Product-Moment Correlations between Dependent Variables (n = 54)*

<table>
<thead>
<tr>
<th></th>
<th>SPR</th>
<th>Gender</th>
<th>SAR</th>
<th>GPA</th>
<th>Course Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPR</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>.057</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>.554***</td>
<td>.188</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPA</td>
<td>.597***</td>
<td>.200</td>
<td>.604***</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Course Level</td>
<td>.600***</td>
<td>.153</td>
<td>.548***</td>
<td>.759***</td>
<td>-</td>
</tr>
<tr>
<td>SQ #2</td>
<td>-.142</td>
<td>.117</td>
<td>.116</td>
<td>-.012</td>
<td>.055</td>
</tr>
</tbody>
</table>

*Note.* SPR = average score on single-point rubric; SAR = score on scaled analytic rubric; GPA = overall grade point average; SQ #2 = survey question number two: To what degree did you find the rubric helpful?; *** *p* < 0.001, one-tailed.

**Remediation of violation of multicollinearity.** Remediation of the violation of multicollinearity for course level and GPA (.759, *p* < 0.001) needed to occur before data analysis could continue. Course level (general education and advanced placement) is inherently connected in the setting due to aforementioned course placement policies based on academic achievement, specifically student grades. Because frequencies revealed unbalanced course level data (Pallant, 2016), and the power of GPA as a scaled continuous variable outweighs the dichotomous nature of course level data (Altman & Royston, 2006), course level was removed as a predictor variable from the stepwise multilinear regression model.

**Suppressor variables.** A suppressor variable correlates with another predictor variable, increasing that variable’s predictive power and thus negating its own beta weight (Meyers, 2006). An analysis of both the Pearson correlations with the criterion variable of the average of
the two raters’ essay scores using a single-point writing rubric as indicated in Table 11 and the beta weights of the regression coefficients as indicated in Table 12 was conducted to determine the existence of differing signs—positive and negative—indicating a suppressor variable.

Table 12

*Stepwise Multilinear Regression Coefficients*

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>-3.617</td>
</tr>
<tr>
<td></td>
<td>GPA</td>
<td>1.012</td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td>8.379</td>
</tr>
<tr>
<td></td>
<td>GPA</td>
<td>.701</td>
</tr>
<tr>
<td></td>
<td>SAR</td>
<td>.203</td>
</tr>
</tbody>
</table>

Analysis of the data revealed no suppressor variables. There was no differentiation of positive and negative signs between any of the predictor variables’ Pearson correlations with the criterion variable and their beta weights as reported in Tables 11 and 12.

**Research Question One Data Analyses**

Data analysis to establish construct validity through the predictive power of multilinear regression followed the removal of participants with missing data, the removal of participants’ whose essay scores from rater one and rater two using the single-point writing rubric differed more than 2 standard deviations from the mean of differences, the elimination of survey question one as a predictor variable due to highly unbalanced frequencies, the transformation of survey...
question two Likert scale responses to dichotomous data, and the elimination of course level as a predictor variable because of its multicollinearity with GPA as a predictor variable.

**Frequencies.** Frequencies are the most effective analysis for categorical variables as means and standard deviations are relatively meaningless for variables such as gender and course level (Pallant, 2016). Frequencies for the remaining categorical predictor variables of gender and responses to survey question two are recorded in Table 8.

**Means, Standard Deviations, and Intercorrelations.** Descriptive statistics of continuous data were calculated for the criterion variable of average score (based on two independent raters) using the single-point writing rubric, as well as for predictor variables of scaled analytic rubric writing score and GPA. Descriptive statistics are recorded in Table 9.

Intercorrelations were reported for each of the variables using a *Pearson Product Moment Correlations* test as indicated in Table 11. Two predictor variables—scaled analytic rubric writing scores and GPA were significantly correlated ($p < .001$) with single-point writing rubric scores. No significant correlation was revealed in the Pearson Product-Moment Correlation between Dependent Variables for the predictor variables of gender and survey question number two (to what degree did you find the rubric helpful?) and the criterion variable of single-point writing rubric scores.

Two of the four predictor variables, student essay scores using a scaled analytic rubric and GPA, positively correlated with the criterion variable of single-point writing rubric scores. As GPAs and student scores using the scaled analytic rubric increased, their scores using the single-point writing rubric increased. No significant correlation was revealed in the Pearson Product-Moment Correlation between Dependent Variables reported in Table 11 for the
predictor variables of gender or survey questions number two (to what degree did you find the rubric helpful?) and the criterion variable of single-point writing rubric scores.

SPSS 25 was used to conduct multicollinearity assessments for intercorrelations of predictor variables. Tolerance for each independent variable is determined by running separate regression analyses for with each predictor variable serving as a criterion variable being predicted by the remaining independent variables in the analysis, with multicollinearity present if tolerance values are .01 or less (Meyers et al., 2016). Tolerance levels for the predictor variables GPA and scaled analytic rubric scores exceeded the .01 threshold.

Variance Inflation Factor (VIF) is the “reciprocal of the tolerance and measures the degree of linear association between a particular independent variable and the remaining independent variables” (Meyers et al., 2006, p. 212). VIF levels for the predictor variables GPA and scaled analytic rubric scores were well within normal bounds, less than 10, indicating the multicollinearity is not present among these independent variables.

SPSS 25 excluded the predictor variables of gender and survey question number two (to what degree did you find the rubric helpful?) from the Coefficients for the Stepwise Multilinear Regression table because gender ($p = .566$) and survey question number two ($p = .219$) lacked statistical significance. Collinearity statistics are reported in Table 13.
### Coefficients for the Stepwise Multilinear Regression

<table>
<thead>
<tr>
<th></th>
<th>Correlations</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T$</td>
<td>Sig.</td>
</tr>
<tr>
<td>(Constant)</td>
<td>.478</td>
<td>.634</td>
</tr>
<tr>
<td>GPA</td>
<td>3.142</td>
<td>.003</td>
</tr>
<tr>
<td>SAR</td>
<td>2.307</td>
<td>.025</td>
</tr>
</tbody>
</table>

*Note.* GPA = overall grade point average; SAR = score on scaled analytic rubric.

The Condition Index (Table 14) was examined to determine dependent relationships between the predictor variables (Meyers et al., 2006) to further rule out multicollinearity of independent variables. No combination of Condition Index numbers above 30 and two variance proportions above 50 (Meyers et al., 2006) for any of the predictor variables exists, further dismissing multicollinearity as a problem in this stepwise multilinear regression analysis.
Table 14

Collinearity Diagnostics for Research Question One

<table>
<thead>
<tr>
<th>Model</th>
<th>Dimension</th>
<th>Eigenvalue</th>
<th>Condition Index</th>
<th>Variance Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Constant)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.998</td>
<td>1.000</td>
<td>.00</td>
</tr>
<tr>
<td>2</td>
<td>.002</td>
<td>31.699</td>
<td>1.000</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2.981</td>
<td>1.000</td>
<td>.00</td>
</tr>
<tr>
<td>2</td>
<td>.017</td>
<td>13.159</td>
<td></td>
<td>.06</td>
</tr>
<tr>
<td>3</td>
<td>.002</td>
<td>44.535</td>
<td></td>
<td>.94</td>
</tr>
</tbody>
</table>

Note. GPA = overall grade point average; SAR = score on scaled analytic rubric.

**Stepwise multilinear regression analysis.** Stepwise multilinear regression was used to assess the ability of the remaining two predictor variables (GPA and scaled analytic rubric scores) to predict student writing scores using a single-point writing rubric. Preliminary analyses were conducted to ensure no violation of assumptions of normality, linearity, multicollinearity, and homoscedasticity.

An analysis of model one revealed that students’ overall grade point average was a significant predictor, $F(1, 54) = 29.971, p < .001, R^2 = .357$, of student scores using the single-point writing rubric, accounting for 35.7% of the variation in single-point writing rubric scores. An analysis of model two revealed that when the predictor variable of students’ scaled analytic rubric writing scores, $F(2, 53) = 18.847, p < .001, R^2 = .394$, was added to the model, an
additional 3.7% of the variance of single-point writing rubric scores was explained as seen in the $R^2$ values in Table 15.

