MEASUREMENT AND FEATURES OF PERSUASIVE WRITING IN UNDERGRADUATE STUDENTS WITH AND WITHOUT WRITTEN LANGUAGE DISORDERS

Stephanie A. Richards

A Dissertation

Submitted to the Graduate College of Bowling Green State University in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2013

Committee:

Lauren A. Katz, Advisor

Wendy Watson
Graduate Faculty Representative

Lynne Hewitt

Elizabeth Burroughs

Miriam Krause
This study utilized a mixed methods design to examine the feasibility of using rating scales to evaluate the writing skills of undergraduate students, the similarities and differences between the persuasive writing samples of undergraduate students with and without written language disorders (WLDs), and the features of persuasive writing that relate to overall writing quality. Both typical undergraduate participants \((n = 50)\) and undergraduate participants with diagnosed WLDs \((n = 4)\) composed persuasive writing samples on the controversial topic of a new printing policy implemented at their university that required students to pay for printing. Persuasive writing samples were scored by three graduate research assistants using holistic and analytic rating scales designed for the current study. Scores on the rating scales were used in both the quantitative and qualitative analyses of the study, and comments provided by the raters about why they selected a particular holistic score for each sample were used for the qualitative analyses. Results of the study revealed problems with inter-rater and intra-rater reliability for both the holistic and analytic rating scales. Additionally, differences were found across several features of writing between the persuasive writing samples of typical participants and participants with WLDs, as well as between the writing samples of participants with high holistic scores and low holistic scores. Research and clinical implications are presented in light of the results of the current study. Limitations of the current study and directions for future research are also discussed.
ACKNOWLEDGMENTS

I would like to start by thanking my advisor, Dr. Lauren Katz, for all of her guidance and support over the last three years. In addition, I would like to thank the other members of my committee – Drs. Elizabeth Burroughs, Lynne Hewitt, Miriam Krause, and Wendy Watson – for all of their valuable feedback and encouragement throughout this process.

I would also like to extend my thanks to all of the faculty and staff in the Department of Communication Sciences and Disorders, as they have all been incredibly kind and helpful during my time at BGSU. Additionally, I have enjoyed getting to know all of the doctoral and bridge students (Elina Banzina, Anna Ehrhorn, Brittany Frazer, Charlie Hughes, Sabiha Parveen, Siva priya Santhanam, Lisa Shattuck, Eric Swartz, Jason Whitfield, and Elizabeth Witter) and have appreciated their willingness to listen and offer suggestions over the years.

I have also been thankful for the suggestions and guidance provided by Dr. Nancy Boudreau and Jason Whitfield regarding the statistical analyses used in this study. Additionally, I have greatly appreciated all of the time and effort given by the undergraduate and graduate research assistants who have helped with various research projects throughout my Ph.D. program (Lauren Everley, Tayler Finsel, Victoria Gora, Lauren Klenk, Laura Knue, Heather Langenkamp, Leah Matuszewski, Kathryn Plath, Cassandra Seemann, Colleen Vetter, and Nancy Walls).

Finally, I would like to thank my family and friends for their love, support, and encouragement throughout the past three years and the rest of my life. I do not know how I would have gotten through this process without you.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Presence of Writing Problems in College Students</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Problems with Existing Diagnostic Writing Measures</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Limited Literature on Assessing and Diagnosing Writing Problems in College Students</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Overview of Literature Review</td>
<td>6</td>
</tr>
<tr>
<td>II</td>
<td>REVIEW OF THE LITERATURE</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Features of Writing</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Writing Process</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Writing Product</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Writing Expectations</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>End of High School Expectations</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>College Expectations</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Summary of Writing Expectations</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Problems with Existing Expectations</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>What is Known about the Writing Abilities of College Students</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Writing of Typical College Students</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Writing of “Basic” College Students</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Differences Between the Writing Abilities of Adolescents and Adults with and without LLDs</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Macrostructure</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Microstructure</td>
<td>33</td>
</tr>
<tr>
<td>Figure/Table</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td>1</td>
<td>History of Disorders and Services of Participants with WLDs</td>
<td>56</td>
</tr>
<tr>
<td>2</td>
<td>Participant Characteristics</td>
<td>57</td>
</tr>
<tr>
<td>3</td>
<td>Means, Ranges, and Standard Deviations of Participants GPAs, ACT Scores, and PPVT-4 Standard Scores</td>
<td>58</td>
</tr>
<tr>
<td>4</td>
<td>Inter-Rater Reliability for 12 Samples Scored Using Original Four-Point Rating Scale</td>
<td>61</td>
</tr>
<tr>
<td>5</td>
<td>Inter-Rater Reliability for Three-Point Rating Scale</td>
<td>62</td>
</tr>
<tr>
<td>6</td>
<td>Distribution of Writing Samples Among Raters for Holistic and Analytic Scoring Procedures</td>
<td>69</td>
</tr>
<tr>
<td>7</td>
<td>Representation of Writing Samples Scored by Each Rater for Intra-Rater Reliability During Holistic and Analytic Scoring Procedures</td>
<td>70</td>
</tr>
<tr>
<td>8</td>
<td>Flowchart of Holistic and Analytic Scoring Procedures</td>
<td>71</td>
</tr>
<tr>
<td>9</td>
<td>Demographic Information for Matched Pairs of Typical Participants and Participants with WLDs</td>
<td>96</td>
</tr>
<tr>
<td>10</td>
<td>Demographic Information for Matched Pairs of Participants with High and Low Holistic Scores</td>
<td>96</td>
</tr>
<tr>
<td>11</td>
<td>Absolute and Reasonable Inter-Rater Reliability for Holistic Scores Before and After In-Depth Training</td>
<td>97</td>
</tr>
<tr>
<td>12</td>
<td>Absolute and Reasonable Intra-Rater Reliability for Holistic Scores Before and After In-Depth Training</td>
<td>98</td>
</tr>
</tbody>
</table>
Means and Standard Deviations of Holistic and Analytic Rating Scale Scores for Participants .......................................................... 108

Means and Standard Deviations of Percentages of Coded Words and Sentences of Participants .......................................................... 109

Prevalence of High and Low Holistic and Analytic Rating Scale Scores in Typical Participants \((n = 50)\) and Participants with WLDs \((n = 4)\) ........................................ 110

Group Differences for Percentages of Word- and Sentence-Level Features and Total Analytic Rating Scale Scores Between Typical Participants and Participants with WLDs .......................................................... 111

Average Number of Comments per Theme and Subtheme for Typical Participants \((n = 50)\) and Participants with WLDs \((n = 4)\) .................................................. 114

Proportions of Types of Spelling Errors Made by Typical Participants \((n = 50)\) and Participants with WLDs \((n = 4)\) .......................................................... 115

Comparison of Performance on Holistic and Analytic Rating Scales of Matched Pairs ........................................................................................................................................ 118

Comparison of Percentages for Matched Pairs ................................................................. 119

Frequency of Positive and Negative Comments in Each Theme for Matched Pairs of Typical Participants \((n = 4)\) and Participants with WLDs \((n = 4)\) ........................................ 120

Frequency of Positive and Negative Comments in Linguistic subthemes for Matched Pairs of Typical Participants \((n = 4)\) and Participants with WLDs \((n = 4)\) ............. 121

Frequency of Positive and Negative Comments in Mechanics subthemes for Matched Pairs of Typical Participants \((n = 4)\) and Participants with WLDs \((n = 4)\) .......... 122
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>Frequency of Positive and Negative Comments in <em>Content</em> subthemes for Matched Pairs of Typical Participants (n = 4) and Participants with WLDs (n = 4) .................. 122</td>
</tr>
<tr>
<td>37</td>
<td>Frequency of Positive and Negative Comments in <em>Quality</em> subthemes for Matched Pairs of Typical Participants (n = 4) and Participants with WLDs (n = 4) .................. 123</td>
</tr>
<tr>
<td>38</td>
<td>Means and Standard Deviations of Analytic Rating Scale Scores for Participants with High (n = 33) and Low (n = 21) Holistic I Scores ......................................................... 124</td>
</tr>
<tr>
<td>39</td>
<td>Means and Standard Deviations of Percentages of Coded Words and Sentences for Participants with High (n = 33) and Low (n = 21) Holistic I Scores ......................... 125</td>
</tr>
<tr>
<td>40</td>
<td>Means and Standard Deviations of Analytic Rating Scale Scores for Participants with High (n = 39) and Low (n = 15) Holistic II Scores ................................................. 126</td>
</tr>
<tr>
<td>41</td>
<td>Means and Standard Deviations of Percentages of Coded Words and Sentences for Participants with High (n = 39) and Low (n = 15) Holistic II Scores ......................... 127</td>
</tr>
<tr>
<td>42</td>
<td>Prevalence of High and Low Analytic Rating Scale Scores in Participants with High (n = 33) and Low Holistic I Scores (n = 21) ......................................................... 128</td>
</tr>
<tr>
<td>43</td>
<td>Prevalence of High and Low Analytic Rating Scale Scores in Participants with High (n = 39) and Low Holistic II Scores ................................................................. 129</td>
</tr>
<tr>
<td>44</td>
<td>Group Differences for Word- and Sentence-Level Percentages and Total Analytic Rating Scale Scores Between Participants with High and Low Holistic I Scores ............. 130</td>
</tr>
<tr>
<td>45</td>
<td>Group Differences for Word- and Sentence-Level Percentages and Total Analytic Rating Scale Scores Between Participants with High and Low Holistic II Scores ............. 131</td>
</tr>
<tr>
<td>46</td>
<td>Average Number of Comments per Theme and Subtheme for Participants with High and Low Holistic Scores ...................................................................................... 133</td>
</tr>
<tr>
<td>Page</td>
<td>Content</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
</tr>
<tr>
<td>47</td>
<td>Frequency of Types of Spelling Errors Made by Participants with High and Low Holistic Scores ......................................................... 135</td>
</tr>
<tr>
<td>48</td>
<td>Comparison of Performance on Holistic and Analytic Rating Scales for Matched Pairs of High and Low Holistic Scores ................................................................. 138</td>
</tr>
<tr>
<td>49</td>
<td>Comparison of Percentages for Matched Pairs of High and Low Holistic Scores ................................................................. 138</td>
</tr>
<tr>
<td>50</td>
<td>Frequency of Positive and Negative Comments in Each Theme for Matched Pairs of Participants with High ( n = 3 ) and Low Holistic Scores ( n = 3 ) ........................................ 139</td>
</tr>
<tr>
<td>51</td>
<td>Frequency of Positive and Negative Comments in Linguistic Subthemes for Matched Pairs of Participants with High ( n = 3 ) and Low Holistic Scores ( n = 3 ) ............................ 140</td>
</tr>
<tr>
<td>52</td>
<td>Frequency of Positive and Negative Comments in Mechanics Subthemes for Matched Pairs of Participants with High ( n = 3 ) and Low Holistic Scores ( n = 3 ) ....................... 140</td>
</tr>
<tr>
<td>53</td>
<td>Frequency of Positive and Negative Comments in Content Subthemes for Matched Pairs of Participants with High ( n = 3 ) and Low Holistic Scores ( n = 3 ) .............................. 141</td>
</tr>
<tr>
<td>54</td>
<td>Frequency of Positive and Negative Comments in Quality Subthemes for Matched Pairs of Participants with High ( n = 3 ) and Low Holistic Scores ( n = 3 ) ......................... 142</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

Presence of Writing Problems in College Students

Writing is a skill that is highly important for academic success at the college level. However, there are currently a high number of students enrolled in colleges and universities across the country that struggle with writing. More specifically, the number of students with learning disabilities (LDs) enrolled in postsecondary institutions is high and continues to rise. It was estimated that there were over 200,000 students with LDs attending colleges and universities in the United States (U.S.) during the 2008-2009 academic year (U.S. Department of Education, 2011). Additionally, it has been reported that the number of students with LDs enrolled in postsecondary institutions in the country has nearly doubled since the 1980s (Wagner, Newman, Cameto, & Levine, 2005). Since researchers estimate that approximately 80% of students with LDs have difficulty with reading and writing (see Rath & Royer, 2002), it seems likely that most college students with LDs will struggle with writing.

Researchers have also reported that problems with reading are more likely to be overcome by individuals with LDs during adolescence and adulthood than problems with writing (Alley & Deshler, 1979). Supporting this claim, studies of adults with a history of language impairment have revealed that writing problems tend to persist into adulthood (e.g., Mortensen, Smith-Lock, & Nickels, 2009; Michelsson, Byring, & Björkgren, 1985). Given that a high number of college students struggle with writing and will likely continue to struggle with writing, it is likely that speech-language pathologists (SLPs) working with college students will be faced with need to assess and treat writing difficulties in this population.
Problems with Existing Diagnostic Writing Measures

However, despite the need, none of the existing standardized diagnostic writing measures sufficiently assess the writing abilities of college students who may present with language disorders and/or learning disabilities (LLDs). One of the biggest problems with the existing diagnostic writing measures normed for college students is that they do not measure writing skills in authentic ways. First of all, few assess writing skills beyond the sentence level, which makes it difficult to determine how examinees would perform on the types of writing assignments typically given at the college level. More specifically, a student may demonstrate adequate or strong skills at the word or sentence level (measured by existing diagnostic writing tests), but struggle to compose an entire writing sample (not measured by most existing diagnostic writing tests). Furthermore, the types of word- and sentence-level tasks that existing diagnostic writing measures use to examine writing skills include discrete activities that stray far from the types of tasks that would be expected of students in a college classroom (e.g., spelling single words, combining sentences, or editing single sentences), making it difficult to know how they would perform on actual college writing assignments. Because existing diagnostic writing measures generally fail to examine students’ writing beyond the sentence level in authentic ways, they do not provide a complete picture of how these students would perform on the types of writing activities typically used in college-level courses.

A second major issue with the existing diagnostic writing measures normed for college students is that they do not measure the types of writing skills that are expected of students at the college level. More specifically, the two existing measures that assess writing skills beyond the sentence level fail to assess the genres of writing most commonly expected of college students. Most writing assignments at the college level require some form of argumentation (or persuasive
writing) (Wolfe, 2011), and persuasive writing skills are heavily emphasized in college writing standards and tests used for undergraduate and graduate admissions (CWPA, 1999; CWPA, 2008; CWPA, NCTE, & NWP, 2011; ACT, 2013; College Board, 2013a; ETS, 2013a).

However, the existing diagnostic writing measures require examinees to compose a paragraph or essay in the narrative genre (rather than the persuasive genre), which individuals tend to have more exposure to throughout life and master earlier and more easily than expository and persuasive genres (e.g., Berman & Katzenberger, 2004; Nippold, 2000). This means that they might perform well when composing a narrative text, but having difficulty composing an expository or persuasive text. Furthermore, most of the existing diagnostic measures focus primarily on examining microstructural elements of writing (e.g., spelling, punctuation, capitalization, or grammar) rather than macrostructural elements (e.g., genre-specific features of text, such as an introduction that includes a statement of the writer’s position and support for the writer’s position in a persuasive text). Because writing standards and admissions tests (e.g., ACT, SAT, and Graduate Record Examination [GRE]) for college students place a heavier emphasis on macrostructural skills than microstructural skills (CWPA, 1999; CWPA, 2008; CWPA et al., 2011), this means that the existing diagnostic measures fail to assess the types of writing skills that are most valued in college-level writing. Additionally, none of the existing writing measures normed for college students assess their ability to complete all phases of the writing process, which is a skill that is expected of students at the college level (CWPA, 1999; CWPA, 2008; CWPA et al., 2011). By failing to examine the types of skills expected of college students in the genres most commonly assigned to them, the existing diagnostic measures do not provide information about how college students would perform on actual college writing tasks.
An additional problem with the existing writing assessments for college students is that many of them have been found to have poor reliability and/or validity. For example, Salvia and Ysseldyke (2007) reported that the *Oral and Written Language Scales* demonstrated adequate reliability, but questionable validity. Additionally, the *Wide Range Achievement Test – 3rd Edition* was found to have so many weaknesses that it is recommended for use only as a screening tool and that even these results be used with caution (Venn, 2007). When measures have poor reliability and/or validity, students’ performance on these measures cannot be considered an accurate representation of their true writing abilities or be used to provide trustworthy diagnoses or recommendations.

Finally, most of the existing diagnostic writing measures are not normed for the full age range of college students (Penner-Williams, Smith, & Gartin, 2009). More specifically, only six of the 19 standardized diagnostic writing measures that are currently available are normed for individuals that fall within the typical age range of undergraduate students (see Appendix A for details about these six measures and three additional assessments that could be used as criterion-referenced measures for undergraduate students). Additionally, three of these measures are only normed for individuals into their early- to mid-twenties. Therefore, these three measures could not be used to assess a large proportion of undergraduate students, as the U.S. Department of Education (2010) reported that more than 40% of undergraduate students were more than 23 years old during the 2007-2008 school year. Although this still leaves three standardized writing measures that are normed for undergraduate students of all ages, none of these remaining measures sufficiently examine the writing skills expected of students at the college level.

Given all of the problems with the existing diagnostic writing measures, SLPs are left with little guidance when it comes to assessing the writing of college students who may present
with disorders of written expression. More specifically, all of the flaws with the existing measures make it difficult (if not impossible) to provide a trustworthy diagnosis of a disorder of written expression in college students. Furthermore, college students’ performance on these measures does not provide SLPs with sufficient information to plan therapy goals and methods, as they fail to measure the types of writing skills expected of college students in authentic ways. Because none of the existing writing measures are able to adequately assess and diagnose writing problems in college students, further research and test development efforts are needed to create and validate a strong and trustworthy writing assessment that can be used by SLPs to examine the writing skills of this population.

**Limited Literature on Assessing and Diagnosing Writing Problems in College Students**

On top of the lack of sufficient writing measures for college-aged students, most of the research on literacy to date has focused on reading rather than writing (Duquès, 1989; Russell, 2006). Furthermore, a majority of the research on writing assessment has focused on younger children rather than adolescents and young adults (see Catts, Bridges, Little, & Tomblin, 2008; Fallon & Katz, 2011; Nippold, 2010). The studies that do focus on assessing writing in college students tend to focus on assessing writing for the purposes of course placement, program/course accountability, or curriculum reform rather than to identify students with writing difficulties (e.g., Brown, 2010; Cho, 2003; Good, Osborne, & Birchfield, 2012; Pagano, Bernhardt, Reynolds, Williams, & McCurrie, 2008; Peckham, 2009). These facts suggest that the literature available to support SLPs in assessing and treating writing problems in college students with writing difficulties is lacking compared to the literature focused on reading, younger students, and writing assessment for purposes other than identifying problems.
Furthermore, research describing the characteristics of typical and atypical writing in college students is limited. While some studies do exist that provide norms for various features of typical college writing (e.g., Crossley, Weston, Sullivan, & McNamara, 2011; Haswell, 2000; Nippold, Ward-Lonergan, & Fanning, 2005), these studies do not include norms for college students with LLDs. This makes it difficult to use the existing literature to determine how to differentiate between college students with and without writing problems. Additionally, many of the studies that do report on the writing characteristics of adolescents and adults with LLDs either do not directly compare their skills to their typically-developing peers (e.g., Dockrell, Lindsay, & Connelly, 2009; Duquès, 1989) or were conducted using participants from other countries (e.g., Harrison & Beres, 2007; Mortensen et al., 2009). While the differences between the writing abilities of college students with and without LLDs in the U.S. and other countries are likely to be similar, studies using U.S. college students are needed to confirm this assumption. These gaps in the research leave SLPs in the U.S. with little information about how to differentiate weak-but-typical college writing from the writing of college students with disorders of written expression.

**Overview of Literature Review**

Although the literature describing the best methods of assessing and diagnosing writing problems in college students is limited, the existing literature will be summarized in the following section. The section will begin with a discussion of the various features of writing that are commonly expected of and assessed in college students. It will then move into a discussion of writing expectations for high school upperclassmen and college students to show the skills that are necessary to be a successful writer at the college-level. Next, the writing abilities of typical and atypical adolescents and adults will be discussed. Then the types of existing writing
measures for older adolescents will be presented in terms of the skills they measure and their strengths and weaknesses. To end the discussion, a summary of the major themes in the literature will be provided, as well as a description of the purposes of the current study.
CHAPTER II
REVIEW OF THE LITERATURE

Features of Writing

Most features of writing are expected to be mastered by college students and are assessed in diagnostic writing measures. Overall, the features of writing can be broken down into two main categories – the writing process and the writing product.

**Writing process.** The stages of the writing process include planning (or pre-writing), drafting (or composing or translating), and reviewing (i.e., editing and revising) (Flower & Hayes, 1981). While researchers may use different terms for these various stages, they typically agree on the types of activities that occur in each of these stages of the writing process. For example, Flower and Hayes suggest that the planning phase involves generating and organizing ideas, as well as setting goals about how to complete the writing task and the types of ideas that need to be presented in the composition. In the translating (or drafting) phase, writers “translate” their ideas into writing, or transfer their ideas to paper. Finally, during the reviewing stage, writers evaluate their writing and make revisions to the text. It is important to note that these phases are not entirely distinct and that writers can move back and forth between these phases throughout the writing process. Additionally, strong writers monitor the processes they use and the progress they make throughout the entire writing process. Once complete, the writing process results in a final written product.

**Writing product.** The writing product refers to the final piece of writing produced by a writer once the writing process is complete and consists of two levels – macrostructure and microstructure.
Macrostructural elements of the writing product include the text-level organizational features of writing for a particular genre (Hall-Mills & Apel, 2012; Heilmann, Miller, Nockerts, & Dunaway, 2010). For example, for narrative texts, macrostructural elements would include the story grammar features that are expected in narrative writing (e.g., characters, plot, or setting). In persuasive texts, macrostructural elements might include an introduction that includes a statement of the writer’s position, support for the writer’s position, acknowledgement of the opposing position, support for the opposing position, presentation of alternative solutions or suggestions to please both sides of an issue, and/or a conclusion that restates the writer’s position and summarizes his or her main points. For persuasive writing, macrostructural elements often vary depending on the instructions given in the writing prompt (e.g., assessment measures or school assignments) or the writer’s purposes for composing the essay (e.g., to convince the writer of something or to suggest an alternative solution). Other features besides organization and text structure that may be considered when examining macrostructural elements of writing include coherence (i.e., clarity or understandability) and cohesion (i.e., connectivity).

Microstructural elements, on the other hand, include the writer’s linguistic form and content at the word, T-unit (i.e., a main clause and any attached subordinate clauses; Hunt, 1970), sentence, and/or sample levels (Heilmann et al., 2010). Some microstructural elements of writing products include productivity (i.e., overall length of a writing sample), lexical diversity (i.e., frequency of unique words or the diversity of vocabulary used in a writing sample), grammatical complexity (i.e., complexity of the utterances in a writing sample; also called sentence complexity or syntactic complexity), grammaticality (i.e., frequency of grammatical errors in a writing sample), spelling accuracy (i.e., frequency of spelling errors in a writing sample), and punctuation usage (i.e., how accurately different forms of punctuation are used,
such as commas, semi-colons, and periods). Just as macrostructural elements differ based on genre, research has suggested that the level and quality of microstructural elements tend to vary from one genre to another. For example, Crowhurst (1980) found that students produced longer T-units in persuasive than narrative writing samples. Therefore, while all of these elements will be present in any genre, a student’s performance in these areas will likely differ depending upon the type of writing task assigned to him or her.

**Writing Expectations**

Because existing writing measures do not provide trustworthy diagnoses of writing disorders in college students or a true picture of how well they would perform on actual college writing assignments, researchers need to work toward developing a tool that can be used to accurately identify writing problems in this population. One of the first steps in developing a diagnostic writing measure for college students is to understand the types of writing skills that are necessary to be successful at the college level. While most of the existing literature on writing assessment and writing expectations focuses on children and adolescents in primary and secondary educational settings, there is some literature that can provide insight into the types of writing skills expected of college students. More specifically, an examination of the writing expectations for high school juniors and seniors can provide some insight into the types of writing skills that should be possessed by incoming college freshmen. Therefore, the types of writing skills outlined in the *Common Core State Standards (CCSS)* for high school upperclassmen and assessed in college entrance exams (e.g., *ACT* and *SAT*) will be discussed in regards to the types of writing abilities that can be expected of college freshmen. Additionally, writing standards for college students and exams used for admittance to graduate programs (e.g., *GRE*) provide information about the types of writing expectations placed upon college students.
at different points in their postsecondary careers. Thus, the existing college writing standards and writing skills measured by the GRE will be presented to show the types of writing abilities expected of college students. The following discussion of the writing skills expected at the college level will shed light on the types of writing skills that should be assessed in college students.

**End of High School Expectations**

**High school standards.** Examining the writing skills expected of high school seniors in the CCSS is one way to determine the types of writing skills that are expected of graduating high school seniors and incoming college freshmen. The CCSS were developed to provide a common framework that could be used across the U.S. to help prepare students for both college and work settings (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). Since their release in June 2010, all but five states in the country have adopted the CCSS. Therefore, the writing and language standards provided in the CCSS for eleventh and twelfth grade students provide a strong idea of the writing skills that should be possessed by most incoming college freshmen across the country.

According to the *Writing* standards in the CCSS, high school juniors and seniors should be able to compose argument, informative/explanatory, and narrative text structures using the appropriate macrostructural elements (Common Core State Standards Initiative, 2012b). Additionally, the *Language* standards indicate that these students should also have mastered the microstructural elements of standard English grammar, usage, capitalization, punctuation, and spelling (Common Core State Standards Initiative, 2012a). Therefore, based on the standards given in the CCSS, it appears that graduating high school seniors and incoming college freshmen
should have strong grasp of the macrostructural and microstructural elements of writing across several genres.

**Achievement and aptitude tests.** Assumptions can also be made about writing expectations for incoming college freshmen based on the writing skills measured by the *ACT* and *SAT*, as students’ scores on these tests influence decisions about college admissions and course placement (*ACT*, 2013; *College Board*, 2013a). Given the scoring procedures of the *ACT* Writing Test, it appears that high school seniors and beginning college freshmen are expected to have a strong grasp of the macrostructural elements of persuasive writing (i.e., appropriate structure, organization, and presentation of ideas) and to be able to effectively use microstructural elements in their writing (i.e., conventions of standard written English) (*ACT*, 2007). Additionally, the *ACT* English Test requires students to read passages and demonstrate their ability to apply usage/mechanics (i.e., punctuation, grammar/usage, and sentence structure) and rhetorical skills (i.e., strategy, organization, and style) while determining what types of changes (if any) need to be made to underlined portions of text. This suggests that the reviewing stage of the writing process (i.e., proofreading, editing, and revising) is an important skill for incoming college freshmen to possess in order to make both macrostructural (i.e., organization and style) and microstructural (i.e., punctuation, grammar, and sentence structure) changes to text. However, it is important to note that very few colleges and universities in the U.S. (i.e., 7.6%) require students to submit scores on the *ACT* Writing Test (*ACT*, 2009), suggesting that most postsecondary institutions in the country do not place a heavy emphasis on the writing abilities of incoming freshmen (or at least the writing abilities suggested by students’ scores on the *ACT*’s writing measures) when making decisions about admissions. Therefore, these
expectations may not accurately represent the expectations that colleges and universities have for their first-year undergraduate students.

On the SAT Essay subtest, students’ scores are based on their ability to compose a writing sample using the macrostructural elements of persuasive writing (i.e., inclusion of position and support for position, clarity, use of critical thinking, organization, cohesion, and coherence) and appropriate microstructure (i.e., appropriate use of language, vocabulary, various sentence structures, grammar, usage, and mechanics) (College Board, 2008; College Board, 2013d). This suggests that incoming college freshmen should possess strong skills in the areas of macrostructure and microstructure, particularly in the genre of persuasive writing. Additionally, the three multiple-choice writing subtests on the SAT (i.e., Improving Sentences, Improving Paragraphs, and Identifying Sentence Errors) require students to identify and/or correct errors in single sentences or paragraphs, suggesting that the reviewing stage (i.e., proofreading, editing, and revising) of the writing process is important for incoming college freshmen. However, an examination of the admissions requirements of colleges and universities across the country revealed that most accept either ACT or SAT scores (College Board, 2013c). Because most schools accept scores on either of these tests and do not require students to submit scores on the ACT Writing Test, it seems likely that they do not strongly consider high school students’ writing skills (or at least the writing skills suggested by students’ scores on the SAT’s writing measures) in their admissions’ decisions. Therefore, the writing skills expected of students on the SAT may not represent the writing expectations placed on college students throughout the country.

College Expectations

College standards. Two different sets of standards for college writing exist (i.e., WPA Outcomes Statement for First-Year Composition and Framework for Success in Postsecondary
These standards have been adopted by numerous universities nationwide. Furthermore, many universities that do not use these sets of standards have opted to create their own standards to reflect the writing skills that they expect from the students enrolled at their particular institution. The examination of these standards provides some insight into the writing skills that are commonly expected of college students and are likely necessary for their success at the college level.

**WPA Outcomes Statement.** The first set of standards, the *WPA Outcomes Statement for First-Year Composition*, was developed by an ad-hoc committee of the Council of Writing Program Administrators (CWPA) in 1999, adopted by the CWPA in 2000, and revised in 2008 (CWPA, 1999; CWPA, 2008). This document was created to provide a standard set of guidelines that could be used by first-year writing instructors to determine which skills should be taught in first-year composition courses, as well as which skills students should have mastered by the end of these courses. The skill areas discussed in this document include rhetorical knowledge (i.e., macrostructure); critical thinking, reading, and writing; processes (i.e., all stages of writing process); knowledge of conventions (i.e., microstructure); and composing in electronic environments. In addition, the *Outcomes Statement* provides suggestions for how instructors should help students learn these skills. As of May 2006, a total of 81 universities across the country were using this document in some form, including defining their first-year composition courses, assessing student skills, training teaching assistants or adjunct faculty, and/or providing a tool for students in these courses (Ericsson, 2006).

**Framework for Success.** The *Outcomes Statement* served as a guide for the development of the *Framework for Success in Postsecondary Writing*, which was released in January 2011 by the CWPA, National Council of Teachers of English (NCTE), and the National Writing Project...
The Framework was developed to provide information about the “experiences, knowledge, and habits of mind that students need to succeed as they begin the first year of college writing” (O’Neill, Adler-Kassner, Fleischer, & Hall, 2012, p. 520). This document was created out of the beliefs that writing instruction should be shared by elementary, secondary, and postsecondary teachers (both at two- and four-year institutions) and that all of these individuals should have common goals and expectations about what students need to be prepared for college-level writing. Using information drawn from the experiences of teachers at all grade levels, the CWPA, NCTE, and NWP worked together to create a statement of the habits of mind and experiences that are essential for students to be successful writers at the college level. The critical habits of mind include curiosity, openness, engagement, creativity, persistence, responsibility, flexibility, and metacognition, and the experiences include rhetorical knowledge (i.e., macrostructure), critical thinking, writing processes (i.e., all stages of the writing process), knowledge of conventions (i.e., microstructure), and ability to compose in multiple environments.

As of July 2012, it was reported that none of the state boards of education had adopted the Framework (McComiskey, 2012). However, it is unclear how many individual universities within the U.S. have adopted the Framework. Additionally, there is no information available about how many of the 81 universities that had adopted the Outcomes Statement as of 2006 chose to switch over to the Framework. Because these documents are nearly identical to each other (with the exception of habits of mind in the Framework), it is possible that many of the universities decided to continue to use the Outcomes Statement rather than adopt the Framework. For example, as of April 2013, the General Studies Writing Program at Bowling Green State University (BGSU) had not updated their Outcomes and Assessment Tool that was adapted from
the Outcomes Statement to match the addition of the habits of mind in the Framework (BGSU GSW, n.d.).

**Summary of college standards.** Based upon the skills listed in the Outcomes Statement and Framework, college students should possess a number of writing skills (CWPA, 1999; CWPA, 2008; CWPA et al., 2011). As far as the writing process, college students are also expected to be able to engage in all aspects of the writing process, including researching, planning and organizing ideas, generating text, editing, revising, and proofreading. There appears to be a strong expectation for students to recognize that writing is a process that takes time and requires multiple drafts. Furthermore, they are expected to be flexible as they progress through the writing process and to have the ability to move back and forth between the different stages as needed.

In terms of macrostructure, the skills that college students are expected to possess relate to genre-specific elements, audience, purpose, context, organization, and text structure (CWPA, 1999; CWPA, 2008; CWPA et al., 2011). In relation to genre-specific elements, students are expected to be able to write in a variety of genres, adapt their writing based upon genre, and write using appropriate genres for their field of study. Students are also expected to be able to identify the appropriate audience for their writing, write for a variety of different audiences (real and imagined), and adapt their writing based upon the audience. For purpose and context, students should possess the skills required to write for a variety of purposes and contexts (real and imagined), demonstrate a focus on a specific purpose in their writing, and adapt their writing to different purposes and contexts. In regard to organization and structure, college students are expected to be able to appropriately format various text types, cite sources, and format paragraphs. In summary, it appears that college students are expected to be flexible in their
writing abilities, as there is a strong emphasis on the need to be able to write in a variety of ways and adapt to specific situations.