Table 15

*Model Summary<sup>c</sup> for Research Question One*

<table>
<thead>
<tr>
<th>Model</th>
<th>$R$</th>
<th>$R$ Square</th>
<th>Adjusted $R$ Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.597&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.357</td>
<td>.345</td>
<td>8.201</td>
</tr>
<tr>
<td>2</td>
<td>.645&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.394</td>
<td>.394</td>
<td>7.891</td>
</tr>
</tbody>
</table>

<sup>a</sup> Predictors: (Constant), Overall weighted GPA.
<sup>b</sup> Predictors: (Constant), Overall weighted GPA, Scaled analytic rubric score (SAR).
<sup>c</sup> Dependent Variable: Single-point writing rubric score.

The model summary of the analysis of variance for this stepwise multilinear regression is reported in Table 16.

Table 16

*ANOVA<sup>a</sup> for Research Question One*

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>$df$</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>2015.592</td>
<td>1</td>
<td>2015.592</td>
<td>29.971</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>3631.533</td>
<td>54</td>
<td>67.251</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5647.125</td>
<td>55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>2347.070</td>
<td>2</td>
<td>1173.535</td>
<td>18.847</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>3300.055</td>
<td>53</td>
<td>62.265</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5647.125</td>
<td>55</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Dependent Variable: Single-point writing rubric (SPR) score
<sup>b</sup> Predictors: (Constant), Overall weighted GPA
<sup>c</sup> Predictors: (Constant), Overall weighted GPA, Scaled analytic rubric score (SAR)
Data Preparation for Research Question Two

In research question two, the researcher sought to determine the inter-rater reliability of a single-point writing rubric. Because it is “the most popular statistic” (Stemler, 2004, p. 4) for the analysis of inter-rater reliability, Pearson’s correlation coefficient (Gall et al., 2003) was conducted to establish inter-rater reliability of overall writing scores determined by raters using a single-point writing rubric. Additionally, item analysis of each item on the single-point writing rubric (19 in total) was measured to determine inter-rater scoring agreement using traditional inter-rater agreement percentages (Fraenkel & Wallen, 2003) as well as the non-parametric Cohen’s kappa (Stemler, 2004) to account for the small sample size (Pallant, 2016).

Data Screening Process

Following procedures described by Pallant (2016), rater one and rater two scores using the single-point rubric for each student’s essay, the average of the raters’ scores for each student’s essay, and the difference between rater one and rater two scores for each essay were recorded in a spreadsheet and then imported into SPSS 25 for coding and analysis. The codebook is reported in Table 6.

Missing data. Data screening was initially conducted to cleanse the data for accurate data analysis. Of the 66 participants, 64 completed both essays. Two students, code numbers 19 and 40, did not complete the second essay using the single-point writing rubric as a guide due to absences from class. Since the scores of this essay contributed to the criterion variable for interrater reliability analysis using Pearson’s $r$, as well as the basis of item analysis for inter-rater agreement using agreement percentages and Cohen’s $k$, all data associated with these two students were omitted from analysis.
Inter-rater Reliability

The development of a new instrument, the single-point writing rubric, requires procedures “to assess the reliability and validity of the scores for the population under study” (Creswell & Plano Clark, 2011, p. 237). The validity of the single-point writing rubric examined in this study was determined using Pearson’s correlation coefficient (Gall et al., 2003). The significant correlation of essay scores using a single-point rubric and those with a scaled analytic rubric with pre-established validity and reliability suggests validity of the single-point rubric. Inter-rater reliability applies when there is more than one judge, or in the case of a classroom setting, teacher, who is assessing student performance using a new instrument. For perfect inter-rater reliability, all raters would assign the same scores to the same performances using the same instrument (Muijs, 2011). Pearson’s correlation coefficient (Gall et al., 2003) was conducted to establish inter-rater reliability of overall writing scores using a single-point writing rubric.

Pearson’s correlation coefficient. The literature (Fraenkel & Wallen, 2003; Jonsson & Svingby, 2007; Meyers et al., 2006; Van Helvoort et al., 2016) supports the selection of Pearson’s correlation coefficient as “the most popular statistic” (Stemler, 2004, p. 4) for the analysis of inter-rater reliability for overall scores for research question two. SPSS 25 was used to perform a correlation of two variables, both continuous, with paired data sets (Meyers et al., 2006)—in this case rater one and rater two scores using a single-point writing rubric.

Pearson’s correlation coefficient assumptions. Data analysis to identify violations of assumptions of Pearson’s correlation coefficient was performed. Pearson’s $r$ operates under the assumption that data being correlated includes two continuous variables, or one continuous and one dichotomous categorical, and that those two variables contain pairs of values for each

**Descriptive statistics.** Descriptive statistics were utilized to analyze continuous variables to provide “basic summary statistics such as mean, median, and standard deviation” (Pallant, 2016, p. 56). Descriptive statistics for rater one and rater two scores using the single-point writing rubric are recorded in Table 17.

Table 17

*Descriptive Statistics for Research Question Two Continuous Variables (n = 64)*

<table>
<thead>
<tr>
<th>Rater</th>
<th>Min.</th>
<th>Max.</th>
<th>Mean</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPR Rater One</td>
<td>60</td>
<td>100</td>
<td>90.16</td>
<td>-1.183</td>
<td>.702</td>
</tr>
<tr>
<td>SPR Rater Two</td>
<td>68</td>
<td>100</td>
<td>91.61</td>
<td>-.940</td>
<td>-.542</td>
</tr>
</tbody>
</table>

**Normality.** The tests for normality required skewness and kurtosis to be determined for each set of rater scores using the single-point writing rubric. While skewness for rater one scores using the single-point writing rubric fell just outside the conventionally acceptable range of -1 to +1 (Meyers et al., 2006), the results fall well within more liberal interpretations of acceptable skewness of -2 to +2 (George & Mallery, 2010), with the acknowledgement that “many scales and measures in the social sciences have scores that are skewed […] revealing] the underlying nature of the construct being measured” (Pallant, 2016, p. 64). Kurtosis values for rater one and rater two single-point rubrics scores all fall within the conventionally acceptable level of -1 to +1. The skewness and kurtosis values for rater one and rater two scores were recorded in Table 17.
Additionally, no major deviations from normality were observed in the Normal Quartile (Q-Q) Plot of the Regression Standardized Residual in Figures 5 and 6.

Figure 5. Normal Q-Q Plot of Single-Point Rubric Scores (1st reader)
Figure 6. Normal Q-Q Plot of Single-Point Rubric Scores (2nd reader)

**Linearity and outliers.** A scatterplot of standardized residuals (Figure 7) had a “roughly rectangular distribution” (Pallant, 2016, p. 160), revealing no violations to linearity between rater one and rater two scores or outliers.
Figure 7. Scatterplot of Rater One and Rater Two SPR scores

**Homoscedasticity.** A matrix of bivariate scatterplots was run to determine the presence of homoscedasticity among rater one and rater two essay scores using the single-point writing rubric. The absence of a curvilinear relationship in the scatterplots in Figure 8 revealed no homoscedasticity was present.
Research Question Two Data Analyses of Inter-Rater Reliability

Reliability is “the consistency of scores obtained—how consistent they are for each individual from one administration to another and from one set of items to another (Fraenkel & Wallen, 2003, p. 165). A Pearson correlation coefficient (Meyers et al., 2006) was used to analyze inter-rater reliability of essay scores from rater one and rater two using the single-point writing rubric. Preliminary analyses were performed to ensure no violation of the assumptions of normality, linearity, and homoscedasticity. There was a strong, positive correlation between
rater one and rater two scores, \( r = .716, n = 64, p = 0.01 \), meaning that scores from rater one were positively associated with scores from rater two. The Pearson’s correlation coefficient for overall scores between rater one and rater two using the single-point rubric is reported in Table 18.

Table 18

*Pearson’s Correlation Coefficient for Single-point Rubric Scores (N = 64)*

<table>
<thead>
<tr>
<th>SPR Rater One</th>
<th>SPR Rater Two</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPR Rater One</td>
<td>-</td>
</tr>
<tr>
<td>SPR Rater Two</td>
<td>.716**</td>
</tr>
</tbody>
</table>

*Note.* SPR = Single-point rubric.

**p < 0.01, two-tailed.

**Inter-Rater Agreement**

Item analysis for each rater’s scores for each item on the single-point writing rubric (19 in total) for each student essay was conducted to determine inter-rater scoring agreement using traditional inter-rater agreement percentages (Fraenkel & Wallen, 2003), as well as the non-parametric Cohen’s kappa (Stemler, 2004) to account for the small sample size (Pallant, 2016).