The *Outcomes Statement* and *Framework* also specify that college students should possess a number of microstructural skills (CWPA, 1999; CWPA, 2008; CWPA et al., 2011). More specifically, it is stated that their writing should demonstrate a strong grasp of the rules of grammar, syntax, spelling, punctuation, and capitalization. Students are also supposed to be able to use the specialized vocabulary of their field of study in their writing.

Finally, both the *Outcomes Statement* and *Framework* emphasize the importance of being able to apply critical thinking skills in writing (CWPA, 1999; CWPA, 2008; CWPA et al., 2011). More specifically, they suggest that students should be able to use critical thinking skills to consider specific texts and/or situations and then write about them in a variety of ways, such as by synthesizing, responding to, analyzing, critiquing, summarizing, and/or interpreting the issues. Furthermore, students are expected to present multiple viewpoints in their writing, as well as demonstrate the ability to write for multiple purposes and audiences.

**Aptitude and achievement tests.** Because college students who are interested in applying to graduate and business school programs take the *GRE* toward the end of their undergraduate career, the scoring procedures of this test can also be used to make assumptions about the writing expectations that are placed on college upperclassmen. Based on the guidelines given about how to score the two writing tasks (i.e., Analyze an Issue and Analyze an Argument) on the *GRE*, college seniors should possess strong macrostructural skills in persuasive writing tasks (e.g., inclusion of genre-specific elements, organization, cohesion, and coherence) and be able to use microstructural elements appropriately in their writing (i.e., variety of sentence structures and conventions of standard written English) (ETS, 2013d; ETS, 2013e).
However, because not all college graduates choose to pursue a graduate degree, the writing abilities expected of students taking the GRE do not necessarily represent the skills of all college students who are close to the end of their undergraduate careers.

**Summary of Writing Expectations**

Overall, it appears that writing expectations for high school and college students depend upon the measure or set of standards being used. While all of the tests and standards emphasize macrostructural and microstructural elements of the writing product, only the standards fully emphasize the importance of the writing process. In fact, only the multiple-choice writing tests of the SAT and the multiple-choice English test of the ACT examine the writing process. On both of these measures, the writing process is only examined through the use of inauthentic editing and revising activities. However, the CCSS, Outcomes Statement, and Framework place heavy emphasis on the writing process, as all of these standards expect students to be able to complete the writing process across a variety of tasks and to demonstrate the ability to adapt their use of the writing process based on task requirements.

Additionally, while all of the tests and standards have expectations about the writing product, these expectations differ from one test or set of standards to the other. On achievement and aptitude tests such as the ACT, SAT, and GRE, there is a strong focus on persuasive writing, and macrostructural abilities appear to be emphasized more than microstructural skills. However, in the CCSS, Outcomes Statement, and Framework, both macrostructural and microstructural skills are emphasized across a variety of genres. Additionally, these standards expect students to be flexible, as they emphasize the need for students to be able to apply their writing skills across a wide range of situations and adapt accordingly.
In conclusion, the differences in expectations among the existing writing measures and writing standards for high school upperclassmen and college students make it difficult to know what writing skills college students are truly expected to possess. If all of the expectations are combined, it appears that students should be able to carry out all stages of the writing process across a variety of tasks, as well as demonstrate the appropriate use of macrostructural and microstructural skills across various genres. However, achievement and aptitude tests primarily emphasize the writing product, while existing standards emphasize both the writing process and writing product. Additionally, different levels of emphasis are given to macrostructural and microstructural elements on the existing tests and standards. More specifically, achievement and aptitude tests primarily emphasize macrostructural elements in persuasive writing, while standards emphasize both macrostructural and microstructural elements across all genres. Finally, the existing standards (but not achievement and aptitude tests) suggest that college students should also be able to adapt to a variety of different writing situations.

Problems with Existing Expectations

While the existing expectations do provide information about the types of writing skills that should be assessed in college students, there are many challenges in attempting to apply these expectations to the writing samples of this population. One issue is that it is difficult to apply the expectations for the writing skills that should be possessed at the end of high school to college students, as the writing activities occurring in high school and college classrooms are considerably different. For example, Fanetti, Bushrow, and DeWeese (2010) reported that the expectations of high school writing assignments tend to be rigid, as students are expected to write in specific ways to perform well on standardized tests. Conversely, writing in college has been described as more flexible, as the “rules” for good writing vary depending upon the course,
professor, and/or discipline (Acker & Halasek, 2008; Fanetti et al., 2010). Furthermore, most writing assignments in high school have “right” or “wrong” answers (Kidwell, 2005), meaning that students must provide the “correct” information without closely analyzing information to make their own interpretations. However, in college, students are frequently expected to form their own opinions about issues and defend their positions, with these responses judged as “better” or “worse” rather than “right” or “wrong.” Finally, Beil and Knight (2007) found that many types of writing assignments expected of college students are rarely assigned to high school students, including critically analyzing written arguments and identifying problems and offering solutions. Given these substantial differences in expectations, trying to apply expectations about the writing abilities of high school students to what should be expected of college students is not particularly helpful.

Additionally, if the ACT and/or SAT were used to determine the types of writing skills that should be possessed by incoming college freshmen, they may not align with what specific colleges and universities actually expect. Few colleges and universities in the country require students to submit scores on the Writing subtest of the ACT (ACT, 2009). Additionally, most institutions accept ACT (without Writing subtest) or SAT scores (College Board, 2013c), meaning that students’ writing scores on these measures likely do not play a strong role in the decision about whether or not they will be accepted into a college or university. If colleges and universities do not strongly consider students’ performance on these measures when admitting them, they likely are not strongly concerned about whether or not students possess the writing skills necessary to receive high scores on these measures or do not feel that students’ scores on these measures accurately reflect how well they will be able to write at the college level. If any of the high-school-level writing expectations are going to be applied to incoming college
fresmen, it seems that the CCSS may be the best option. Because these standards have been adopted by all but five states in the country and more closely align to the expectations listed in the Outcomes Statement and Framework than the writing skills measured by the ACT and SAT, it seems likely that they would represent the desired abilities of incoming college freshmen better than writing tests that do not seem to be considered highly important by most colleges and universities.

Another issue is the fact that the expectations described in the CCSS, Outcomes Statement, and Framework are broad. Because these standards do not provide specific details about how to measure these skills or the specific skill level that students should achieve in each of the described areas, they are not particularly helpful in determining where students stand in relation to their peers or how to differentiate between the typical and atypical writing in college students.

Additionally, the existing writing expectations have been designed for students at various academic levels. For example, the ACT, SAT, and CCSS focus on the writing skills of high school juniors and seniors, while the GRE measures the writing abilities of college juniors and seniors hoping to pursue a graduate degree. The Outcomes Statement describes skills expected of incoming freshmen, while the Framework focuses on skills that should be mastered by the end of composition courses. Therefore, it would be difficult to apply any of these standards to all college students. Furthermore, none of the existing expectations clearly define the skills expected of college students during each year of their education.

Finally, because the writing standards for college students place a heavy emphasis on the writing process, which is difficult to measure, it is challenging to determine how well college students are able to complete all stages of the writing process. Most of the strategies used during
the writing process take place in the writer’s mind, meaning that they cannot easily be witnessed by an outside observer. This makes it difficult to measure someone’s ability to complete the various stages of the writing process. To understand the strategies that a writer uses during the writing process, an examiner would have to ask the writer to think aloud while engaging in the writing process and/or ask questions after the writing process is complete to determine what types of strategies were used while planning, drafting, and reviewing. With both of these approaches, it is difficult to know whether or not a full picture of a writer’s strategy use during the writing process has been obtained, as it is impossible to determine the accuracy of their responses. Therefore, the challenges of measuring individuals’ abilities to complete the writing process make it difficult to determine how well college students are meeting expectations relating to the writing process.

What is Known about the Writing Abilities of College Students

Over the years, a number of researchers have conducted studies to examine the typical writing abilities seen at various stages in life. However, most of these studies have focused on students in primary or secondary education settings (e.g., Crowhurst, 1980; Crowhurst, 1987; Crowhurst & Pichè, 1979; Harrell, 1957; Hunt, 1965; Hunt, 1970; Knudson, 1992a; Knudson, 1992b; Loban, 1976; McCann, 1989; Rubin & Pichè, 1979; Sun & Nippold, 2012), or adults that were not enrolled in college at the time of the study (e.g., Hunt, 1970). Because a limited number of studies focusing on the writing skills of typical college students exist, it is difficult to know what “typical” writing looks like at this level. Without knowing what typical writing looks like, it is also difficult to identify atypical writing and know when a diagnosis of a written language disorder is justified. In the following sections, the existing literature on the writing abilities of typical college students, “basic” college students, and adolescents and young adults
with LLDs will be discussed to provide some insight into the features of writing that show promise for distinguishing between typical and atypical writing in college students.

**Writing of Typical College Students**

While a number of studies of writing have been published in relation to college students (see Juzwik et al., 2006), few of these studies provide useful information regarding the writing abilities of typical college students. More specifically, few of the existing college writing studies examine the abilities of typical college students to complete all stages of the writing process or use specific macrostructural and microstructural elements in their writing. Instead, they tend to focus on other topics related to college writing, such as the writing abilities of non-native English speakers or the effectiveness of various writing instructional practices. However, the studies that do discuss the specific features of writing in typical college students usually examine persuasive writing samples rather than narrative or expository samples.

One of the studies that examined the writing abilities of typical college students was conducted by Haswell (2000). His study included 64 undergraduate students from Washington State University (WSU) who were freshmen in 1991 and juniors in 1993. An equal number of male and female participants were randomly selected for this study, and an analysis of their junior level grade point averages and achievement on a writing exam showed that they were similar to those of other juniors at WSU. All participants completed persuasive writing samples as placement measures for their first-year and upper-division composition courses. To complete these measures, participants were given two hours to respond to various writing prompts, which all included a paragraph-long quote and instructions for the type of response they were supposed to write about the quote. For this study, two different sets of instructions (i.e., Frame 1 and Frame 3) were chosen to be used with all participants, although the quotes that the participants
received with these instructions differed. Frame 1 required participants to read one of several quotes that discussed a controversial topic and then respond by (1) resolving differences among the opinions of the controversial issue, (2) describing the main issue, (3) summarizing the author’s view of the issue, and (4) comparing the author’s view to at least two other possible views. For Frame 3, participants were required to read one of many quotes that provided a misleading picture of a controversial issue and then respond by (1) suggesting ways that the issue could be presented in a more accurate and honest manner, (2) describing the main issue, (3) summarizing the author’s view, and (4) comparing the author’s view with at least two other possible views.

Results of t-tests revealed significant differences in performance from freshman to junior year for the writing features of productivity (i.e., total number of words), grammatical complexity (i.e., average number of words per sentence and average number of words per clause), semantic development (i.e., percentage of words containing nine or more letters), and overall writing quality of the writing samples (i.e., score on eight-point holistic scale, where “1” was low and “8” was high) (Haswell, 2000). These results indicate that persuasive writing abilities appear to increase in typical college students from the freshman to junior year in the areas of productivity, grammatical complexity, semantic development, and overall writing quality.

Nippold et al. (2005) also conducted a study that examined the persuasive writing skills of college students. In this study, researchers collected persuasive writing samples from 60 adolescents in 11th and 12th grades (mean age = 17;3, range = 16;3 – 18;6), 60 undergraduate and graduate students (mean age = 24;10, range = 19;2 – 43;5), and 60 children (mean age = 11;9, range = 10;6 – 13;5) from western Oregon and northern California. Nearly all of the participants
reported European American ancestry (95%) and all reported that they were native speakers of English. To collect the writing samples, all participants were given a booklet that contained black-and-white images of animals and clowns performing at a circus, as well as lined paper. Participants were then read a persuasive writing prompt that asked them to discuss whether or not they thought that animals performing in circuses was a good or a bad idea, as well as provide support for their opinion. They were then given 20 minutes to compose their writing samples.

Several one-way ANOVAs revealed significant effects for all measures of productivity (i.e., total number of words, total number of reasons, total number of T-units and fragments, and mean length of utterance in words [MLU-W]), one measure of clausal development (i.e., percentage of utterances that contained relative clauses), and three types of literate words (i.e., percentage of words that were adverbial conjuncts, abstract nouns, and metaverbs). Large effect sizes were seen for total words, total reasons, MLU-W, and abstract nouns, while medium effect sizes were seen for total utterances (i.e., T-units and fragments), relative clauses, adverbial conjuncts, and metaverbs. Tukey tests revealed that college students outperformed adolescents and children in total words, total utterances, total reasons, and abstract nouns; college students and adolescents outperformed children in adverbial conjuncts and metaverbs; college students outperformed children in relative clauses; and adolescents outperformed children in abstract nouns and total words. These results suggest that persuasive writing abilities increase in a number of areas from elementary school to college, and that some gains also occur from high school to college.

Finally, Crossley et al. (2011) collected persuasive writing samples from 62 ninth grade students in a suburban area of upstate New York, 70 eleventh grade students in a suburban area of Washington, D.C., and 70 college freshmen at Mississippi State University. To collect
writing samples, the researchers had participants write in response to different persuasive prompts from the SAT. All eleventh grade students responded to the same prompt, while ninth grade students and college freshmen responded to one of two prompts. Participants were given 25 minutes to compose their persuasive writing samples.

To analyze these samples, Crossley et al. (2011) used the Coh-Metrix program, an automated text analysis tool, to obtain calculations that examined performance differences between groups in the areas of productivity (i.e., total number of words and total paragraphs), grammatical complexity (i.e., number of modifiers per noun phrase), semantic development (i.e., word frequency of all words, lexical diversity D, content word overlap, word polysemy [words that have multiple meanings, such as foot or table], and word concreteness), and cohesion (i.e., incidence of positive logical connectives). Additionally, all samples were scored for overall writing quality using the six-point holistic rating scale used to score SAT essays.

ANOVA were run to compare the performance among groups on the lexical indices calculated by the Coh-Metrix program (Crossley et al., 2011). In regard to productivity, these results revealed that college freshmen produced more total words than ninth and eleventh grade students, and eleventh grade students produced more words than ninth grade students. Additionally, college freshmen and eleventh grade students produced more total paragraphs than ninth grade students. For grammatical complexity, the results showed that college students used more modifiers per noun phrase than ninth and eleventh grade students, and eleventh grade students used more modifiers per noun phrase than ninth grade students. Semantic development results revealed that college freshmen used the least frequent words and ninth grade students used the most frequent words, while eleventh grade students did not different in their word frequency use from college freshmen or ninth grade students. Additionally, college freshmen
had higher lexical diversity than eleventh and ninth grade students, and eleventh grade students
had higher lexical diversity than ninth grade students; ninth grade students had the most content
word overlap, followed by eleventh grade students and then college freshmen. College freshmen
also had the least polysemous words, ninth grade students had the most polysemous words, and
eleventh grade students’ use of polysemous words did not differ from college freshmen or ninth
grade students; college freshmen used the most concrete words, followed by eleventh grade and
then ninth grade students. Results for cohesion showed that ninth grade students used more
logical positive connectives than eleventh grade and college freshmen. Finally, ANOVA results
for overall quality revealed that college freshmen outperformed ninth and eleventh grade
students, and eleventh grade students outperformed ninth grade students. These results suggest
that changes in writing abilities occur throughout high school and into college.

Overall, the findings of these studies suggest that growth in persuasive writing abilities
increases from elementary school to college, as well as throughout the college years. Because
changes in persuasive writing abilities continue throughout college, this indicates that diagnostic
writing assessments normed for high school students should not be used to evaluate the writing
skills of college students who may present with writing difficulties. More specifically, college
students’ scores on measures developed for primary or secondary students will likely over-
estimate how well they would perform if compared to other college students. Therefore, it is
important to determine what college-level writing looks like and how to best assess the writing
abilities of college students who may have writing disorders.