**Simple percentage agreement.** Given some vast discrepancies observed when rater one and rater two scores were visually analyzed and standard deviations from the mean of those differences were calculated during data analysis for research question one in Table 7, item analysis was performed on the 19 dichotomous items of the single-point writing rubric. Rater one and rater two agreement for each item on each essay was initially measured using simple percentage-agreement figures (McHugh, 2012). The sum of the number of times rater one and rater two assigned identical scores for each of the 19 variables on the rubric was divided by the
number of potential agreements (i.e., total number of participants) and multiplied by 100. For reliability, scoring agreement of at least 80% is “desired” on instruments and items that are “susceptible to differences in administration, scoring, or both, such as essay evaluations (Fraenkel & Wallen, 2003, p. 169). A benefit of the simple percentage analysis is that it identifies variables that may be problematic through the strong intuitive appeal” of percentages (Stemler, 2004, p. 2).

**Assumptions of simple percentage agreement.** Simple percentage agreement assumes there are pairs of scores on the same variables from two different raters (Fraenkel & Wallen, 2003). Those pairs of scores are reported in Table 19.
Table 19

Simple Percentage Agreements of Single-point Rubric Items on Student Essays

<table>
<thead>
<tr>
<th>Domain</th>
<th>Criterion</th>
<th>N Disagreement</th>
<th>% Agreement</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction</strong></td>
<td>Creative Opening</td>
<td>3</td>
<td>95</td>
<td>.377**</td>
</tr>
<tr>
<td></td>
<td>Identifies Speaker</td>
<td>5</td>
<td>92</td>
<td>.245*</td>
</tr>
<tr>
<td></td>
<td>Identifies Subject</td>
<td>1</td>
<td>98</td>
<td>.000a</td>
</tr>
<tr>
<td></td>
<td>Identifies Audience</td>
<td>2</td>
<td>97</td>
<td>.000a</td>
</tr>
<tr>
<td></td>
<td>Identifies Purpose</td>
<td>17</td>
<td>72</td>
<td>.193</td>
</tr>
<tr>
<td></td>
<td>Identifies Techniques</td>
<td>10</td>
<td>83</td>
<td>.245**</td>
</tr>
<tr>
<td><strong>Body</strong></td>
<td>Transition Expression</td>
<td>2</td>
<td>97</td>
<td>.000a</td>
</tr>
<tr>
<td></td>
<td>Establishes Analytical Point</td>
<td>3</td>
<td>95</td>
<td>.642***</td>
</tr>
<tr>
<td></td>
<td>Relevant Evidence</td>
<td>13</td>
<td>80</td>
<td>.390**</td>
</tr>
<tr>
<td></td>
<td>Synthesis/Conclusion</td>
<td>5</td>
<td>92</td>
<td>.245*</td>
</tr>
<tr>
<td><strong>Conclusion</strong></td>
<td>Transition Expression</td>
<td>3</td>
<td>95</td>
<td>.000a</td>
</tr>
<tr>
<td></td>
<td>Summary Statement</td>
<td>6</td>
<td>91</td>
<td>-.049</td>
</tr>
<tr>
<td></td>
<td>Implications</td>
<td>23</td>
<td>64</td>
<td>.169</td>
</tr>
<tr>
<td></td>
<td>Dramatic Flourish</td>
<td>9</td>
<td>86</td>
<td>.553***</td>
</tr>
<tr>
<td><strong>Conventions</strong></td>
<td>Organization</td>
<td>11</td>
<td>83</td>
<td>.074</td>
</tr>
<tr>
<td></td>
<td>Syntax</td>
<td>11</td>
<td>83</td>
<td>.074</td>
</tr>
</tbody>
</table>

*Note.  N = 64.*

a = no statistics because rater one and/or rater two is a constant.

*p < 0.05, two-tailed.  **p < 0.01, two-tailed.  ***p < 0.001, two-tailed.*

(continued)
Table 19

*Simple Percentage Agreements of Single-point Rubric Items on Student Essays*

<table>
<thead>
<tr>
<th>Domain</th>
<th>N Disagreement</th>
<th>% Agreement</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quotations</td>
<td>0</td>
<td>100</td>
<td>.000a</td>
</tr>
<tr>
<td>Tone &amp; Style</td>
<td>6</td>
<td>91</td>
<td>.200</td>
</tr>
<tr>
<td>Diction</td>
<td>14</td>
<td>78</td>
<td>.176</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>19</strong></td>
<td><strong>144</strong></td>
<td><strong>88</strong></td>
</tr>
</tbody>
</table>

*Note. N = 64.

a = no statistics because rater one and/or rater two is a constant.

*p < 0.05, two-tailed. **p < 0.01, two-tailed. ***p < 0.001, two-tailed.*

**Cohen’s Kappa.** Stemler (2004) cautions that percentage “consensus estimates” can be inflated with dichotomous data despite its “strong intuitive appeal” (p. 2). Cohen’s kappa statistic, “an estimate of the degree of consensus between two judges after correcting the percent-agreement figure for the amount of agreement that could be expected by chance alone,” was analyzed to abet item-analysis reliability as it is “a highly useful statistic when one is concerned that the percent-agreement statistic may be artificially inflated due to the fact that most observations fall into a single category” (Stemler, 2004, p. 2). McHugh (2012) argues that Cohen’s suggested interpretation of kappa values, with .41 and above considered moderate agreement is too low, as such results imply the acceptability of raters disagreeing 60% of the time. McHugh’s “more logical interpretation” (2012, para. 13) raises the level of moderate agreement to .60. Table 20 displays these competing interpretations of Cohen’s kappa values.
Table 20

*Cohen’s and McHugh’s Interpretations of Cohen’s Kappa*

<table>
<thead>
<tr>
<th>Cohen’s Level of Agreement</th>
<th>Value of Kappa</th>
<th>McHugh’s Level of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>None to Slight</td>
<td>0-.20</td>
<td>None</td>
</tr>
<tr>
<td>Fair</td>
<td>.21-.39</td>
<td>Minimal</td>
</tr>
<tr>
<td>Moderate</td>
<td>.40-.59</td>
<td>Weak</td>
</tr>
<tr>
<td>Substantial</td>
<td>.60-.79</td>
<td>Moderate</td>
</tr>
<tr>
<td>Almost Perfect</td>
<td>.80-.90</td>
<td>Strong</td>
</tr>
<tr>
<td>Above .90</td>
<td></td>
<td>Almost Perfect</td>
</tr>
</tbody>
</table>


**Assumptions of Cohen’s Kappa.** Cohen’s kappa assumes that data are nominal, paired, and that those pairs scores are exclusive of one another, that cross tabulations are symmetrical, and that all scores are conducted by the same two raters independent of one another (Laerd Statistics, 2018). Those pairs of scores are reported in Table 7.

**Research Question Two Data Analyses of Inter-Rater Agreement**

Data analysis followed the removal of participants with missing data. The balance of the participants ($N = 64$) met the assumptions of single percentage agreement and Cohen’s kappa with pairs of scores that are dichotomous, mutually exclusive, and were assigned by the same two raters independent of one another. Rater one and rater two simple percentage agreements and Cohen’s kappa statistics on individual items for the two raters using the single-point rubric are reported in Table 19.
Item score agreement between rater one and rater two totaled 88%, with 17 of 19 items meeting or surpassing the minimum threshold of 80% agreement (Fraenkel & Wallen, 2003), reinforcing instrument reliability conclusions drawn from the Pearson’s correlation coefficient determined for rater one and rater two essay scores. Kappa values for each item, which remove inflated percentages of agreement in order to provide a valid measure of the level of agreement beyond chance, ranged from none-to-slight to substantial, with most variables displaying little agreement beyond chance. However, a limitation of Cohen’s kappa is that “when both raters report a very high prevalence of the condition of interest” (as in 17 of the 19 items analyzed), the overlap of scores “should be considered ‘true’ agreement, but it is attributed to chance agreement (i.e., \( K = 0 \)), which results in lower kappa values (Tang, Hu, Zhang, Wu, & He, 2015, p. 67).

**Chapter Summary**

This quantitative, correlational research study utilized statistical analysis to establish validity and inter-rater reliability of a teacher-generated single-point writing rubric for instruction and assessment. For research question one, the non-directional hypothesis indicating analytic rubric scores, overall GPA, and self-reported student perceptions of single-point rubric efficacy would significantly predict writing scores for students who were provided a single-point writing rubric during the writing process was supported by the data, establishing construct validity of the single-point rubric as a scoring instrument. Course enrollment, gender and self-reported single-point rubric use were not supported by the data as predictor variables.

For research question two, the non-directional hypothesis indicating there will be a significant degree of inter-rater reliability of a single-point rubric was supported by data analysis conducted using Pearson’s correlation coefficient. Cumulative item analysis using simple
percentage agreement further supported inter-rater reliability but exposed troublesome items on the single-point rubric and the overall statistical frailty of dichotomous scoring.
CHAPTER FIVE: SUMMARY AND CONCLUSIONS

This chapter begins with a summary of the study and results for each research question. This is followed by a discussion of the data findings, supported by the theory and research studies found in the review of literature. These findings shape the implications for use of single-point rubrics as pedagogical tools in the writing classroom. The study’s results are then substantiated by an assessment of the threats to internal and external validity. The conclusion of the chapter contains suggestions for future research on the use of rubrics as instructive tools for the teaching of writing.

Synopsis of the Study

The purpose of this study was to establish validity and inter-rater reliability of the teacher-generated single-point writing rubric. Student-performance using the single-point writing rubric during the writing process was correlated with student-performance using a scaled analytic rubric that has pre-established validity and reliability. Course level (general education or advanced placement), gender, overall grade point average (GPA), self-reported single-point rubric use (Survey Question #1), and self-reported student perceptions of single-point rubric efficacy (Survey Question #2) were also initially correlated with student performance using a single-point writing rubric. Rater one and rater two essay scores from the single-point rubric and item analysis of each rubric component the two raters scored were also correlated to determine inter-rater reliability and inter-rater agreement of the single-point writing rubric. Two research questions were developed to address the purposes of this study.

1. To what degree and in what manner do scaled analytic rubric scores, course level, gender, GPA, self-reported single-point rubric use, and self-reported student
perceptions of single-point rubric efficacy predict writing scores for students who were provided a single-point writing rubric during the writing process?

2. What is the inter-rater reliability of a single-point writing rubric?

A sample of accessible high school seniors (n = 86) was derived from a small, suburban, relatively affluent high school where graduation rates at the setting hover near 100% annually and 90% of students attend four-year colleges immediately after high school. A convenience sampling method was utilized for several established reasons in educational research (Gall et al., 2003), including researcher access, familiarity, and administrative approval. The setting chosen was both purposive and convenient—the geographic location, socioeconomic makeup of the school, the characteristics of student groupings, and the students’ educational experiences with rubrics were reflective of the study’s design, and the researcher had potential access to the participants in this school. Of the 86 students in the accessible population, 66 students opted to participate in the study, and complete data were collected from 64.

A stepwise multilinear regression (Meyers, Gamst, & Guarino, 2006) was utilized to determine if the non-directional hypothesis for research question one was supported by the data—that is, if scaled analytic writing rubric scores, course level, gender, GPA, self-reported single-point writing rubric use, and self-reported student perceptions of single-point writing rubric efficacy predicted writing scores for students who were provided a single-point writing rubric during the writing process.

To determine if the non-directional hypothesis for research question two was supported by the data, a Pearson’s correlation coefficient (Gall et al., 2003) was calculated to establish inter-rater reliability of overall writing scores using a single-point writing rubric. Additionally, item analysis of each item on the single-point writing rubric (19 in total) was measured to
determine inter-rater scoring agreement using traditional inter-rater agreement percentages (Fraenkel & Wallen, 2003) as well as the non-parametric Cohen’s kappa (Stemler, 2004) to account for the small sample size (Pallant, 2016).

Results

The findings for this study were a consequence of thorough data analysis that identified and attempted to resolve all data violations. The findings, implications, and recommendations for future research are supported by the data and a thorough review of the literature.

Research Question One. Data screening and data analysis resulted in the removal of participants with missing data, the removal of participants’ whose essay scores from rater one and rater two using the single-point writing rubric differed more than 2 standard deviations from the mean of differences, the elimination of survey question one as a predictor variable due to highly unbalanced frequencies, the transformation of survey question two Likert scale responses to dichotomous data, and the elimination of course level as a predictor variable because of its multicollinearity with GPA as a predictor variable. Stepwise multiple regression was used to assess the ability of the remaining four remaining predictor variables (gender, survey question two, GPA and scaled analytic rubric scores) to predict student writing scores using a single-point writing rubric for the remaining participants (n = 54).

A stepwise multiple regression analysis revealed that students’ overall grade point average was a significant predictor, $F(1, 54) = 29.971, p < .001, R^2 = .357$, of student scores using the single-point writing rubric, accounting for 35.7% of the variation in single-point writing rubric scores. When the predictor variable of students’ scaled analytic rubric writing scores, $F(2, 53) = 18.847, p < .001, R^2 = .394$, was added to the model, an additional 3.7% of the variance of single-point writing rubric scores was revealed.
The non-directional hypothesis selected to answer research question one was partially supported by the data, which revealed that the independent variables of grade point average (GPA) and the participants’ scaled analytic rubric scores were significant predictors of participants’ single-point rubric scores. Gender and student responses to survey question number two (self-reported student perceptions of single-point rubric efficacy) did not reveal statistically significant correlations. Course level and student responses to survey question number one (self-reported single point rubric use) were omitted due to violations of assumptions of multilinear regression.

**Research Question Two.** Data screening was initially conducted to cleanse the data for accurate data analysis. Of the 66 participants, 64 completed both essays. The scores of this essay contributed to the criterion variable for inter-rater reliability analysis using Pearson’s $r$, as well as the basis of item analysis for inter-rater agreement using agreement percentages and Cohen’s $k$.

Pearson’s correlation coefficient (Gall et al., 2003) was conducted to establish inter-rater reliability of overall writing scores using a single-point writing rubric. The non-directional hypothesis selected to answer research question two was supported by the data, which revealed a strong, positive significant correlation between rater one and rater two scores, $r = .716, n = 64, p = 0.01$, meaning that scores from rater one were positively associated with scores from rater two. Preliminary analyses were performed to ensure no violation of the assumptions of normality, linearity, and homoscedasticity.

Rater one and rater two agreement for each item on each essay was initially measured using simple percentage-agreement values (McHugh, 2012). For reliability, scoring agreement of at least 80% is “desired” on instruments and items that are “susceptible to differences in
administration, scoring, or both, such as essay evaluations (Fraenkel & Wallen, 2003, p. 169). Item score agreement between rater one and rater two totaled 88%, with 17 of 19 items meeting or surpassing the minimum threshold of 80% agreement (Fraenkel & Wallen, 2003), reinforcing instrument reliability conclusions drawn from the Pearson’s correlation coefficient determined for rater one and rater two overall essay scores.

Kappa values for each item, which remove inflated percentages of agreement in order to provide a valid measure of the level of agreement beyond chance, ranged from none-to-slight to substantial, with most variables displaying little agreement beyond chance. However, a limitation of Cohen’s kappa is that “when both raters report a very high prevalence of the condition of interest” (as in 17 of the 19 items analyzed), the overlap of scores “should be considered ‘true’ agreement, but it is attributed to chance agreement (i.e., \( K = 0 \))” (Tang, Hu, Zhang, Wu, & He, 2015, p. 67). Thus, low kappa values were likely due to the high frequency of identical scores and should be contextualized in analysis.

**Discussion**

The theoretical foundation provided within the review of literature in chapter two of this research study addressed Vygotsky’s (1978) theory of the zone of proximal development as operationalized through the pedagogical construct of formative assessments, specifically using rubrics. The zone of proximal development was Vygotsky’s seminal work that distinguished between a student’s actual developmental level with that of his or her proximal developmental level after teacher assistance. A pedagogy utilizing rubrics as formative assessments—that is “supporting learning through the instructional scaffolding, including feedback, and the active involvement of students in the assessment/learning process” (Heritage, 2010, p. 15)—is reflective of the basic tenets of the zone of proximal development. Ash and Levitt (2003)
conducted two case studies of the zone of proximal development as the “core feature” (p. 23) of formative assessment processes. They concluded that transformative learning for both teachers and students was the outcome of a pedagogy rooted in formative assessments. Teachers and students engaged in an “ever-changing zone of mutual understanding” (p. 41) revealing the expanding upper limits of the zone of proximal development. Potentially germane to this study, given the high frequency of students who self-reported that they found the single-point rubric helpful (see Table 8), was Ash and Levitt’s (2003) observation that the teachers’ “ever-increasing fine-tuning of scaffolding tool use”—that is, their revision of standard rubric use—within the context of formative assessments “led the learner to move towards more independent understanding” (p. 42).