Writing of “Basic” College Students

One of the first individuals to closely examine the writing abilities of college students
was Shaughnessy (1977). She used her nine years of experience teaching basic writing
(sometimes referred to as remedial or developmental writing) to open admissions students and an analysis of over 4,000 of their essays to write a book that focused on the writing problems experienced by under-prepared college freshmen. This book contained an introductory chapter followed by seven chapters that each focused on one of the most prevalent problems that she noticed in the writing of her students. Within each of these seven chapters, Shaughnessy provided examples of the problem, an explanation of what likely caused the problem, and suggestions about how to help remedy the problem.

The second chapter of Shaughnessy’s (1977) book was dedicated to a discussion of handwriting and punctuation. In this chapter, she briefly noted that handwriting problems impairing a writer’s ability to convey his or her message are common. She then provided a more detailed discussion of punctuation errors, which described how basic writers in college struggle to use punctuation to communicate their message. Errors in the use of periods, commas, and capital letters were the primary focus of the chapter. Little attention was given to other forms of punctuation, as Shaughnessy noted that these features are usually missing from the writing of basic writers. Overall, she reported that the primary cause behind the punctuation difficulties of basic writers is likely a limited understanding of the differences between spoken and written language.

In the third chapter, Shaughnessy (1977) discussed common errors of syntax made by basic writers, including accidental errors (i.e., errors made unintentionally), blurred patterns (i.e., patterns that incorrectly combine features of multiple patterns, such as combining “At least I can say that…” and “The least I can say is…” into “At least I can say…”), consolidation errors (i.e., errors made when combining and shrinking sentences, such as converting, “I learned a lot in elementary school,” and, “For example, I learned how to read in third grade,” into, “In
elementary school I learned a lot for example how to read I learned in third grade” instead of, “In elementary school, I learned a lot, such as how to read.”), and inversions (i.e., errors made when organizing sentences with subordinate structures, such as, “I think that the smarter a person is how much sooner he’ll get done with school,” instead of, “I think that the smarter a person is, the sooner he’ll get done with school.”). She noted that these errors could stem from limited knowledge of vocabulary, sentence structure, differences between speech and writing, and/or the writing process.

The fourth chapter focused on common errors made by basic writers that readers find disturbing, including problems with verbs (e.g., using the wrong form of “to be” given the number or tense of a sentence, such as, “They was going…”), nouns (e.g., problems using inflection for number or possession, such as, “My three sister…”), pronouns (e.g., using wrong pronoun for number, possession, or case, such as, “They walked to they car.”), and subject-verb agreement (e.g., “There is only two days left in the week.”) (Shaughnessy, 1977). While these types of errors are common in non-native English speakers and speakers of African American English, Shaughnessy noted that they also often appear in the writing of native-born American students who have grown up speaking a dialect besides Standard American English. She also mentioned that typical college writers make a tolerable number of these types of errors in their writing (i.e., five or six in a 300 word essay), but that the number of these errors made by basic writers (i.e., 10 to 30 in a 300 word essay) becomes intolerable to the reader and makes it difficult for them to focus on anything else. Again, she reported that these errors are likely the result of students not recognizing the difference between their speech and formal written English.

Spelling was the focus of chapter five, as Shaughnessy (1977) discussed the different types of spelling errors commonly made by basic writers. She distinguished between accidental
misspellings (e.g., those due to “slips of the pen” or ambiguous letter formations), misspelled words (e.g., those due to the unpredictability of English spelling [“educashun” for “education”], differences between pronunciations and spellings [“aditude” for “attitude”], misuse of homophones [“to” for “two”], unfamiliarity with the structure of words [“instution” for “institution”], lack of mental representations of written words [“after noon” for “afternoon” or “change” for “chance”), and grammatical errors that appear as misspelled words (e.g., “five hat” for “five hats”). It was proposed that these errors could be the result of weak spelling skills, differences between spoken and written language, and lack of visual experience with words.

Chapter six focused on the weak vocabulary skills of basic writers (Shaughnessy, 1977). In this chapter, Shaughnessy noted that basic writers enter college with limited academic vocabularies, which take years to develop and cannot successfully be taught word by word at this level. She found that these weak vocabulary skills often led basic writing students to make many vocabulary errors while writing, including errors of word form (i.e., create words that do not exist or incorrect forms of a word by incorrectly using derivational affixes, such as, “He got home safably [safely].”), word substitution (i.e., substituting phonetically similar words for one another, such as, “I really liked how he floormatted [formatted] his paper.”), and prepositional phrases (i.e., inappropriate use of prepositions, such as, “He should be wary with [of] the dog.”). Additionally, she noted that the writing of basic writers often contains the frequent use of vague nouns (e.g., “They have lots of those things.”), ambiguous pronouns (e.g., “They really liked it.”), basic verbs (e.g., have, be, or get), and basic modifiers (e.g., very, too, best, more, or worst). She claimed that these vocabulary difficulties likely stem from a lack of exposure.

In chapter seven, Shaughnessy (1977) focused on errors made beyond the sentence. For example, she described how basic writers tend to produce their thoughts sentence by sentence,
rather than connecting these ideas throughout an entire passage. More specifically, she stated that they seem to struggle to elaborate on their thoughts and often do not use more than one or two sentences to discuss a single thought. Therefore, their essays tend to be short and full of many thoughts with weak connections and transitions. Shaughnessy thought that many of these problems occur because basic writers believe that their readers understand their thoughts, so they do not provide the appropriate introductions, transitions, or explanations to make their ideas clear to readers.

Finally, chapter eight presented the idea that teacher and learner expectations are strongly related to achievement (Shaughnessy, 1977). Shaughnessy noted that writing is a skill that writers are always learning to do better and that students can make tremendous progress in their writing skills when expectations are high and instruction appropriately meets their needs. Overall, Shaughnessy concluded that there are common patterns to the errors made by basic writers and that, although it takes time, these students can make progress in their writing skills.

While Shaughnessy’s (1977) book was published over three decades ago, the postsecondary educational climate of that time was not unlike our current postsecondary educational climate. In the 1960s, four-year universities began admitting an increased number of students who would not have been considered “college material” in the past. Today, with the increased demand for individuals to earn college degrees in order to obtain decent jobs, more and more individuals are choosing to pursue a postsecondary education. This means that colleges and universities across the country are likely admitting higher numbers of the “basic writers” that Shaughnessy described than ever before. Therefore, it is quite possible that these “basic writing” traits accurately reflect the writing abilities of many of today’s typical college students. This could make it difficult to differentiate between college students who struggle with writing due to
a lack of education or experience versus those who struggle due to underlying language and/or learning difficulties.

**Differences Between the Writing of Adolescents and Adults with and without LLDs**

The literature about the differences between the writing abilities of adolescents and adults with and without LLDs serves as a starting point for identifying whether or not a disorder of written expression is present in college students, as it suggests specific features of writing that would be most helpful in differentiating between typical and atypical writing.

**Macrostructure**

A few researchers have focused on examining differences in the presence and quality of macrostructural elements (i.e., expected content, organization, and structure of a particular genre, as well as cohesion and coherence) between the writing samples of adolescents and young adults with and without LLDs. Overall, these researchers have found that adolescents and adults with LLDs demonstrate problems with various features related to macrostructure. For example, when Dockrell et al. (2009) had adolescents with language impairments complete the *Wechsler Objective Language Dimensions (WOLD)* test that required them to write a letter describing their ideal house, they found that these individuals received scores in the low range for ideas and development, as well as organization. Additionally, Harrison and Beres (2007) found that college students with writing difficulties received lower scores for organization and theme development on the *Wechsler Individual Achievement Test – 2nd Edition (WIAT-2)* compared to their peers without writing difficulties. Finally, Hall-Mills and Apel (2012) examined the difference in macrostructure (i.e., number of macrostructure elements present in each sample) between the narrative and expository writing samples of adolescents (grades six through twelve) with LLDs. In the narrative samples, they looked for elements of characters, plot, sensory detail,
and context. For the expository texts, assignment, logical sequence, introduction, body, and conclusion were examined. Overall, they found that these students used an equal number of genre-specific macrostructural features in each genre. However, the participants included only an average of 2.5 of the 5 macrostructural elements that were examined for each genre, suggesting that they struggle to include all of the necessary macrostructural elements in their writing across genres.

**Microstructure**

Researchers have also found differences in the microstructural abilities of adolescents and adults with and without LLDs. As discussed previously, microstructure refers to semantic, syntactic, and mechanical elements present in all writing genres, such as productivity, lexical diversity, grammaticality, and spelling accuracy. The following discussion will present the findings of researchers who have examined the differences between adolescents and/or adults with and without LLDs in their use of microstructural elements in writing.

**Productivity.** Many researchers have found that individuals with LLDs have deficits in the area of productivity (i.e., how much text a writer generates) when compared to their typically-developing peers. More specifically, on a task that required 11- to 21-year-olds to listen to an expository passage and then write about what they remembered from the passage, Puranik, Lombardino, & Altmann (2007) found that individuals with language impairments produced fewer total words and fewer total T-units than their typically developing peers and their peers with dyslexia. Additionally, Gregg, Coleman, Stennett, and Davis (2002) found that college students with LDs and college students with both LDs and attention-deficit hyperactivity disorder (ADHD) produced fewer total words on an expository writing task than their typical peers. Harrison and Beres (2007) had college students with and without writing difficulties
compose an expository writing sample using the essay task (Prompt A) from the WIAT-2 and found that individuals with a history of writing problems composed fewer words during the 15-minute time limit than their peers without a history of writing problems. All of these findings suggest that individuals with LLDs produce significantly fewer words than their typical peers on expository writing tasks.

In addition, Hall-Mills and Apel (2012) focused on the writing skills of only students with LLDs (grades six through twelve) and examined the differences between their narrative and expository writing samples. They found that these students produced a higher number of total words in narrative than expository texts, but that there was no significant difference between the total number of T-units produced in narrative and expository writing samples. It was suggested that the productivity difference between narrative and expository samples was likely due to differing levels of difficulty and familiarity with the two genres. More specifically, many researchers have suggested that expository text structures tend to have heavier cognitive and linguistic demands than other genres, making them more difficult to master than narrative structures (e.g., Berman & Katzenberger, 2004; Nippold, 2000). Furthermore, they have also reported that students tend to have more exposure to narrative than expository text structures beginning early in life. Therefore, it is likely that these students have an easier time composing narrative texts than expository texts, which impacts some of the microstructural elements of their writing.

**Lexical diversity.** Researchers have found that adolescents and adults with LLDs receive lower scores on measures of lexical diversity (i.e., variety of words used in a writing sample) than their peers without LLDs. More specifically, Gregg et al. (2002) found that college students with LDs and both LDs and ADHD had lower type-token ratios (number of different
words divided by number of total word roots) than their typically-developing peers in their expository writing samples. Additionally, Morris and Crump (1982) conducted a study that required individuals between the ages of nine and 15 years to watch two silent videos and produce two narrative writing samples about what happened in the films. Their results indicated that the students with LDs had lower type-token ratios than their peers without LDs. Therefore, it appears that it is common for individuals with LLDs to use a less diverse vocabulary in their writing than their typical peers.

Finally, when comparing the narrative and expository writing samples of students with LLDs (sixth to twelfth grades), Hall-Mills and Apel (2012) found differences in lexical diversity between the two genres. More specifically, students used a higher number of different words in narrative than expository texts. As with the findings for productivity, it was suggested that this difference was due to the fact that expository texts tend to be more difficult for students to compose and they tend to have less exposure to expository than narrative text structures throughout life (e.g., Berman & Katzenberger, 2004; Nippold, 2000). Therefore, the increased difficulty of composing expository texts results in lower performance on microstructural measures when compared to narrative texts.

**Grammatical complexity.** Several researchers have found that individuals with LLDs show reduced grammatical complexity in their writing when compared to their typically-developing peers. More specifically, Morris and Crump (1982) found that individuals with LDs received lower Syntactic Density Scores (i.e., scored derived by a computer program that considered T-unit length, clause length, embeddings, number of subordinate clauses, and verb expansions) than their peers without LDs on a narrative writing task. Furthermore, Smith-Lock, Nickels, and Mortensen (2009) had participants between the ages of 15 and 84 years write the
story of Cinderella. When they compared the narrative samples of participants with a history of language impairment to unimpaired participants, that found that those with a history of language impairment showed reduced grammatical complexity (measured by mean length of T-unit) compared to their unimpaired peers. These findings suggest that individuals with LLDs demonstrate weaker grammatical complexity skills in the writing compared to their typical peers.

Additionally, Hall-Mills and Apel (2012) found that students with LLDs in grades six through twelve produced a higher number of complex correct sentences in their narrative writing samples than in their expository writing samples, but that there was no difference in the average length of T-unit between these two types of samples. The authors concluded again that the reduced number of complex sentences and the higher number of grammatical errors (measured by complex correct sentences) in the expository samples compared to the narrative samples were due to differences in the students’ experiences with the two text structures. As mentioned previously, it has been suggested that students tend to have less exposure to and more difficulty composing expository than narrative text structures (e.g., Berman & Katzenberger, 2004; Nippold, 2000). These factors influence the level and complexity of the microstructural elements seen in the writing of these two genres.

Grammaticality. Several researchers have found that adolescents and adults with LLDs tend to produce a high number of grammatical errors in their writing. For example, Duquès (1989) found that 25 college students with LDs produced a high number of grammatical errors in their writing (i.e., average of 35.5% of T-units contained errors) when she examined their written responses to a moral dilemma. Additionally, Smith-Lock et al. (2009) found that adolescents and adults with a history of language impairment produced significantly more grammatical errors (calculated as average number of grammatical errors per T-unit) in their narrative writing
samples than their peers without a history of language impairment. Furthermore, Suddarth, Plante, and Vance (2012) found that adults (ages 18 to 42) with language impairment produced a higher percentage of total errors based on total number of words (including spelling, capitalization, punctuation, verb tense, self-corrections, word choice, and other errors) in their narrative writing samples when compared to their peers (ages 18 to 53) with typical language. Finally, Dockrell et al. (2009) examined the writing skills of students with language impairment at ages eight, 11, 12, 14, and 16 years by administering the *WOLD*. On this measure, the students were given 15 minutes to compose a writing sample describing their ideal house. In their samples, these students demonstrated a high number of errors in sentence structure that reduced the clarity and/or fluency of the sentences, as well as scores toward the low end of the scale for grammar.

**Spelling accuracy.** It has also been found that individuals with LLDs have spelling difficulties and produce a high number of spelling errors in their writing. For example, Duquès (1989) found that college students with diagnosed LDs produced a higher percentage of spelling errors in their narrative writing samples than college students without LDs. Additionally, Harrison and Beres (2007) found that college students with writing difficulties received lower scores for mechanics (partially due to a preponderance of spelling errors) on the *WIAT-2* than their peers without writing difficulties. Smith-Lock et al. (2009) also found that adults with a history of language impairment produced a greater number of spelling errors in their narrative writing samples than their peers without language problems. Finally, Suddarth et al. (2012) found that adults with language impairment made more total errors (including spelling, capitalization, punctuation, verb tense, self-corrections, word choice, and other errors) in their narrative writing samples than their peers with typical language.
**Punctuation.** Adolescents and adults with LLDs have also been found to have difficulties using correct punctuation in their writing. More specifically, Smith-Lock et al. (2009) found that adults with language impairment produced a greater number of punctuation errors in their narrative writing samples than adults with typical language. Additionally, Harrison and Beres (2007) found that college students with writing problems received lower scores for mechanics on the *WIAT-2* than their peers without writing problems partially due to a high number of punctuation errors. Finally, it was found that adults with language impairment make more total errors in their narrative writing samples (including spelling, capitalization, punctuation, verb tense, self-corrections, word choice, and other errors) than adults without language impairments (Suddarth et al., 2012).

**Overall Writing Quality**

Finally, the overall quality of the writing samples produced by adolescents and adults with LLDs has been found to be lower than their peers without LLDs. Overall quality is usually measured using a holistic rating scale, which provides an overall score for the entire writing sample without focusing on just one specific element. Gregg et al. (2002) found that college students with LDs or LDs and ADHD received lower scores on a holistic rating scale examining their expository writing samples when compared to college students without disabilities. This particular rating scale did not take into account spelling, punctuation, and grammar (i.e., the original hand-written samples were typed to be error-free by the researchers), suggesting that the overall quality of the writing produced by individuals with LLDs is lower than their peers even when microstructural elements are not considered.
How Writing is Assessed

Over the years, writing has been assessed in a variety of ways. The following section discusses how specific features of the writing process and writing product are commonly assessed. In addition, an overview of the different types of existing writing measures for high school and college students is provided to provide insight into the possible methods of evaluating writing skills at the college level.