The review of literature was provided to inform the rationale and findings of this research study. The conclusions drawn for the following research questions are grounded in the theory and studies analyzed in chapter two of this study.

**Research Question One.** Research question one was used to investigate the degree to which student-performance using the single-point writing rubric during the writing process was correlated with student-performance using a scaled analytic rubric that had pre-established validity and reliability. Course level (general education or advanced placement), gender, overall grade point average (GPA), self-reported single-point rubric use (Survey Question #1), and self-reported student perceptions of single-point rubric efficacy (Survey Question #2) were also initially correlated with student performance using a single-point writing rubric. The purpose of correlational analysis was to establish validity for the single-point rubric.

Citing the theoretical constructs of Vygotsky’s (1978) zone of proximal development and self-regulated learning (SRL), Fluckiger (2010) concluded that single-point rubrics are an
effective tool for qualitative feedback in a formative setting where students utilize the rubric for self-assessment throughout the writing process. Fluckiger (2010) found that the quality of work by students using a single-point rubric was equal to or better than work completed using a scaled rubric, and that participants in the study perceived that they learned more using a single-point rubric and had a better understanding of the traits of good writing. These results support the findings of this study, which reported a strong correlation among student writing scores based on the scaled analytic writing rubric and the single-point writing rubric.

Van Helvoort, Brand-Gruwel, Huysmans, and Sjoer (2016) utilized Pearson correlations to assess validity of an instructor-designed information literacy rubric. High correlations between overall scores with an existing assessment rubric supported validity conclusions. However, Rezaei and Lorvorn (2010) concluded that raters are ultimately influenced by their intuitive impressions of student writing, even when provided with a rubric, diminishing its validity as an assessment instrument. Germane to this study is Rezaei and Lorvorn’s (2010) belief that a limiting factor was the broadness of the rubric used in their study. They recommended further research utilizing a rubric with “more detailed descriptors” (p. 28), like the single-point rubric used in this study. These results and conclusions support the findings of this study, which, like Van Helvoort, Brand-Gruwel, Huysmans, and Sjoer (2016), utilized a rubric with pre-established validity to establish validity for the new single-point rubric through Pearson’s correlation coefficient. Further, the single-point rubric relies on specific and nuanced descriptors (see Appendix B) to guide and measure student performance, as Rezaei and Lorvorn (2010) recommended.
Research Question Two. Research question two investigated the inter-rater reliability, the consistency of scores among multiple raters, of a single-point writing rubric for the purpose of establishing the reliability of the instrument for the evaluation of student writing.

Jonsson and Svingby (2007) conducted a comprehensive article review of seventy-five papers that presented empirical evidence of rubric reliability to determine if rubrics “enhance the reliability of scoring” (p. 132). The researchers assert that while classroom assessments do not require the same high degree of reliability, some degree of scoring consistency is necessary. Inter-rater reliability was examined in some of the studies using Pearson’s correlation coefficients, which were generally moderate to strong, with most above the .70 threshold (Stemler, 2004). These results support the findings of this study, which used Pearson’s correlation coefficient to analyze inter-rater reliability of overall essay scores, finding a strong, positive correlation for rater one and rater two scores using the single-point writing rubric \((r = .716)\).

Additionally, more than half the studies examined by Jonsson and Svingby (2007) relied upon consensus agreement—the percentage of agreed upon scores between raters to determine inter-rate reliability. Few of the studies achieved the .80 threshold (Fraenkel & Wallen, 2003) for exact agreement, only reaching good levels of consistency with adjacent agreement. These results partially support the findings of this study, which found item score agreement on components of the single-point writing rubric between rater one and rater two totaled 88%, with 17 of 19 items on the rubric revealing strong agreement percentages, albeit using a dichotomous scoring method.

Jonsson and Svingby (2007) noted that Cohen’s kappa, which considers how consensus agreements vary from chance agreements when using rubrics with few or dichotomous scoring
levels, was applied in some studies with most only reaching fair agreement levels above .20 and below .70. These results partially support the findings of this study, which found kappa values ranged from none-to-slight to substantial, with most variables displaying little agreement beyond chance. However, the high percentage of identical scores (simple agreement) on 17 of the 19 items lowered kappa values despite true agreement, thus limiting the implications of kappa analysis (Tang, Hu, Zhang, Wu, & He, 2015).

**Implications for Education**

The findings of this research study, that the single-point writing rubric is a valid and reliable evaluation instrument, suggest that teachers should consider using single-point writing rubrics as a complement to or in lieu of traditional scaled analytic writing rubrics to teach and assess writing. The results are consistent with Fluckiger’s (2010) findings that the quality of work by students using a single-point rubric was equal to or better than work completed using a scaled analytic rubric.

However, the context of any instrument’s use impacts its effectiveness. Rubrics have been found to be particularly effective when utilized as an instrument of both instruction and evaluation. A pedagogy that relies on formative assessments and teacher-student collaboration to help students achieve their proximal developmental levels was present and, the literature suggests, is necessary for improved performance by students (Andrade, 2005; Andrade et al., 2008; Ash & Levitt, 2003; Edens & Shields, 2015). In short, single-point writing rubrics need to be introduced at the beginning and incorporated throughout the writing process for students to reap their benefits as instructional tools.

Additionally, rubric design is paramount to its effectiveness as a teaching instrument, and its potential validity and reliability as an evaluative tool (Orsmond & Merry, 1996; Panadero &
Jonsson, 2013; Rezaei & Loro, 2010). At the risk of being too constraining and potentially invalid (Farenga, Ness, & Sawyer, 2015; Mabry, 1999), single-point writing rubric design, like the one in this study, need to compensate for the educational shortcomings of traditional scaled analytic rubrics. Clearly described, transcendent criteria that promote the tenets of good writing are the hallmarks of a well-designed single-point rubric (Fluckiger, 2010). And there is merit to the additional criteria that could be included in a single-point rubric. The rubric in this study, for example, utilized 19 criteria to instruct and assess student composition. Scaled rubrics traditionally rely no more than six domains. Concomitantly, nuanced conversations regarding the criteria of the rubric to promote understanding, as well as qualifiers to preclude formulaic and stilted writing-to-the-rubric must occur. Lest we forget, rubrics “are not a replacement for good instruction” (Andrade, 2005, p. 29).

**Limitations of the Study**

This quantitative study utilizing a convenience sample was developed to investigate the correlation between student performance using a teacher-generated single-point rubric and student performance using a scaled analytic rubric with pre-established validity and reliability (among other variables) to establish the validity of the single-point rubric. Student scores from a single-point rubric were also utilized to determine overall inter-rater reliability of single-point rubrics and inter-rater agreement of items assessed using single-point rubrics. The following threats to internal and external validity exist in a correlational design (Fraenkel & Wallen, 2003).

**Internal Threats.** Internal validity is “the extent to which the investigator can conclude that there is a cause and effect relationship among variables” (Creswell & Plano Clark, 2011, p. 211). Extraneous variables need to be controlled so that conclusions can be attributed to the relationship between the independent and dependent variables (Gall et al., 2003). With
Correlational studies, the primary threat to internal validity is that “other characteristics” than those demarked as variables “can explain relationships” (Fraenkel & Wallen, 2003, p. 348).

That the students in this study were consensual participants raises the potential for subject-attitude threats to internal validity (Fraenkel & Wallen, 2003). Students could have performed better (or worse) on one or both essays because they were conscious of the study rather than as a result of the instruments under scrutiny. The standardization of the tasks and their inherent relationship to the course curriculum helped mitigate this threat.