How Features of Writing are Commonly Assessed

Writing process. Although writers must engage in the stages of the writing process to complete any writing task, few of the existing formal writing measures actually examine the writing process. Part of the reason why is likely the fact that much of the writing process is invisible to the outside observer. More specifically, while it is possible to make some observations of a writer’s abilities to engage in the writing process by watching him or her write, much of the writing process takes places inside the writer’s mind (Flower & Hayes, 1981). Therefore, examiners must go beyond visual observations to fully understand a writer’s ability to engage in the writing process. For example, to get an idea of what is going on in a writer’s mind during the writing process, the Writing Process Test (Warden & Hutchinson, 1992) has examiners ask the examinee several questions about how often he or she used various writing strategies while planning, composing, and revising his or her writing sample composed for the test. Another possible strategy to obtain information about a writer’s use of the writing process is to have him or her engage in a “think aloud” while writing, or describe what he or she is thinking as she writes (e.g., Flower & Hayes, 1981).

Writing product. The writing product can be evaluated at two levels – macrostructure and microstructure. The most common methods for evaluating macrostructure include simply
noting the presence or absence of various features of a text (e.g., story grammar elements for narratives) or making judgments about the quality of each of the included macrostructural elements in a text (Heilmann et al., 2010).

To evaluate microstructure, each microstructural element is typically examined individually. To measure productivity (i.e., overall length of a sample), researchers and diagnostic measures commonly count the total number of words (e.g., Gregg et al., 2002; Hall-Mills & Apel, 2012; Harrison & Beres, 2007; Puranik et al., 2007) or total number of T-units in a sample (e.g., Hall-Mills & Apel, 2012; Puranik et al., 2007). Additionally, some researchers have counted the number of different words (e.g., Hall-Mills & Apel, 2012) or total number of paragraphs in a sample (e.g., Crossley et al., 2011) to examine productivity.

Lexical diversity (i.e., diversity of the vocabulary in a sample) is commonly measured using type-token ratio (i.e., number of different words divided by number of total words) (e.g., Gregg et al., 2002; Morris & Crump, 1982) or the number of different words in a sample. However, several researchers have found that number of different words is a stronger measure of lexical diversity than type-token ratio, as it shows more developmental change (i.e., continued growth as individuals mature) and is better able to differentiate between individuals with and without LLDs than type-token ratio (see Scott & Windsor, 2000).

Grammatical complexity (sometimes called sentence complexity or syntactic complexity) has been measured using mean length of utterance (i.e., total number of words divided by total number of sentences or T-units) (e.g., Hall-Mills & Apel, 2012; Smith-Lock et al., 2009), number of clauses per T-unit (i.e., total number of clauses divided by total number of T-units), and percentage of complex sentences. Grammaticality (or grammatical accuracy), on the other hand, has been measured by calculating the percentage of sentences or T-units that are free of or
contain grammatical errors (e.g., Duquès, 1989) or the average number of grammatical errors produced per sentence or T-unit (e.g., Smith-Lock et al., 2009).

The features of spelling accuracy and punctuation usage are sometimes combined into a single measure that focuses on mechanics (i.e., spelling, punctuation, and capitalization) or total errors (e.g., Suddarth, Plante, & Vance, 2012). Individual measures of spelling accuracy examine a writer’s ability to spell words correctly by calculating the percentage of words misspelled in a writing sample (e.g., Duquès, 1989), while individual measures of punctuation usage examine how accurately individuals use various forms of punctuation in their writing (e.g., commas, semicolons, ending punctuation, and apostrophes) by measuring the average number of punctuation errors per sentence or T-unit (e.g., Smith-Lock et al., 2009) or the percentage of sentences or T-units containing punctuation errors.

Finally, the overall quality of writing samples is usually examined using holistic rating scales. Holistic scales provide an overall score for the entire writing sample without focusing on just one specific element (e.g., Gregg et al., 2002) and may take into account features of macrostructure and/or microstructure.

**Existing Writing Measures: Diagnostic versus Achievement and Aptitude Tests**

As mentioned in the introduction, few formal diagnostic measures currently exist to evaluate the writing skills of college-aged students who may present with writing problems, and there are many problems with these existing measures. However, achievement and aptitude tests for high school upperclassmen and college students (i.e., ACT, SAT, and GRE) measure writing skills using more authentic methods (i.e., composing essays) and genres (i.e., persuasive writing) than the existing diagnostic measures. Therefore, these types of tests may serve as models for the types of writing tasks that should be included on diagnostic writing measures for college
students. The following sections will describe existing diagnostic and achievement/aptitude writing tests for college students to reveal the strengths and weaknesses of these measures, as well as provide insight into methods and features that should be considered in the development of diagnostic writing measures for college students.

**Diagnostic tests.** Currently, several diagnostic writing measures are available to identify individuals who have disorders of written expression. However, as mentioned in the introduction, only six of these measures are normed for college-aged students. For two of these tests, the only aspect of writing measured is spelling of single words (i.e., *Spelling Performance Evaluation for Language and Literacy – 2nd Edition* [SPELL-2]; Masterson, Apel, & Wasowicz, 2006; *Wide Range Achievement Test – 4th Edition* [WRAT-4]; Wilkenson & Robertson, 2006). Another two of these measures only examine writing skills at the word and sentence level (i.e., *Woodcock-Johnson II Normative Update* [WJ III NU]; Woodcock, McGrew, Shrank, & Mather, 2007; Woodcock, Shrank, Mather, & McGrew, 2007; *Test of Adolescent and Adult Language – 4th Edition* [TOAL-4]; Hammill, Brown, Larsen, & Wiederholt, 1994) using tasks such as spelling dictated real words and/or nonwords, generating sentences using given words or phrases, and combining sentences. Additionally, one measure examines writing at the word, sentence, and paragraph levels (i.e., *Oral and Written Language Scales – 2nd Edition* [OWLS-2]; Carrow-Woolfolk, 2011). However, because none of these five measures require examinees to compose an entire writing sample, they do not provide information about well college students would perform on actual writing assignments. Neither has it been investigated whether performance on these types of measures correlates with success in college writing.

The *Kaufman Test of Educational Achievement – 2nd Edition* (KTEA-2) is the only diagnostic writing measure normed for college-aged students that assesses writing at the essay
level (Kaufman & Kaufman, 2004). However, there are problems with the essay task on the *KTEA-2*. First of all, it assesses writing in the narrative genre, which is a genre less commonly expected of college students than expository or persuasive genres. Additionally, the task requires examinees to write about a story that they helped finish during the completion of previously-administered discrete writing activities. For example, the story for examinees in 6th grade or above is about spending a day on a movie set. Before they compose a story, examinees are required to complete several tasks relating to the story that they will have to write later, such as editing a brief news release about a film or composing single sentences about pictures showing events from the day. Therefore, most of the content for the story has already been provided to examinees; they do not have an opportunity to demonstrate their ability to compose an original piece of writing using their own ideas.

In conclusion, because none of the existing diagnostic writing measures for college students examine their ability to compose the types of writing activities commonly assigned at the college level, they do not provide a true picture of how well college students would perform on authentic writing activities. Thus, other measures will need to be developed in order to accurately identify college students with writing disorders.

**Achievement and aptitude tests.** Currently, a number of the achievement and aptitude tests that are used in the process of admitting students to undergraduate or graduate programs measure writing skills. More specifically, many colleges across the country look at high school students’ performance on the *ACT* or *SAT* when making decisions about admission and course placement (ACT, 2013; College Board, 2013a). Both of these tests include subtests that measure writing. Additionally, many graduate programs require students to take the *GRE* General Test (and/or Subject Tests) and consider their scores on this measure when making decisions about
admissions to graduate or business school (ETS, 2013a). All three of these tests require students to complete more authentic writing tasks than the existing diagnostic writing measures normed for college students, making them good examples of what should be examined during the assessment of the writing abilities of college students with writing difficulties.

**ACT.** Students taking the *ACT* have the option of taking the *ACT* (No Writing) or the *ACT* Plus Writing (ACT, 2007). The optional Writing Test was introduced in February 2005, but is not required by most colleges in the U.S. More specifically, based on information presented in the *ACT*’s search tool titled “What Colleges Have Decided about the Writing Test”, only 7.6% (283 of 3,728) of colleges and universities require *ACT* writing scores for admissions (ACT, 2009).

Students who opt to take the *ACT* Writing Test are given 30 minutes to compose an essay in response to a persuasive writing prompt that explains an issue and provides two different points of view on this issue (ACT, 2007). The students are required to pick a position on the topic and support their position. The position chosen can be one of the two positions presented in the prompt or a different position that is developed by the student. Essays are scored by two independent raters using a six-point holistic rating scale. The scores of the two raters are combined to create a score ranging from two to 12, which is then converted to an *ACT* scaled score ranging from one to 36. The raters who score the writing samples each provide a holistic score based upon their overall impression of each writer’s ability to “make and articulate judgments; develop and sustain a position on an issue; organize and present ideas in a logical way; and communicate clearly and effectively using the conventions of standard written English (ACT, 2007, p. 16).”
The development of all the components of the ACT Writing Test was a lengthy process that involved careful examination of existing practices of writing instruction and assessment in secondary and postsecondary setting across the country (ACT, 2007). Additionally, a panel of the nation’s experts on writing instruction, writing assessment, ESL writing, and developmental writing was used to determine the genre of writing to be measured, the format and structure of the prompt, and the criteria to be included on the holistic scoring rubric. Revisions and refinements were made to the scoring rubric during an extensive process of scoring student writing samples to ensure that the scoring descriptions were clear and resulted in reliable scores. As a result of this extensive process, the developers of the ACT Writing Test feel that students’ scores on this measure provide a fairly accurate representation of the writing skills and knowledge that they possess, which will, in turn, assist colleges and universities in making decisions about admissions and course placement.

**SAT.** The SAT contains four writing subtests, including Improving Sentences, Improving Paragraphs, Identifying Sentence Errors, and Essay (College Board, 2013b). The first three subtests are multiple-choice in nature, while the fourth subtest requires students to compose a persuasive essay. Raw scores on the multiple choice subtests and the Essay subtest are used to derive a composite score (ranging from 200 to 800) for the Writing portion of the test (College Board, 2013i). Additionally, the raw scores of the three multiple-choice subtests are converted to individual subtest scaled scores ranging from 20 to 80.

The Improving Sentences subtest of the SAT requires students to read a sentence and select one response from five options that provides the best way to write an underlined portion of the sentence (College Board, 2013g). On the Improving Paragraphs subtest, students must read short passages and make decisions about how to best revise select sentences in the passages; five
multiple-choice options are presented and students must select the option that provides the best revision (College Board, 2013f). To complete the Identifying Sentence Errors subtest, students must read single sentences and then make a decision about which of four underlined words or phrases in the sentence needs to be changed to correct the sentence or if the sentence is correct as written (College Board, 2013e).

The SAT’s Essay subtest requires students to compose an essay on a given topic within a 25-minute time limit (College Board, 2013h). Students must pick a side on the topic and defend their position by providing clear reasoning for their position and examples. Essays are scored by two independent raters (i.e., high school and college teachers with experience and training) using a six-point holistic rating scale. Scores from the two raters are combined, so students can receive scores from two to 12 points on the Essay subtest. Factors that are considered while scoring writing samples include level of adherence to the given topic, development of a position on the issue (i.e., clarity, demonstration of critical thinking, use of examples and other evidence), organization, cohesion and coherence, use of language and vocabulary, use of various sentence structures, and presence of errors (i.e., grammar, usage, and mechanics) (College Board, 2008; College Board, 2013d).

GRE. The GRE is taken by prospective graduate and business school applicants. Scores on the GRE are used by graduate programs as they make decisions about admissions (ETS, 2013a). During the 2012-2013 school year, a total of 2,834 institutions and fellowship sponsors in the United States were approved to accept GRE scores (ETS, 2012). Subtests on the GRE General Test include Verbal Reasoning, Quantitative Reasoning, and Analytical Writing (ETS, 2013f).
The Analytical Writing subtest requires examinees to complete two separate 30-minute timed writing tasks, including Analyze an Issue and Analyze an Argument (ETS, 2013b). For the Analyze an Issue task, examinees are given a statement and asked to pick a side on the issue, provide support for their position, and discuss the opposing side of the issue. Depending on the instructions given, the examinee may be required to consider ways that the provided statement may or may not be true and discuss how these factors influence their position; explain situations where adopting the given recommendation would or would not be beneficial and how these situations influence their position; provide and refute the strongest arguments that could be used to challenge their position on the provided claim; explain which provided view on the issue most aligns with their opinion, provide support for their opinion, and explain why they do not support the opposing opinion; explain how much they agree or disagree with the provided claim and discuss the reasons behind the given claim; or provide and support their position on a given policy, discuss the possible problems that could occur if the policy was implemented, and explain how these issues influence their position on the policy.

On the GRE’s Analyze an Argument task, examinees must read a short passage that provides a position on an issue that is backed with support and evidence (ETS, 2013b). After reading this passage, examinees are asked to evaluate the soundness of the author’s argument. Depending on the instructions given, examinees may be required to focus on specific evidence that would been needed to strengthen or weaken the argument; the assumptions that the argument relies on and how the argument would be impacted if these assumptions were proven wrong; questions that need to be answered to determine the reasonableness of the given argument, advice, recommendation, or prediction and how the answers to these questions would help in the evaluation of the provided sample; or alternative explanations that could challenge the given
Both the Analyze an Issue and Analyze an Argument task are scored in the same manner (ETS, 2013c). For both tasks, holistic scores are awarded in half-point increments on a zero-to-six-point scale by one or two trained raters. On the computer-based test, each sample is scored by one rater and a computer program. If the scores agree, then the original human rater score is given. If they disagree, a second rater is brought in and the average of the two human rater scores is rounded to the nearest half-point and given as the final score. For the paper-based test, each sample is scored by two raters. Scores that are in within one point of each other are averaged and rounded to the nearest half-point for a final score. If there is more than a one-point discrepancy, a third rater is brought in to determine the final score. To obtain an overall score for the entire Analytical Writing subtest, the scores on the two separate tasks are averaged and rounded to the nearest half-point.

When determining which holistic score to award to a sample, raters are required to consider several different features of writing. Some of the features are considered for both the Argument and Issue tasks, while other features are unique to each individual task. Features that are considered when awarding holistic scores for both tasks include adherence to the instructions of the assigned task, flow and conciseness of ideas, effectiveness of vocabulary usage, effectiveness and variety of sentence types used, and capability of using standard conventions of written English (ETS, 2013d; ETS, 2013e). For the Argument task, unique features that are considered include clarity and insightfulness of presented issues, clarity of the development of ideas, logicalness of organization, connectedness of ideas, use of transitions, quality and level of support given for main ideas, (ETS, 2013d). Features unique to the scoring of the Issue task
include clarity and insightfulness of position taken on the issue, level of development of position, persuasiveness of provided support, level of focus and organization of ideas, and logical connectedness of ideas (ETS, 2013e).

**Summary of writing measures.** While there are diagnostic writing measures normed for college students, they fail to measure the types of writing tasks expected of college students in authentic ways. Therefore, they provide little information about how well college students would perform on actual writing tasks in a college classroom. The existing achievement and aptitude tests for high school and college students measure writing in more authentic ways than the diagnostic measures, but provide limited information about the specific strengths and weaknesses of students’ writing abilities. More specifically, achievement and aptitude tests provide holistic scores for students’ writing abilities, meaning that it is impossible to determine which features of writing would need to be addressed in therapy for students who struggle with writing. Overall, the tasks required of students on achievement and aptitude tests provide models for the types of tasks that should be required of college students on diagnostic writing measures, although different methods of scoring the writing samples would be needed to provide enough information for SLPs to design individualized therapy plans for college students with writing difficulties.

**Summary of Major Themes**

An analysis of the literature on writing assessment and existing writing measures shows that writing can be assessed in a variety of ways. For example, features of the writing process and/or the writing product can be evaluated using formal or informal measures. However, while several standardized writing measures exist, none of them adequately assess the writing abilities
of college students in a way that allows for the accurate identification of writing disorders and the appropriate guidance for writing skills that should be targeted during intervention.

Although there are not any strong diagnostic writing measures for college students, the existing literature relating to college-level writing does provide some information about the writing skills expected of typical college students. For example, the existing writing standards for college students indicate that they are expected to have mastered many macrostructural and microstructural writing skills and have flexibility in their use of these skills (CWPA, 1999; CWPA, 2008; CWPA et al., 2011; ETS 2013d; ETS, 2013e). More specifically, they are expected to be able to apply their writing skills in a variety of situations and have the ability to adapt accordingly. Knowing what skills are expected of college students provides some insight into the types of writing skills that should be assessed in this population.

However, while standards exist for college-level writing, literature describing the features of “typical” college-level writing is limited. The literature that does exist suggests that writing skills continue to develop from high school to college, and even through college (e.g., Crossley et al., 2011; Haswell, 2000; Nippold et al., 2005). Additionally, Shaughnessy’s (1977) work provides some insight into the writing skills of some of today’s college students, as the increasing number of students admitted into college likely include some of the basic writers that Shaughnessy described in her book. Differentiating between students with writing problems due to underlying language and/or learning difficulties and basic writers could be challenging, as the characteristics of basic writing that Shaughnessy described resemble many of the same challenges faced by individuals with writing disorders.

While the features of “typical” college writing remain unclear, there are several studies that indicate that the writing skills of adolescents and adults with LLDs are weaker than their
typical peers. Based on the existing literature, it appears that the writing samples of individuals with LLDs contain weaker macrostructural and microstructural elements than their typically-developing peers. More specifically, adolescents and adults with LLDs appear to have weaker macrostructural skills than their typical peers in the areas of idea development (Dockrell et al., 2009), organization (Dockrell et al., 2009; Harrison & Beres, 2007), and theme development (Harrison & Beres, 2007). Microstructurally, they appear to have weaknesses with productivity (Gregg et al., 2002; Harrison & Beres, 2007; Puranik et al., 2007), lexical diversity (Gregg et al., 2002; Morris & Crump, 1982), grammatical complexity (Morris & Crump, 1982; Smith-Lock et al., 2009), grammaticality (Dockrell et al., 2009; Duquès, 1989; Smith-Lock et al., 2009; Suddarth et al., 2012), spelling accuracy (Duquès, 1989; Harrison & Beres, 2007; Smith-Lock et al., 2009; Suddarth et al., 2012), and punctuation (Harrison & Beres, 2007; Smith-Lock et al., 2009; Suddarth et al., 2012). Finally, adolescents and adults with LLDs have been found to produce writing samples with weaker overall writing quality when compared to their typical peers (Gregg et al., 2002). Given the weaknesses in these areas, it seems likely that an examination of skills in these areas would assist in differentiating between the writing of college students with and without disorders of written expression.

How Present Study Will Extend Literature

The results of the current study will extend the literature about the persuasive writing skills of typical college students and college students with written language disorders (WLDs) in a number of ways. First of all, results regarding the ability of holistic and analytic rating scales to reliably measure persuasive writing skills at the college level will provide information about whether or not these are acceptable methods for assessing the writing abilities of college students. Because the current study utilized an authentic writing task (i.e., composing an essay)
in a genre commonly assigned at the college level (i.e., persuasive), the results discussing the effectiveness of the rating scales will provide insight about possible methods of measuring writing skills at the college level in ways that more accurately reflect college students’ writing abilities on the types of writing activities they will be assigned than existing diagnostic writing measures. Additionally, comparisons between the writing samples of typical college students and college students with WLDs will provide more information about the similarities and differences between these groups. This information will help begin to show what atypical writing looks like at the college level, which, down the road, will assist in identifying college students with disorders of written expression. Finally, the results discussing the features of writing that relate to overall writing quality will provide some guidance to clinicians who are trying to determine which features of writing to target in therapy to make the largest gains in overall writing quality.

**Purpose of Current Study**

Overall, the primary purposes of the current study were to learn more about the feasibility of using rating scales to measure college-level writing and the features of persuasive writing in undergraduate students. More specifically, it was hoped that examining the reliability of holistic and analytic rating scales would provide insight into the possibility of using these measures to examine the writing abilities of undergraduate students who may present with writing disorders. Furthermore, it was anticipated that the similarities and differences in the writing features of undergraduate students with and without WLDs would help determine how to best differentiate between typical and disordered writing at the college level. Additionally, it was expected that these findings would also provide information about which features of writing should be included on diagnostic measures used to identify college students with writing disorders.
Finally, the examination of the features of writing relating to overall writing quality was expected to provide guidance to SLPs regarding which features of writing should be targeted in therapy to result in the greatest improvements to overall writing quality. The following research questions were addressed in the current study to help achieve these goals:

1. To what extent is it possible to reliably measure the persuasive writing of undergraduate students using holistic and analytic rating scales?

2. Which features of persuasive writing seem to best distinguish between typical undergraduate students and undergraduate students with WLDs?

3. Which features of persuasive writing relate to overall quality of writing among typical undergraduate students and undergraduate students with WLDs?
CHAPTER III

METHODS

Subject Selection

Several different methods of subject recruitment were used to locate participants at a four-year, mid-sized public university in the Midwest United States. To recruit three typical participants used to collect pilot data, an announcement was made at a monthly meeting of the university’s National Student Speech Language and Hearing Association (NSSLHA) chapter by the group’s vice president (see Appendix B). Additionally, the vice president of the university’s NSSLHA chapter sent an email to all members of this group. The email and announcement provided a brief description of the study and contact information for the primary investigator. Individuals interested in participating in the study were asked to email the primary investigator. Once an email was received by the primary investigator, a data collection session was scheduled for each participant.

Following the pilot study, an additional 50 typical participants were recruited through two sections of the Introduction to Communication Disorders course at the same university. To recruit these participants, an announcement was made in class by the course instructor, who provided students with the option of participating in the study or completing an alternative assignment for extra credit. All students were given scheduling sheets to record their top three preferred times for participation in the study. Three undergraduate research assistants reviewed these scheduling sheets to create a schedule of participants. To get started, 50 interested students were contacted via email to schedule data collection sessions and all additional students who expressed interest in participating in the project were placed on a waiting list. An initial email that provided information about the scheduled data collection session was sent to each of the first
50 participants. Each participant was asked to respond to this initial email to confirm that the scheduled day and time would fit into his or her schedule. After a confirmation email was received from each participant, two additional emails were sent. The first email served as a confirmation of the scheduled appointment and the second was a reminder to the participant about the upcoming data collection session. When a participant failed to attend his or her scheduled session, an additional student from the waiting list was contacted until a total of 50 data collection sessions had been conducted.

A total of six college students (four undergraduate and two graduate) with diagnosed WLDs were recruited through the speech and hearing clinic at the same university. Each of these participants had come to the clinic for an evaluation of his or her language and literacy skills. At the beginning of each diagnostic session, the participant was asked if he or she would be interested in participating in the study. A brief description of the study was provided and the participant was informed that his or her decision regarding participation would not impact the diagnostic procedure. When the participant agreed to participate, a consent form was provided. The participant was given time to review the consent form, ask questions about the study, and sign the consent form.

**Participants**

A total of 59 participants were recruited for this study, including 53 college students recruited as typical participants and six recruited as having WLDs. Two participants with diagnosed WLDs were excluded from all analyses of this study because they were graduate rather than undergraduate students. Writing samples from the remaining 57 undergraduate participants were used to answer the first research question, as this question addressed inter-rater and intra-rater reliability rather than specific features of writing of participants.
To answer the second and third research questions, three participants recruited as typical undergraduate students were excluded, as these questions focused on the specific writing skills of various participant groups. The excluded participants included two of the typical participants who reported a history of language and/or learning disability and one typical participant who reported being a high school senior enrolled in college classes.

Table 1 provides background information on the educational difficulties and services of the four participants with WLDs, while Tables 2 and 3 provide demographic information for the 54 participants included in all analyses of the current study.

Table 1

<table>
<thead>
<tr>
<th>Disorder or Treatment</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Previous diagnoses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning disability</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attention deficit hyperactivity disorder</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dyslexia</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asperger’s</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td><strong>Prior use of special education services</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elementary school</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle school</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Junior high</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accommodations, but no service (7th – 12th grade)</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td><strong>Prior professional services sought</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writing</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Accommodations through university disability services</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Currently</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In past</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
### Table 2

**Participant Characteristics**

<table>
<thead>
<tr>
<th></th>
<th>Typical Participants</th>
<th>WLD Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Numbers</strong></td>
<td>$n$</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>44</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>4</td>
</tr>
<tr>
<td><strong>Year in School</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Freshman</td>
<td>26</td>
<td>3</td>
</tr>
<tr>
<td>Sophomore</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>Junior</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Senior</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-20</td>
<td>45</td>
<td>2</td>
</tr>
<tr>
<td>21-25</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>26-30</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>31-35</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>36-40</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>46</td>
<td>2</td>
</tr>
<tr>
<td>Black/African American</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Hispanic/Latino(a)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>No Response</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Current Major</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communication disorders</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>Intervention specialist</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Accounting</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Special education</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Mild/moderate intervention</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Physical therapy</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Violin performance</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Applied health science</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Adolescent-young adult education</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Psychology</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Early childhood education</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Political science</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Gerontology</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Sports management/marketing</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Medical technology</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Undecided</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Not reported</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3

Means, Ranges, and Standard Deviations of Participants’ GPAs, ACT Scores, and PPVT-4

Standard Scores

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Typical M</th>
<th>Range</th>
<th>SD</th>
<th>WLD M</th>
<th>Range</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPA</td>
<td>3.27(^a)</td>
<td>1.6 – 4.0(^a)</td>
<td>0.60(^a)</td>
<td>2.94(^b)</td>
<td>2.4 – 3.5(^b)</td>
<td>0.80(^b)</td>
</tr>
<tr>
<td>ACT Score</td>
<td>23.18(^c)</td>
<td>18 – 29(^c)</td>
<td>3.13(^c)</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>PPVT-4 Standard Score</td>
<td>103.00</td>
<td>80 – 123</td>
<td>10.08</td>
<td>98.75</td>
<td>94 – 104</td>
<td>4.27</td>
</tr>
</tbody>
</table>

Note. ACT scores were not provided by participants with WLDs. Participants who were freshmen were unable to report current college GPAs, as data collection occurred during the fall semester and they had not yet received any final grades at the college level.

\(^a\) n = 33. \(^b\) n = 2. \(^c\) n = 44.

Procedures

Writing Prompt Development

One of the tasks required of participants in this study was to write a response to a persuasive writing prompt. The prompt used in this study was adapted from a persuasive writing prompt originally developed by Dr. Lauren Katz in the fall of 2009 for use in her specialty language and literacy clinic. In the clinic, this prompt was used to evaluate the writing skills of college students who reported language and/or literacy difficulties. The initial prompt asked students to write about their opinion of a new printing policy that their university was considering putting into action, which would require students to pay for printing to help the university save money. Because it was found that students had strong opinions about this issue, it was decided that this topic would also be used in the prompt for the current study.

To adapt the original prompt for use in this study, four changes were made. Firstly, the prompt was re-worded to reflect the fact that the policy had been enacted (rather than was
planning to be enacted). Secondly, an outline of expectations was added so that it would be clear to participants what they were expected to include in their writing samples. These expectations stated that the participants needed to (1) write a letter to the university president, (2) state whether or not they agreed or disagreed with the printing policy change, (3) explain both sides of the argument, and (4) provide alternatives that would both help the university reduce printing costs and be more accepted by students who disagree with the printing policy. Thirdly, a 20-minute time limit was put into place, as this is a common time limit on existing writing assessments that require participants to compose a writing sample (e.g., *PIAT-R/NU*; Markwardt, 1998). Finally, a standard set of directions was added to the prompt. These directions alerted participants of the time limit, informed them that they could use scrap paper to plan and organize their ideas before writing, reminded them that they would be required to handwrite their response, let them know that they would not be provided with any assistance, and offered them the opportunity to ask questions before starting the task. Overall, these changes were made to ensure that all of the writing samples were collected in the same manner (see Appendix C for writing prompt).

**Analytic Rating Scale Development**

To develop an analytic rating scale that would be used to analyze the writing samples collected for the current study, several existing rating scales were examined to determine the features of persuasive writing that are most commonly measured (e.g., Westby & Clauser, 2005). Additionally, features measured by existing writing assessments (e.g., *Test of Written Language – 4th Edition*; Hammill & Larsen, 2009) were considered for inclusion. Finally, the prompt was reviewed to determine which features of writing participants would be expected to produce.
Initially, the features of writing measured by the rating scale included macrostructural elements of *The Claim, Reasons for the Claim, Reasons Against the Claim, Solutions/Suggestions,* and *Style/Sense of Audience.* Microstructural elements included *Length/Content, Word Choice/Written Language, Sentence Complexity, Syntax, Organization, Transitions/Cohesion,* and *Spelling/Punctuation/Capitalization.* Each feature was rated on a four-point scale (i.e., “1” = poor and “4” = excellent) that included detailed descriptions of each point value.

To test the scale and resolve wording issues, 12 writing samples were scored independently by three raters (two undergraduate research assistants and the primary investigator). The raters came together after scoring every three samples to compare scores, resolve discrepancies, and calculate inter-rater reliability. Inter-rater reliability was calculated by comparing the scores of two raters using all possible combinations of raters. Overall, absolute inter-reliability (i.e., scores were exactly the same) for all 12 samples ranged from 25% to 92%, while reasonable inter-rater reliability (i.e., scores within one point of each other) for all 12 samples ranged from 83% to 100%. Absolute and reasonable inter-rater reliability between each pair of raters across all categories on the rating scale can be seen in Table 4.

Due to problems achieving acceptable levels of inter-rater reliability, the analytic rating scale was adapted to a three-point scale. Additionally, several of the categories on the original scale were condensed and/or separated. On the second version of the rating scale, macrostructural elements included *Presence of Prompt Components, Logical Reasoning, Sense of Audience,* and *Organization,* while microstructural elements included *Word Choice, Sentence Complexity, Sentence Soundness, Spelling,* and *Punctuation.* In this version, *Sentence Complexity* and *Spelling* were not given point values on the scale, but were instead calculated as percentages (i.e., percentage of complex sentences and percentage of total words misspelled).
Table 4

Inter-Rater Reliability for 12 Samples Scored Using Original Four-Point Rating Scale

<table>
<thead>
<tr>
<th>Category</th>
<th>Rater 1 vs. Rater 2</th>
<th>Rater 2 vs. Rater 3</th>
<th>Rater 1 vs. Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute</td>
<td>Reasonable</td>
<td>Absolute</td>
</tr>
<tr>
<td>Claim</td>
<td>83%</td>
<td>100%</td>
<td>58%</td>
</tr>
<tr>
<td>Reasons for Claim</td>
<td>58%</td>
<td>100%</td>
<td>67%</td>
</tr>
<tr>
<td>Reasons Against Claim</td>
<td>50%</td>
<td>100%</td>
<td>42%</td>
</tr>
<tr>
<td>Solutions/ Suggestions</td>
<td>42%</td>
<td>100%</td>
<td>58%</td>
</tr>
<tr>
<td>Style/Sense of Audience</td>
<td>83%</td>
<td>100%</td>
<td>33%</td>
</tr>
<tr>
<td>Length/ Content</td>
<td>75%</td>
<td>100%</td>
<td>58%</td>
</tr>
<tr>
<td>Word Choice/ Written Language</td>
<td>50%</td>
<td>100%</td>
<td>25%</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>83%</td>
<td>100%</td>
<td>83%</td>
</tr>
<tr>
<td>Syntax</td>
<td>67%</td>
<td>100%</td>
<td>75%</td>
</tr>
<tr>
<td>Spelling/ Punctuation/ Capitalization</td>
<td>67%</td>
<td>100%</td>
<td>75%</td>
</tr>
<tr>
<td>Organization</td>
<td>75%</td>
<td>100%</td>
<td>58%</td>
</tr>
<tr>
<td>Transitions/ Cohesion</td>
<td>92%</td>
<td>100%</td>
<td>58%</td>
</tr>
</tbody>
</table>

All 58 writing samples (including four samples excluded for the current study) were scored using the three-point rating scale to calculate reliability. Presence of Prompt Components and Logical Reasoning were scored by two graduate research assistants, Sense of Audience and Organization were scored by two different graduate research assistants, Word Choice and
Spelling were scored by two undergraduate research assistants, and Sentence Complexity, Sentence Soundness, and Punctuation were scored by a clinical fellow (i.e., also a doctoral student and the primary investigator of this study) and an expert clinician (i.e., also a professor in communication disorders and the primary investigator’s advisor). Inter-rater reliability was calculated for absolute agreement and reasonable agreement. Overall, absolute agreement ranged from 56.9% to 93.2% and reasonable agreement ranged from 93.1% to 100.0% across all categories (Table 5).