To avoid internal threats to validity that may arise from instrument decay and testing (Fraenkel & Wallen, 2003), participants completed similar but not identical New York State Common Core Regents writing tasks (Appendices F and G) utilizing a scaled analytic writing rubric (Appendix A) and a single-point writing rubric (Appendix B), respectively. The research design was used to deliberately space the relatively brief tasks two weeks apart, and the structure of the tasks themselves are inherent to the curriculum and instruction of the setting (i.e., senior English classes).

Threats of data collector characteristics and data collector bias were mitigated by having one teacher administer the task for consistency, and two other teachers scoring the papers blindly—unaware of participant names or characteristics (e.g., gender, course enrollment, GPA).

**External Threats.** Mortality emerged as an external threat, as limitations in sampling impact the generalizability of this study and pose the greatest threat to external validity (Gall et al., 2003). The homogenous sample, while convenient and purposive to the study and reflective of the location’s regional demographics, threatens population validity and, thus, limits generalizability. The setting is a small high school, with relatively little racial, ethnic, or socio-economic diversity. For increased generalizability, studies with larger, more diverse populations
are needed, and samples should be drawn from multiple age groups in multiple disciplines.

Sampling also limits the scope of statistical analysis, as low enrollment in general education senior English classes affected data analysis of the correlational study (Fraenkel & Wallen, 2003). The researcher avoided losing participants to mitigate this threat. Novelty and disruption effects were controlled for by the instrumentation: The tasks and rubrics are grade-level appropriate and familiar curricular staples to students.

**Suggestions for Future Research**

The findings from this research study serve as impetus for future research on the use of rubrics, particularly single-point rubrics, as instruments in a pedagogical approach that utilizes formative assessment strategies.

**Student Perceptions of Single-Point Writing Rubrics.** Statistical analysis can explain significance and predict outcomes, but it certainly cannot tell the whole story. Qualitative and mixed methods research designs could add depth and nuance to researchers’ understandings of how students use the various rubric structures, particularly single-point writing rubrics and the scaled analytic writing rubrics in contrast, offering valuable insights that could influence how teachers design and use rubrics to improve student outcomes.

**Teacher Attitudes Toward Single-Point Writing Rubrics.** Rezaei and Lorvorn’s (2010) findings concluded that raters are ultimately influenced by their intuitive impressions of student writing, even when using traditional scaled analytic rubrics. This impacts the validity and reliability of these rubrics as an evaluative tool. Additional research could determine if single-point rubric design impacts how teachers perceive and use rubrics as instructional and assessment instruments, potentially influencing writing instruction at the classroom level. For example, the amount of time needed to complete each type of rubric may impact teacher attitudes.
as well as influence the fatigue factor if one type of rubric takes more type to complete. This is an area of future research.

**Effectiveness of Single-Point Writing Rubrics in Diverse Populations of Learners.** Limitations of the convenience sample impact the generalizability of this study and pose the greatest threat to external validity (Gall et al., 2003). The setting is a small high school with relatively little racial, ethnic, or socio-economic diversity. For increased generalizability, studies with larger, more diverse populations are needed, and samples should be drawn from multiple age groups in multiple disciplines.

**Additional Validity and Reliability Studies of Single-Point Rubrics.** This research study was motivated in part by a dearth of research on the validity and reliability of single-point rubrics. Brookhart’s (2003) seminal argument for developing a measurement theory for classroom assessment purposes challenges the use of “borrowed theory” (p. 5) from large-scale assessments in classroom settings. Additional studies measuring the validity and reliability of single-point rubrics could foster greater acceptance as instructional and evaluative tools.

**Chapter Summary**

This chapter provided the research questions and findings of this study, abetted by references to the literature reviewed in chapter two. Analysis and implications for classroom practice followed. Limitations and subsequent remediation were addressed in support of the validity and reliability of the findings. Based on the literature and results of this study, suggestions for future research on the use of single-point rubrics as instructional and evaluative tools were provided.
References


College Board. SAT suite of assessments. Retrieved from https://collegereadiness.collegeboard.org/sat


doi: 10.1016/j.jclinepi.2010.12.001


Appendices

Appendix A-NYS Regents Scaled Analytic Rubric

<table>
<thead>
<tr>
<th>Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>Criteria</td>
<td>Content (90%)</td>
<td>Content (90%)</td>
<td>Content (90%)</td>
<td>Content (90%)</td>
<td>Content (90%)</td>
</tr>
<tr>
<td></td>
<td>Analysis (10%)</td>
<td>Analysis (10%)</td>
<td>Analysis (10%)</td>
<td>Analysis (10%)</td>
<td>Analysis (10%)</td>
</tr>
<tr>
<td></td>
<td>Use of Language (10%)</td>
<td>Use of Language (10%)</td>
<td>Use of Language (10%)</td>
<td>Use of Language (10%)</td>
<td>Use of Language (10%)</td>
</tr>
</tbody>
</table>

# Appendix B-Single-Point Rubric

<table>
<thead>
<tr>
<th>Area</th>
<th>Description</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction</strong> (SOAP +RT)</td>
<td>Opens creatively (e.g., interrupted quote, adjective clause, participial phrase, infinitive phrase, noun clause)</td>
<td>0 1</td>
</tr>
<tr>
<td>Speaker</td>
<td>Identifies speaker.</td>
<td>0 1</td>
</tr>
<tr>
<td>Occasion</td>
<td>Identifies subject matter.</td>
<td>0 1</td>
</tr>
<tr>
<td>Audience</td>
<td>Identifies primary intended audience.</td>
<td>0 1</td>
</tr>
<tr>
<td>Purpose</td>
<td>Precisely, specifically, and comprehensively identifies author’s intent that addresses prompt.</td>
<td>0 1</td>
</tr>
<tr>
<td>Uses expression “rhetorical techniques” or “literary techniques” or identifies specific technique(s) that will be discussed in essay</td>
<td>0 1</td>
<td></td>
</tr>
<tr>
<td><strong>Body</strong></td>
<td>Transition expression that reveals relationship between this paragraph and previous paragraph.</td>
<td>0 2</td>
</tr>
<tr>
<td>1 well developed point of analysis of technique(s)</td>
<td>0 3</td>
<td></td>
</tr>
<tr>
<td>Provides comprehensive evidence to support analytical point.</td>
<td>0 3</td>
<td></td>
</tr>
<tr>
<td>Concluding statement that reveals how this specific technique / part of passage contributes to author’s overall intent.</td>
<td>0 2</td>
<td></td>
</tr>
<tr>
<td><strong>Conclusion</strong></td>
<td>Implicit transition into conclusion (noun clause / rhetorical question / short sentence / demonstrative pronoun).</td>
<td>0 1</td>
</tr>
<tr>
<td>Summary statement of author’s main point.</td>
<td>0 1</td>
<td></td>
</tr>
<tr>
<td>Implications of author’s main point for audience (we / us / our)</td>
<td>0 1</td>
<td></td>
</tr>
<tr>
<td>Dramatic flourish (partial quote, short sentence).</td>
<td>0 1</td>
<td></td>
</tr>
</tbody>
</table>
### Appendix B-Single-Point Rubric (continued)

| Command of organization (logical structure of analysis and evidence) | 0 | 1 |
| Command of syntax (No run-ons or fragments, variety, punctuation) | 0 | 1 |
| Command of quotations (proper structure, variety, punctuation) | 0 | 1 |
| Command of tone and style (appropriate and consistent for task) | 0 | 1 |
| Command of diction (rhetorical jargon, good verbs, avoids language repetition) | 0 | 1 |

| Total Points | 25 possible points |
| Final Score | Total points x 4= |
Appendix C - Student’s Assent

Department of Education and Educational Psychology
Western Connecticut State University
181 White Street
Danbury, CT 06810
January 23, 2019

Dear Student,

I am in a doctoral program at Western Connecticut State University. I am doing an exciting research study about writing assessments. I would like you to be a part of my study. I will send a permission slip home with you. But first, I would like you to know more about my project.

The study is about the ways in which you use rubrics during the writing process. I will ask you to complete two writing tasks. Your teacher will incorporate these tasks into your curriculum so you will not be missing out on any other lessons. Other English teachers (not me or your teacher) will score your essays blindly to protect your anonymity. Your classroom teacher and I will be the only people who know your identity for your essays or the follow-up survey. Your guidance counselor will also provide me your high school GPA so I can see how it relates to your essay performance on the two tasks.