Table 5

<table>
<thead>
<tr>
<th>Rating Scale Category</th>
<th>Absolute Agreement</th>
<th>Reasonable Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence of Prompt Components</td>
<td>60.3%</td>
<td>98.3%</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>56.9%</td>
<td>98.3%</td>
</tr>
<tr>
<td>Sense of Audience</td>
<td>82.8%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Organization</td>
<td>86.2%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Word Choice</td>
<td>56.9%</td>
<td>93.1%</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>93.2%</td>
<td></td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td>80.0%</td>
<td>99.9%</td>
</tr>
<tr>
<td>Spelling</td>
<td>59.3%</td>
<td></td>
</tr>
<tr>
<td>Punctuation</td>
<td>72.4%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Note. Reasonable agreement was unable to be calculated for Sentence Complexity and Spelling because these categories were coded based on sentence type or type of spelling error. Therefore, it was not possible for scores to be within one point of each other. However, absolute agreement was calculated based on the type of sentence and type of spelling error that was coded.

When the same or similar categories were compared from the original four-point scale to the adapted three-point scale, reliability increased in all areas except Spelling. Absolute agreement for Spelling was low on the new rating scale, likely because raters were not given specific instructions for the types of errors that should be coded as spelling errors. Instead, they were simply told to code all spelling errors. When they began coding, the raters quickly realized
that they had many questions about which types of errors should be coded as spelling errors. Therefore, during the process of coding for spelling errors, they kept track of the nature of the errors that they coded and the questions that they had about what should be coded. After all coding was completed, they came together to discuss the patterns that they had found and the questions that they had in relation to the types of errors that should be coded as spelling errors. The types of errors that they had coded or had questions about included general spelling errors (e.g., *nessessary* for *necessary*), apostrophe errors (e.g., *dont* for *don’t*), morphosyntactic errors (e.g., *student* for *students* or *look* for *looked*), spacing issues (e.g., *some times* for *sometimes*), and word usage errors (e.g., *then* for *than*). These issues were helpful in developing the analytic rating scale used for the current study.

Although the use of the three-point rating scale improved reliability, researchers have suggested that scales with an even number of point values are preferable because they prevent raters from defaulting to the midpoint of the scale (Pagano et al., 2008). Additionally, it has been reported that the highest levels of reliability are reached when five- to nine-point scales are used (Knoch, 2011). However, Knoch also reported that a smaller number of levels can be used when there is limited variability among the overall quality writing samples (or the overall writing abilities of the participants). Because all of the writing samples analyzed in this study were collected from undergraduate students at the same university within a fairly narrow age range (i.e., all participants were between the ages of 17 and 27 years, with the exception of one 37-year-old participant), it was expected that the variability among samples would be limited compared to instances where samples come from multiple sites and/or a wide range of ages or educational levels. Therefore, for the purposes of this study, the analytic rating scale was adjusted back to a four-point scale.
The categories on the four-point rating scale for this study included the macrostructural elements of Presence of Prompt Components, Logical Reasoning, Sense of Audience, and Organization, as well as the microstructural elements of Word Choice, Sentence Complexity, Sentence Soundness, Spelling, and Punctuation (Appendix D). The categories of Sense of Audience, Sentence Complexity, Sentence Soundness, Spelling, and Punctuation were initially calculated as percentages (e.g., percentage of total sentences that were inappropriate for the audience of a university president) and were later converted to point values on the four-point rating scale based on criteria developed by the primary investigator and her advisor.

**Instrumentation**

Although the data collection process differed for typical participants and participants with WLDs, all participants completed the same measures for this study. These measures included the *Peabody Picture Vocabulary Test – 4th Edition (PPVT-4)* and a persuasive essay. Additionally, both groups completed demographic questionnaires that differed slightly due to the nature of their use (i.e., diagnostic versus research).

**Demographic questionnaire.** Two separate questionnaires were used to gather demographic information from the participants in this study. The questionnaires differed due to the different purposes for their use. For participants with WLDs, the demographic questionnaires served as both clinical and research tools. Therefore, these questionnaires were longer than those completed by typical participants, as they requested additional information pertaining to the participants’ present concern(s) and history of language and literacy problems. Questionnaires completed by typical participants served only a research purpose, so substantially less information was requested on these questionnaires than those given to participants with WLDs.
The demographic questionnaire for participants with WLDs was developed by Dr. Lauren Katz for use as a case history form for college students with reading and/or writing problems in clinical settings (Appendix E). Participants were asked to provide information about their gender, age, year in school, major, Honors program enrollment, university probation program membership, race/ethnicity, history of disorders/disabilities, special education history, college accommodations history, academic assistance history, reading/writing/organization skills, attitudes toward reading/writing/organization, beliefs regarding ability to benefit from therapy, willingness to participate in therapy, and amount of money willing to be spent on therapy.

For typical participants, the original demographic questionnaire was condensed to include only questions that pertained to the current study (Appendix F). The information requested from typical participants included gender, age, year in school, major, current GPA, ACT/SAT scores, enrollment in Honors program, membership in university probation program, race/ethnicity, history of disorders/disabilities, college accommodations history, special education history, academic assistance history, attitudes towards reading and writing, history of college writing courses, and reason(s) for attending college.

**Peabody Picture Vocabulary Test – 4th Edition (PPVT-4).** The PPVT-4 (Dunn & Dunn, 2007) is an untimed, individually administered, norm-referenced test that measures the understanding of single-words in individuals aged 2.5 to 90+ years. To administer this measure, the examiner says a word that describes one of four pictures on a page, and the examinee is expected to either point to or say the number of the picture that best represents that word. The PPVT-4 (Form A) was used in this study to provide a measure of general functioning, as scores on the PPVT may be used as a proxy for verbal intelligence (Anderson & Flax, 1968; Pool & Brown, 1970).
**Persuasive essay.** The persuasive writing task required participants to write a letter to the university president about whether they were for or against a change in the university’s printing policy, which would now require students to pay for printing. Participants were instructed to include a statement of their claim, reasons for their claim, reasons against their claim, and a suggestion for how the university could adjust the policy to make both sides happy.

To collect the persuasive writing samples, participants were each given a copy of the prompt, which was read aloud to them as they followed along. Additionally, each participant was provided with one blank sheet of scratch paper, two double-sided lined sheets of paper, and one pen. All participants were informed that they would be given 20 minutes to complete their persuasive essays and were given a warning when they had only five minutes remaining. At 20 minutes, if a participant had not completed his/her sample, he/she was asked to draw a line across the paper and continue writing below this line until he/she was finished. If a participant completed his/her writing sample before the 20-minute time limit expired, the timer was stopped and the number of minutes he/she had spent writing was recorded at the top of his/her sample.

Overall, a total of four typical participants (8% of typical) and one participant with a WLD (25% of WLDs) spent more than 20 minutes composing their writing samples. For the typical participants going over the 20-minute time limit, two spent an additional one minute writing, one spent an additional two minutes, and one spent an additional five minutes writing. However, the participant with a WLD spent 20 minutes composing a rough draft and an additional 19 minutes copying this draft word-for-word to another sheet of paper to produce a final draft. A visual inspection of these two samples revealed that they were identical (i.e., this WLD participant had really only spent 20 minutes producing a sample). In the end, all of these writing samples were scored and coded in their entirety, as it would have been difficult to score
for all categories on the analytic rating scale without the full samples (e.g., *Presence of Prompt Components*) and none of the participants had drastically exceeded the time limit.

**Data Collection**

**Data collection for typical participants.** Individual data collection sessions were scheduled with each of the typical participants via email. To begin each of these sessions, consent forms (Appendix G) were reviewed and signed. Participants then completed the demographic questionnaire. Next, the *PPVT-4* (Form A) was administered. Finally, participants composed a persuasive writing sample using the provided prompt. Each data collection session typically lasted about 45 minutes.

**Data collection for participants with WLDs.** Individual diagnostic sessions were held with each of the participants with WLDs. To begin each of these sessions, participants signed both a video consent form (for diagnostic purposes) and a study consent form (Appendix G). Next, interviews were conducted with each of the participants with WLDs to obtain additional information that would be relevant to the diagnostic process (e.g., present concern, goal of evaluation, etc.). Procedures following the interview varied for each participant depending upon his/her specific concerns. Both formal and informal measures of spoken and written language were administered. In all cases, *PPVT-4* (Form A) scores and persuasive writing samples were obtained. Diagnostic session length varied, ranging from three to five and a half hours over one or two sessions during a single day.

**Analysis of Writing Samples**

All hand-written writing samples were typed verbatim by three undergraduate research assistants and the primary investigator. Any identifying information present in the writing samples was typed as XXX. After each persuasive writing sample was typed verbatim by one of
these researchers, it was reviewed by another to ensure accuracy. The primary investigator then segmented all typed samples into sentences based upon the punctuation provided by the participants. The unsegmented samples were used to obtain holistic scores and scores for several categories on the analytic rating scale (i.e., Presence of Prompt Components, Logical Reasoning, Organization, Word Choice, and Spelling), while the segmented samples were used for Sense of Audience, Sentence Complexity, Sentence Soundness, and Punctuation on the analytic rating scale.

**Rater recruitment.** To recruit raters to score the writing samples using both a holistic rating scale and the four-point analytic rating scale, faculty members in the Department of Communication Sciences and Disorders were contacted to determine if they had any graduate research assistants (RAs) that they would be willing to share with the primary investigator of this study for a few hours per week. A total of five graduate RAs were available. However, it was decided that only three would be used, as this would ensure that there were enough raters for the scoring process to be completed in a timely manner, but also that there were few enough raters for the scoring process to be completed cleanly.

**Rater training.** To prepare the raters to complete the holistic scoring task, a brief meeting was held to teach them about the process. During this meeting, the raters were given instructions about how to complete the holistic scoring task (Appendix H), a copy of the persuasive writing prompt, data sheets for each writing sample, and two highlighters (i.e., one yellow and one pink). Additionally, each rater was given a total of 27 unsegmented writing samples that were randomly-selected from the 57 samples included in this study. Twelve of these samples were the same for all raters (for inter-rater reliability), while 15 were unique for each rater (see Figure 6). Raters were blind to the type of participant that composed each writing
sample. After materials were dispersed, the primary investigator discussed the procedures outlined in the instructions, answered any questions that the raters had, provided a deadline for the completion of the holistic scoring task, and reminded the raters not to discuss the scoring process with each other until all scoring-related tasks had been completed. Limited information was given to the raters about the study at this time to ensure that they were able to complete the holistic scoring task without focusing on the specific aspects of writing that would be examined using the analytic rating scale. A week after the raters had scored all of the samples holistically, they were asked to re-score six of their samples so that intra-rater reliability could be calculated (see Figure 7).

![Pie Chart](image)

**Figure 6.** Distribution of writing samples among raters for holistic and analytic scoring procedures.

A total of four meetings were held over the course of four weeks after the holistic scoring process had been completed to teach the raters how to score the writing samples using the
analytic rating scale (see Figure 8). During the first meeting, instructions for completing the scoring procedure (Appendix I), the analytic rating scale (Appendix D), and guidelines for using the rating scale (Appendix J) were reviewed. After these materials were reviewed, the primary investigator showed the raters five writing samples that had already been scored using the rating scale and discussed how scores and/or codes for *Presence of Prompt Components, Logical Reasoning,* and *Sense of Audience* were decided upon for each of these samples. The raters were then given six unscored samples and were asked to work together to score these samples for *Presence of Prompt Components, Logical Reasoning,* and *Sense of Audience.* During this time, the primary investigator answered questions and provided feedback about their scoring and coding decisions. At the end of this meeting, the raters were assigned to score and/or code their 27 samples for these rating scale categories before the next meeting.

![Figure 7](image)

**Figure 7.** Representation of writing samples scored by each rater for intra-rater reliability during holistic and analytic scoring procedures
Figure 8. Flowchart of holistic and analytic scoring procedures.
At the next meeting, the guidelines for scoring and/or coding for Organization, Word Choice, Spelling, and Punctuation were reviewed. The raters were then given the opportunity to score/code six samples for these features on their own. The primary investigator answered questions and provided feedback as appropriate during this task. To end the meeting, the raters were assigned to score/code their 27 samples for these rating scale categories before the next meeting.

During the third meeting, the raters were trained how to identify complex sentences (see Appendix K for guidelines given to raters). They then practiced identifying complex sentences in six samples together with feedback given from the primary investigator as needed. At the end of this meeting, the raters were assigned to re-score/re-code six samples for Presence of Prompt Components, Logical Reasoning, and Sense of Audience (for intra-rater reliability), as well as code all 27 of their samples for Sentence Complexity before the next meeting.

The fourth meeting was used to review the guidelines for coding for Sentence Soundness and allow the raters to practice coding six samples for Sentence Soundness on their own. The primary investigator answered questions and provided feedback as appropriate as the raters practiced coding. At the end of the meeting, the raters were assigned to re-score/re-code six samples for Organization, Word Choice, Spelling, and Punctuation, as well as code all 27 of their samples for Sentence Soundness within the next week.

After training was completed, two weeks passed before another meeting was held, as one of the raters was not available to meet during the first week and the university’s spring break occurred during the second week. However, during the first week between meetings, the primary investigator provided the raters with tasks to complete before the following meeting. These tasks included re-coding six of their samples for Sentence Complexity and Sentence
Soundness, as well as re-scoring all 27 of their samples using the holistic scoring procedure discussed previously. When the next meeting was held, the raters were assigned to re-score six of their sample holistically so that intra-rater reliability could be calculated. Additionally, raters began to resolve discrepancies in scores (procedure described in Resolving Discrepancies in Scoring section below).

**Holistic scoring procedure.** During the holistic scoring procedure, a total of 57 writing samples were scored using a four-point holistic rating scale developed for this study. The holistic rating scale required raters to provide a quick rating of the overall quality of the samples without focusing on specific features of writing, as one of the aims of the current study was to determine which features of persuasive writing (i.e., scores on analytic rating scale) relate to overall writing quality (i.e., scores on holistic rating scale). As mentioned previously, all three of the raters were given the same twelve randomly-selected writing samples in the same, random order so that inter-rater reliability could be calculated. Additionally, each rater was given fifteen different randomly-selected writing samples from the remaining 45 samples that were randomly ordered. Raters were blind to the type of participant that composed each sample.

To complete the holistic scoring task, the raters were first required to read the persuasive writing prompt that was given to the participants in this study. This helped the raters develop an understanding of what the participants were expected to include in their writing samples, which would help in determining the overall quality of each writing sample. Next, the raters were asked to read through twelve randomly-selected writing samples without scoring them so that they could develop a sense of the range of quality seen in the writing samples. All three raters read the same twelve samples (i.e., those used for inter-rater reliability) in the same order.
Once they had read through the writing prompt and twelve samples without scoring them, they were instructed to re-read each of the twelve samples individually and focus on the overall quality of each sample. They were reminded that the goal of this process was to get their gut reaction about the overall quality of each sample, so they should neither take too much time examining each sample nor rush through each sample. Instead, they were told that they should be sure to read each sample quickly and carefully enough to be able to provide an accurate score. While reading, they were asked to highlight any sections of the text that they felt were especially strong (yellow) or weak (pink). After they finished reading each sample, they were instructed to record their initials, the sample number, a holistic score (i.e., 4 = excellent, 3 = good, 2 = fair, and 1 = poor), and comments on a provided data sheet. The comments were supposed to help explain why the rater had selected a particular score for each writing sample. After all of the raters scored the same twelve initial samples, they were each asked to complete the same process with their unique set of fifteen samples. The raters were asked to try to complete the entire scoring process in one sitting and to place the samples and data sheets in the primary investigator’s mailbox once all of the samples had been scored.

One week after the raters had completed the holistic scoring task, they were asked to re-score six (~20%) of the samples that they had previously scored using the same procedures. At this time, each rater was given six samples that had been randomly-selected from the 27 samples that she had previously scored and six data sheets. This task was completed so that intra-rater reliability could be calculated.

After the raters had scored and coded all of the samples using the analytic rating scale (described below), they were asked to complete the holistic scoring procedure again. This process was completed a second time to see if the intense training that the raters had experienced
when learning how to use the analytic rating scale had changed how they scored the writing samples holistically and/or changed their inter-rater and/or intra-rater reliability for the holistic rating scale. The process was completed in the same manner as the first round of holistic scoring except that two weeks passed between the initial scoring of the samples and the re-scoring of the samples for intra-rater reliability.