I will not use your name in the study; I will use numbers. The task will have nothing to do with report card grades and the ratings will not be reported to your parents or teachers. All of the
information will be kept private. You have the option of not participating in the study with no impact on your grade; your teacher will provide you with an alternative assignment. If you have any questions, please ask me.

If you would like to be in my study, please print and sign your name below:

___________________________________________________
Print student name

X ____________________________________________
Student signature

Sincerely,
David Popken

Sally Dobyns, PhD
Adjunct Professor, EdD in Instructional Leadership

popken001@connect.wcsu.edu
dobynss@wcsu.edu
Appendix D-Parent’s Consent

Department of Education and Educational Psychology
Western Connecticut State University
181 White Street
Danbury, CT 06810
January 23, 2019

Dear Parent:

I am currently enrolled in the doctoral program for Instructional Leadership at Western Connecticut State University. This program requires that I design and implement a dissertation research study. The purpose of this study, which will require approximately six class periods of instructional time with senior English students, is to determine the validity and reliability of teacher-generated single-point writing rubrics utilized by teachers in the North Salem English department. The study will also measure the predictive quality of several variables: student writing scores when using a scaled analytic rubric; gender; student GPA; student course enrollment (AP vs. regular English); students’ feedback on if they used each rubric structure and how much they found each helpful during the writing process.

Two New York State Regents Examination in English Language Arts writing tasks will be utilized to collect data on student performance. The first will be completed by students using a New York State Regents Examination in English Language Arts Part 2 Rubric as a guide during the writing process. The second will be completed by students using a teacher-generated single-
point rubric as a guide during the writing process. All students will be asked to complete a five-minute survey at the conclusion of each of the two tasks regarding demographic information and student feedback on whether they used each rubric during the writing process and rating how helpful they found it as a guide. The risk of lost instructional time will be mitigated in that the task and its inherent lessons are consistent with the North Salem English department writing curriculum and the benefits of the study—determining validity and reliability of an instructional and evaluation instrument and student perceptions of the helpfulness of that instrument—could influence curriculum and instruction.

This research study has been reviewed and approved by Western Connecticut State University’s Institutional Review Board. Participation in this study is completely voluntary, and students who choose not to participate will not have to leave the classroom. Instead, they will be provided an alternative assignment by the classroom teacher that is consistent with curricular objectives of the course. Study participation will not impact a student’s grades or standing in the class. Students who agree to participate will submit all information to their teacher and the researcher will then collect it. Student GPAs will be attained by the researcher through the Guidance Department as an intermediary to shield the researcher from any other confidential information about students. While in any study there is always a risk of loss of confidentiality, privacy will be protected. Student names will be numerically coded. All student identities will be maintained in a secure location to protect confidentiality. Your child’s name, individual performance, and specific feedback will be kept confidential. Results will only be reported in aggregate form.
I wish to thank the parents in the North Salem Central School District for considering participation in this study. If you have any questions, please feel free to contact me.

Sincerely,

David Popken
Sally Dobyns, Pd.D.
Adjunct Professor, EdD in Instructional Leadership

popken001@connect.wcsu.edu  dobynss@wcsu.edu

If you agree to have your child participate in the study, please sign the attached statement below, return it to me by 01/28/19 and keep the attached letter for your records.

Thank you.

David Popken, Ed.D. Candidate
Instructional Leadership
Western Connecticut State University

I am the parent of _________________________________. I acknowledge that Mr. Popken has made clear to me the purpose of this research, identified any risks involved, and offered to answer any questions. I voluntarily grant permission for my child to participate.

Printed Name of Parent: __________________________________________________________

Signature of Parent: ___________________________________ Date: __________
Appendix E - Student Demographic Survey

Department of Education and Educational Psychology
Western Connecticut State University
181 White Street
Danbury, CT 06810
January XX, 2019

Name ________________________________ Date __________

Thank you for completing this survey. Please circle each response.

1. In which English class are you enrolled?
   - o English 12
   - o Advanced Placement English Literature

   *If you answered English 12, move on to #3. If you answered Advanced Placement English Literature, please answer #2.*

2. How many years have been enrolled in honors/advanced placement English in high school?
   - o 1
   - o 2
   - o 3
   - o 4
3. With which gender do you primarily identify?
   - Female
   - Male
   - Transgender
   - Other
   - Prefer not to answer

4. With which ethnicity do you primarily identify?
   - African American
   - Asian
   - Hispanic
   - Native American
   - White
   - Other
Appendix F-First Regents Writing Task

Text-Analysis Response

Your Task: Closely read the text* provided on pages 18 and 19 and write a well-developed, text-based response of two to three paragraphs. In your response, identify a central idea in the text and analyze how the author's use of one writing strategy (literary element or literary technique or rhetorical device) develops this central idea. Use strong and thorough evidence from the text to support your analysis. Do not simply summarize the text. You may use the margins to take notes as you read and scrap paper to plan your response.

Guidelines:

Be sure to:

• Identify a central idea in the text

• Analyze how the author's use of one writing strategy (literary element or literary technique or rhetorical device) develops this central idea. Examples include: characterization, conflict, denotation/connotation, metaphor, simile, irony, language use, point-of-view, setting, structure, symbolism, theme, tone, etc.

• Use strong and thorough evidence from the text to support your analysis

• Organize your ideas in a cohesive and coherent manner

• Maintain a formal style of writing

• Follow the conventions of standard written English

*Text omitted due to copyright

Appendix G-Second Regents Writing Task

Text-Analysis Response

Your Task: Closely read the text* provided on pages 19 and 20 and write a well-developed, text-based response of two to three paragraphs. In your response, identify a central idea in the text and analyze how the author's use of one writing strategy' (literary element or literary technique or rhetorical device) develops this central idea. Use strong and thorough evidence from the text to support your analysis. Do not simply summarize the text. You may use the margins to take notes as you read and scrap paper to plan your response.

Guidelines:

Be sure to:

• Identify a central idea in the text

• Analyze how the author's use of one writing strategy (literary element or literary technique or rhetorical device) develops this central idea. Examples include: characterization, conflict, denotation/connotation, metaphor, simile, irony, language use, point-of-view, setting, structure, symbolism, theme, tone, etc.

• Use strong and thorough evidence from the text to support your analysis

• Organize your ideas in a cohesive and coherent manner

• Maintain a formal style of writing

• Follow the conventions of standard written English

*Text omitted due to copyright

Appendix G-Superintendent’s Consent

Department of Education and Educational Psychology
Western Connecticut State University
181 White Street
Danbury, CT 06810
January XX, 2019

Dear [Redacted]:

I am currently enrolled in the doctoral program for Instructional Leadership at Western Connecticut State University. This program requires that I design and implement a dissertation research study. The purpose of this study, which will require approximately six class periods of instructional time with senior English students, is to determine the validity and reliability of teacher-generated single-point writing rubrics utilized by teachers in the [Redacted] English department. The study will also measure the predictive quality of several variables: student writing scores when using a scaled analytic rubric; gender; student GPA; student course enrollment (AP vs. regular English); students’ feedback on if they used each rubric structure and how much they found each helpful during the writing process.

Two New York State Regents Examination in English Language Arts writing tasks will be utilized to collect data on student performance. The first will be completed by students using a New York State Regents Examination in English Language Arts Part 2 Rubric as a guide during the writing process. The second will be completed by students using a teacher-generated single-
point rubric as a guide during the writing process. Each task will be completed over three class periods. The first and second tasks will take place two weeks apart. All students will be asked to complete a five-minute survey at the conclusion of each of the two tasks regarding demographic information and student feedback on whether they used each rubric during the writing process and rating how helpful they found it as a guide. High school English teachers will be asked to voluntarily participate after school hours in assessing student performance on the tasks. The risk of lost instructional time will be mitigated in that the task and its inherent lessons are consistent with the North Salem English department writing curriculum and the benefits of the study—determining validity and reliability of an instructional and evaluation instrument and student perceptions of the helpfulness of that instrument—could influence curriculum and pedagogy.

This research study has been reviewed and approved by Western Connecticut State University’s Institutional Review Board. To avoid coercion, students and parents will be informed that participation in this study is completely voluntary, and students who choose not to participate will not have to leave the classroom. Instead, they will be provided an alternative assignment by the classroom teacher that is consistent with curricular objectives of the course. Study participation will not impact a student’s grades or standing in the class. Students who agree to participate will submit all information to their teacher and the researcher will then collect it. With parental consent and student assent, student GPAs will be attained by the researcher through the Guidance Department as an intermediary to shield the researcher from any other confidential information about students. While in any study there is always a risk of loss of confidentiality, privacy will be protected. Student names will be numerically coded. All student
identities will be maintained in a secure location to protect confidentiality. Results will only be reported in aggregate form.

I wish to thank administrators in the School District for considering participation in this study. If you have any questions, please feel free to contact me.

Sincerely,

David Popken

Sally Dobyns, PhD

Adjunct Professor, EdD in Instructional Leadership

popken001@connect.wcsu.edu
dobynss@wcsu.edu
If you agree to have your school district participate in the study, please sign the attached statement below, return it to me by (date) and keep the attached copy for your records.

Thank you.

David Popken, Ed.D. Candidate
Instructional Leadership
Western Connecticut State University

I, _________________________________, am the superintendent of __________________. I acknowledge that Mr. Popken has made clear to me the purpose of this research, identified any risks involved, and offered to answer any questions. I voluntarily grant permission for our district’s students and teachers to participate.

Printed Name of Superintendent: _________________________________________________

Signature of Superintendent: ___________________________ Date: ____________
Appendix H-Principal’s Consent

Department of Education and Educational Psychology
Western Connecticut State University
181 White Street
Danbury, CT 06810
January XX, 2019

Dear Mr. DiGrandi:

I am currently enrolled in the doctoral program for Instructional Leadership at Western Connecticut State University. This program requires that I design and implement a dissertation research study. The purpose of this study, which will require approximately six class periods of instructional time with senior English students, is to determine the validity and reliability of teacher-generated single-point writing rubrics utilized by teachers in the English department. The study will also measure the predictive quality of several variables: student writing scores when using a scaled analytic rubric; gender; student GPA; student course enrollment (AP vs. regular English); students’ feedback on if they used each rubric structure and how much they found each helpful during the writing process.

Two New York State Regents Examination in English Language Arts writing tasks will be utilized to collect data on student performance. The first will be completed by students using a New York State Regents Examination in English Language Arts Part 2 Rubric as a guide during the writing process. The second will be completed by students using a teacher-generated single-
point rubric as a guide during the writing process. Each task will be completed over three class periods. The first and second tasks will take place two weeks apart. All students will be asked to complete a five-minute survey at the conclusion of each of the two tasks regarding demographic information and student feedback on whether they used each rubric during the writing process and rating how helpful they found it as a guide. High school English teachers will be asked to voluntarily participate after school hours in assessing student performance on the tasks. The risk of lost instructional time will be mitigated in that the task and its inherent lessons are consistent with the North Salem English department writing curriculum and the benefits of the study—determining validity and reliability of an instructional and evaluation instrument and student perceptions of the helpfulness of that instrument—could influence curriculum and pedagogy.

This research study has been reviewed and approved by Western Connecticut State University’s Institutional Review Board. To avoid coercion, students and parents will be informed that participation in this study is completely voluntary, and students who choose not to participate will not have to leave the classroom. Instead, they will be provided an alternative assignment by the classroom teacher that is consistent with curricular objectives of the course. Study participation will not impact a student’s grades or standing in the class. Students who agree to participate will submit all information to their teacher and the researcher will then collect it. With parental consent and student assent, student GPAs will be attained by the researcher through the Guidance Department as an intermediary to shield the researcher from any other confidential information about students. While in any study there is always a risk of loss of confidentiality, privacy will be protected. Student names will be numerically coded. All student identities will
be maintained in a secure location to protect confidentiality. Results will only be reported in aggregate form.

I wish to thank administrators in the School District for considering participation in this study. If you have any questions, please feel free to contact me.

Sincerely,

David Popken

Sally Dobyns, PhD
Adjunct Professor, EdD in Instructional Leadership

popken001@connect.wcsu.edu
dobynss@wcsu.edu
If you agree to have your school district participate in the study, please sign the attached statement below, return it to me by (date) and keep the attached copy for your records.

Thank you.

David Popken, Ed.D. Candidate

Instructional Leadership

Western Connecticut State University

I, _________________________________, am the principal of [school name]. I acknowledge that Mr. Popken has made clear to me the purpose of this research, identified any risks involved, and offered to answer any questions. I voluntarily grant permission for our school’s students and teachers to participate.

Printed Name of Principal: ________________________________________________

Signature of Principal: ______________________________________ Date: __________
Appendix I-Teacher’s Consent

Department of Education and Educational Psychology
Western Connecticut State University
181 White Street
Danbury, CT  06810
January XX, 2019

Dear [Teacher]:

I am currently enrolled in the doctoral program for Instructional Leadership at Western Connecticut State University. This program requires that I design and implement a dissertation research study. The purpose of this study, which will require approximately six class periods of instructional time with senior English students, is to determine the validity and reliability of teacher-generated single-point writing rubrics utilized by teachers in the [Department Name] English department. The study will also measure the predictive quality of several variables: student writing scores when using a scaled analytic rubric; gender; student GPA; student course enrollment (AP vs. regular English); students’ feedback on if they used each rubric structure and how much they found each helpful during the writing process.

Two New York State Regents Examination in English Language Arts writing tasks will be utilized to collect data on student performance. The first will be completed by students using a New York State Regents Examination in English Language Arts Part 2 Rubric as a guide during the writing process. The second will be completed by students using a teacher-generated single-
point rubric as a guide during the writing process. Each task will be completed over three class periods. The first and second tasks will take place two weeks apart. All students will be asked to complete a five-minute survey at the conclusion of each of the two tasks regarding demographic information and student feedback on whether they used each rubric during the writing process and rating how helpful they found it as a guide. High school English teachers will be asked to voluntarily participate after school hours in assessing student performance on the tasks. The risk of lost instructional time will be mitigated in that the task and its inherent lessons are consistent with the English department writing curriculum and the benefits of the study—determining validity and reliability of an instructional and evaluation instrument and student perceptions of the helpfulness of that instrument—could influence curriculum and pedagogy.

This research study has been reviewed and approved by Western Connecticut State University’s Institutional Review Board. To avoid coercion, students and parents will be informed that participation in this study is completely voluntary, and students who choose not to participate will not have to leave the classroom. Instead, they will be provided an alternative assignment by the classroom teacher that is consistent with curricular objectives of the course. Study participation will not impact a student’s grades or standing in the class. Students who agree to participate will submit all information to their teacher and the researcher will then collect it. With parental consent and student assent, student GPAs will be attained by the researcher through the Guidance Department as an intermediary to shield the researcher from any other confidential information about students. While in any study there is always a risk of loss of confidentiality, privacy will be protected. Student names will be numerically coded. All student identities will
be maintained in a secure location to protect confidentiality. Results will only be reported in aggregate form.

I wish to thank teachers in the School District for considering participation in this study. If you have any questions, please feel free to contact me.

Sincerely,

David Popken
Sally Dobyns, PhD
Adjunct Professor, EdD in Instructional Leadership

popken001@connect.wcsu.edu dobyss@wcsu.edu
If you agree to have your school district participate in the study, please sign the attached statement below, return it to me by (date) and keep the attached copy for your records.

Thank you.

David Popken, Ed.D. Candidate

Instructional Leadership

Western Connecticut State University

---

I, _________________________________, am the principal of ___________________. I acknowledge that Mr. Popken has made clear to me the purpose of this research, identified any risks involved, and offered to answer any questions. I voluntarily grant permission for our school’s students and teachers to participate.

Printed Name of Teacher: ______________________________

Signature of Teacher: ______________________________ Date: _________
EdD in Instructional Leadership
Department of Education and Educational Psychology
Dissertation Registration Form

Student David Peter Popken Date: 04/13/2020

Dissertation Title: THE VALIDITY AND RELIABILITY OF A SINGLE-POINT RUBRIC TO ASSESS STUDENT WRITING PERFORMANCE

Dissertation Committee Members: See attached Dissertation Approval Page

For Office Use Only.

Sally M. Dobyns, Ph.D. Sally M. Dobyns 4/14/20
Dissertation Committee Chair Signature Date

Interim Program Coordinator Signature Date

Joan S. Palladino, Ed.D. May 11, 2020
Interim Dean, School of Professional Studies Signature Date

Christopher Shankle, Ed.D. May 11, 2020
Associate Director, Division of Graduate Studies Signature Date