Analytic scoring procedure. Once the initial holistic scoring process was complete, each of the same three raters used the analytic rating scale to re-score the same 27 samples that they had originally been assigned. Again, the raters were blind to the type of participant that composed each writing sample. To complete this task, the raters were provided with an instruction sheet (Appendix I), a copy of the analytic rating scale (Appendix D), guidelines for how to use the rating scale (Appendix J), guidelines for how to identify complex sentences (Appendix K), five examples of scored samples, six unscored samples for practice, two copies of each of their 27 writing samples (one segmented and one unsegmented), and 27 data sheets.

Before they started to score the samples using the analytic rating scale, the raters were first asked to re-read the persuasive writing prompt to remind them of what participants were supposed to include in their writing samples. After they re-read the prompt, they were asked to re-read the guidelines for scoring or coding the category that they would be working on and then score/code all 27 of their samples one component at a time. In other words, they would start by scoring all of their samples for Presence of Prompt Components and then continue through the categories on the rating scale. While practicing SLPs would likely score one sample at a time for all categories on a rating scale, they likely would not be scoring 27 samples in a single sitting. Therefore, for the purposes of the current study, raters were asked to score all samples one
category at a time to ensure that they were consistent in how they scored/coded each category and that the training for each category was fresh in their minds as they scored.

To provide a score for each category, they were instructed to read through each sample at least once. However, they were told that they should read through the samples as many times as they needed to provide an accurate score (or accurate codes). In an attempt to maintain consistency in scoring across samples, the raters were required to complete scoring/coding for each category within a short time period (ideally in one sitting). While scoring/coding, they were required to write specific codes on the writing samples for certain categories (i.e., Sense of Audience, Spelling, Punctuation, Sentence Complexity, and Sentence Soundness), but were also encouraged to write on the samples while scoring for other categories if they felt that it would be helpful.

After they finished scoring/coding each sample for a particular category, they were required to record their score or frequency count of codes on provided data sheets. Additionally, they were asked to track the dates that they completed scoring/coding for each category so that they would know when they needed to re-score six of their samples for intra-rater reliability. All re-scoring for intra-rater reliability for the analytic rating scale took place two weeks after the initial scoring date. As they were scoring, the raters were told to record anything they found to be frustrating or confusing while they were scoring the samples so that their concerns could be addressed in future revisions of the analytic rating scale. Finally, the raters were reminded not to talk to each other about the scoring/coding procedures until all of their work with the writing samples was complete.

To score each sample for Presence of Prompt Components, raters were required to read the unsegmented samples and determine which of the following were included in each sample:
(1) claim (i.e., for or against the printing policy), (2) reason(s) for the claim, (3) reason(s) against the claim, and (4) solution(s) or suggestion(s) that would make both sides happy. Claims that were not explicitly stated and left the reader wondering where the writer stood on the issue were not counted. The number of these components included in each sample was used to derive a score for this category.

Scoring for Logical Reasoning required raters to read the unsegmented samples and closely examine the arguments and reasons provided in the samples for each of the prompt components. They then had to decide whether or not the arguments presented by the writer were clear and persuasive. To judge the clarity and persuasiveness of arguments, the raters were given several questions to ask themselves while reading (see guidelines in Appendix J). The level of clarity and persuasiveness of the arguments and the amount of work that would be needed to improve the arguments were used to derive scores for this category.

For Sense of Audience, raters had to read the segmented samples and determine the number of sentences that were inappropriate for the audience of a university president. Sentences were coded as “inappropriate” if they contained the use of slang, were too conversational in nature, and/or were rude. Raters reported the total number of sentences that were coded as inappropriate for this category and the primary investigator later calculated the percentage of sentences in each sample that were deemed inappropriate by the raters. These percentages were then used to derive scores on the rating scale using criteria developed by the primary investigator and her advisor.

Organization required raters to determine how well organized each sample was and how easy it was to follow. Examples of features to look for included presence or absence of introductory and concluding elements, flow within and between sentences, flow within and
between paragraphs, overall flow, appropriateness of breaks in paragraphs, appropriateness of content location, appropriate use of paragraph structure, and use of transitional words or phrases. Raters were told that samples that were either perfectly organized or horribly organized were likely to be rare, so scores of “4” and “1” were unlikely to be given often. Scores were determined based on the level of organization, amount of work needed to improve organization, and how well the raters were able to follow the sample.

To score for **Word Choice**, raters had to read the unsegmented samples and determine how frequently sophisticated words were used correctly in each sample. Raters were informed that sophisticated words could include low frequency words, words that were more advanced than typical college vocabulary words, or words that were morphologically complex, multisyllabic, and/or long. Additionally, raters were told that some morphologically complex words might not be sophisticated, while words that were not morphologically complex might be sophisticated. Furthermore, raters were instructed that words that were only morphologically complex in nature due to the addition of a morpheme for possession (‘s or s’), plurality (plural -s), or verb tense (-ing or -ed) should not be considered sophisticated in nature. Raters were also reminded that sophisticated words that were repeated multiple times throughout a sample became less sophisticated as a result of this repetition. Finally, raters were asked not to consider words mentioned in the writing prompt as sophisticated (i.e., *mindful*, *unnecessary*, *alternative(s)*, *disagree*, and *acknowledge*). Scores for this category were determined based on the frequency of occurrence of sophisticated words in a sample, as well as whether or not any of the sophisticated words in a sample were used incorrectly. The results from a previous study that focused on the best methods of measuring vocabulary sophistication in the persuasive writing samples of undergraduate students and was conducted by the primary investigator, her advisor, and several
undergraduate research assistants were used to develop the criteria for this category (Richards, Gora, Finsel, Langenkamp, & Katz, 2011).

For Sentence Complexity, raters were required to read through the segmented samples and code all sentences that were complex. Raters were told not to code sentences as complex based solely upon the use of an infinitive (e.g., I want to go to the store.). To identify complex sentences, raters were encouraged to first look for the presence of subordinating conjunctions. Raters were given a list of subordinating conjunctions to assist them in this process and were reminded to look for instances of implied “that”. If they were unable to determine whether or not a sentence was complex by looking for a subordinating conjunction, they were then told to locate all of the verbs or verb units in a sentence. Once they located the verbs or verb units, they were told to (1) determine whether or not they occurred in separate clauses, and (2) determine the nature of the clauses that contained the separate verbs/verb units. If at least one of the clauses was dependent, they were instructed to code the sentence as complex (unless it was a run-on sentence or sentence fragment). After all of the sentences were coded, raters totaled up the number of complex sentences they had coded for each sample and recorded this number on a provided data sheet. The primary investigator later calculated the percentage of sentences that were complex in each sample. These percentages were then used to derive scores on the rating scale using criteria developed by the primary investigator and her advisor.

Segmented samples were also used to score for Sentence Soundness. For this category, raters were required to score each sentence on a three-point scale based upon how sound it was when considering clarity, syntax, word choice, and awkwardness. Some examples of features that raters might penalize a sentence for include sentence fragments, run-on sentences, multiple redundancies, semantic errors, morphological errors, syntactic errors, flawed logic, weirdness or
awkwardness, or issues of sentence clarity. Raters were instructed to consider the previous and subsequent sentences when judging for Sentence Soundness and to think about how each sentence fit with the surrounding sentences. Additionally, any errors that would be captured by another category on the rating scale were ignored (e.g., spelling or punctuation). A sentence was given a score of “3” when it was completely sound or acceptable as written and its point was entirely clear. A score of “2” was given when a sentence contained some errors or instances of awkwardness that were easily fixable and the reader was still able to grasp the point of the sentence. Sentences that contained frequent errors or instances of awkwardness, needed a lot of work to be fixed, and were very confusing and difficult to understand without multiple readings were given a score of “1”. The three-point scale was used for this category because it was found that adequate inter-rater reliability could not be met using a four-point scale even when expert clinicians were coding samples. After all sentences were coded, the primary investigator calculated the percentage of total sentences that had been given a score of “3” in each sample. These percentages were used to determine which point value to award each sample on the analytic rating scale using criteria that had been developed by the primary investigator and her advisor.

For Spelling, raters read through each unsegmented sample and coded words that were misspelled, homophones that were used incorrectly (e.g., “then” for “than”), or contractions that demonstrated problems with apostrophe usage (e.g., “dont” or “do’nt”). They were told not to code words that were misspelled solely due to errors of capitalization, spacing, and/or hyphen usage, as it was possible that these were not errors but were instead a typist’s interpretation of a participant’s handwriting. For example, while transcribing samples, the poor handwriting of certain participants often made it difficult to determine whether a letter was capitalized or written
in lower case, or whether or not spaces were present within or between some words. Furthermore, participants sometimes used hyphens to split up words that ran on two lines in their handwritten sample, which ended up looking like errors in their typed samples because these words no longer fell on two separate lines. Additionally, raters were instructed to only code morphological errors (e.g., omitted plural –s) as spelling errors if they were not frequent throughout the sample, as this would suggest that the errors were more related to spelling problems than morphosyntactic issues. Once all spelling errors were coded, raters totaled up the number of misspelled words they had coded and recorded this number on a data sheet. The primary investigator later divided this number by the total number of words to determine the overall percentage of words that were misspelled in each sample. This percentage was then compared to criteria on the rating scale developed by the primary investigator and her advisor to determine the appropriate Spelling score for the sample.

*Punctuation* was coded for using the segmented samples. Raters were instructed to read through each sample and code sentences for errors of ending punctuation or other punctuation. Raters coded a sentence as having an error of ending punctuation if incorrect ending punctuation was used or it was a run-on sentence that required more than the addition of a semicolon to be fixed. A sentence was coded for other punctuation errors if there were any of the following: omitted or misused apostrophes in instances of possession, omitted or misused commas, omitted or misused semicolons (including run-on sentences), omitted or misused quotation marks, and/or omitted or misused periods in abbreviations. Raters reported the total number of sentences in each sample that were coded for each type of error on their data sheets. The primary investigator later calculated the percentage of sentences with each type of error in each sample. These
percentages were later compared to criteria on the rating scale developed by the primary investigator and her advisor to determine the appropriate *Punctuation* score for each sample.

**Resolving discrepancies in scoring.** After all of the samples had been scored using the holistic and analytic rating scales, three meetings were held for raters to resolve discrepancies in their scores/codes and come to agreed-upon scores/codes for each sample. The categories of *Presence of Prompt Components* and *Logical Reasoning* were resolved during the first meeting. *Organization, Word Choice, Sense of Audience, Ending Punctuation,* and *Other Punctuation* discrepancies were resolved during the second meeting. The final meeting focused on resolving discrepancies for *Sentence Complexity, Sentence Soundness, Spelling,* and holistic scores. These agreed-upon scores were used during the quantitative and qualitative analyses that addressed the second and third research questions.

**Research Design**

The current study utilized a mixed methods approach in order to best examine how reliably the persuasive writing skills of undergraduate students can be measured using rating scales, the similarities and differences between the persuasive writing samples of typical undergraduate students and undergraduate students with WLDs, and the features of persuasive writing that relate to overall writing quality. Tashakkori and Creswell (2007) define mixed methods research “…as research in which the investigator collects and analyzes data, integrates the findings, and draws inferences using both qualitative and quantitative approaches or methods in a single study (p. 4).” The primary goal of mixed methods research is to draw on the strengths of qualitative and quantitative methods of research in a single study to minimize their weaknesses (Johnson & Onwuegbuzie, 2004). A mixed methods approach was selected for this
study to provide a fuller picture and deeper understanding of the data being studied than either method could provide on its own.

Since the primary reason for using a mixed methods approach was to obtain a more complete picture of the data, a complementarity design was selected for this study. Greene, Caracelli, and Graham (1989) stated that the reason for combining qualitative and quantitative methods in a complementarity design “is to use the results from one method to elaborate, enhance, or illustrate the results from the other (p. 266-267).” This design is used to examine different aspects of the same phenomenon. Therefore, in the current study, the phenomenon was persuasive writing in undergraduate students and the aspects of this phenomenon that were examined included the overall quality of the writing samples (i.e., scores on holistic rating scale), specific features of the writing samples (i.e., scores on analytic rating scale), and comments about the various strengths and weaknesses of the writing samples (i.e., comments given about why a particular holistic score was given to each sample).

In this study, features of undergraduate persuasive writing were analyzed using both quantitative and qualitative methods. Using quantitative methods, the overall quality of the writing samples was scored using a holistic writing scale, while specific features of persuasive writing (e.g., presence of prompt components and sense of audience) were scored using an analytic rating scale. Comparisons of scores on these rating scales were made using various descriptive and inferential statistical analyses. Additionally, qualitative methods were used to code and group comments provided by raters during the holistic scoring process into various themes and subthemes that described reasons why raters had awarded each sample a particular holistic score. These themes and subthemes were used to examine the consistency in scoring procedures within and between raters, as well as compare the types of comments given to
specific groups of participants (i.e., typical participants versus participants with WLDs, and participants with high holistic scores versus participants with low holistic scores).

To ensure that a complementarity mixed methods design was followed, the qualitative and quantitative methods examined different aspects of the same phenomena and were implemented in a simultaneous and interactive fashion. For example, both qualitative and quantitative methods were used to examine the features of persuasive writing that seem to distinguish between typical undergraduates and undergraduates with WLDs. The quantitative analyses primarily focused on the scores awarded to the writing samples, while the qualitative analysis examined the comments provided by raters about why they had given a particular holistic score to each sample. Both the quantitative and qualitative analyses examined the same phenomena (i.e., features of persuasive writing), but focused on different aspects of the phenomena (i.e., holistic and analytic scores versus comments about strong and weak features of samples). Additionally, these analyses were conducted during the same time period and were highly related to each other.

**Analysis of the Data**

**Quantitative Methods**

The quantitative portion of this study consisted of three raters scoring the 54 persuasive writing samples using both the four-point holistic and analytic rating scales designed for this study. As discussed previously, all writing samples were scored independently by three separate raters, who then came together to resolve any discrepancies in the 12 samples that they all scored. Scores on the rating scales were used to quantitatively answer all three research questions.
To answer the first research question regarding to what extent rating scales can be used to reliably measure the persuasive writing abilities of undergraduate students, inter-rater and intra-rater reliability were calculated for both the holistic and analytic rating scales. Measures of reliability for the holistic rating scale were calculated twice – both before and after raters completed the in-depth training on how to score various features of persuasive writing using an analytic rating scale.

To determine the features of persuasive writing that seemed to best distinguish between typical undergraduates and undergraduates with WLDs, means and standards deviations were calculated for the performance of each group of participants on the rating scales. Additionally, statistical analyses were run to compare the performance of typical participants and participants with WLDs on the holistic and analytic rating scales. Fisher’s exact test was used to compare the proportion of participants in each group scoring high (i.e., three or four) and low (i.e., one or two) on the rating scales, while t-tests were used to compare total scores on the analytic rating scale and percentages calculated for specific categories on the rating scale (i.e., Sense of Audience, Sentence Complexity, Sentence Soundness, Spelling, and Punctuation). Effect sizes for the t-tests were calculated using the following formula for Cohen’s d:

\[ \text{Cohen’s d} = \frac{(M_1 - M_2)/ \sqrt{([SD_1^2 \times N_1 - 1] + [SD_2^2 \times N_2 - 1]) / N_1 + N_2 - 2}} \]

This formula was selected because it is suggested for times when the sample sizes of the two groups being compared are not relatively equal (Dunst, Hamby, & Trivette, 2004). In this particular instance, the sample sizes were not similar (i.e., 50 typical participants versus four participants with WLDs).
For the final research question, which sought to determine which features of persuasive writing relate to overall writing quality, statistical tests were used to compare the performance of participants receiving high holistic scores (i.e., scores of three or four) versus low holistic scores (i.e., scores of one or two) in each area of the analytic rating scale and total analytic rating scale scores. More specifically, Fisher’s exact test was used to compare the proportion of participants in each group (i.e., high vs. low holistic scores) receiving high (i.e., scores of three or four) and low scores (i.e., scores of one or two) in each category on the analytic rating scale. Additionally, t-tests were used to compare the groups’ total analytic rating scale scores and percentages within specific categories of the rating scale (i.e., Sense of Audience, Sentence Complexity, Sentence Soundness, Spelling, and Punctuation). Again, the Cohen’s d formula provided above was used to calculate effect sizes for the t-tests.

**Qualitative Methods**

The qualitative portion of the current study involved having raters score the 54 persuasive writing samples using the holistic and analytic rating scales, as well as having them provide comments about why they selected a particular holistic score for each sample. Figures were created that showed the proportion of participants in each group receiving specific scores on the holistic rating scale and within each category of the analytic rating scale. Additionally, comments provided by the raters during the holistic scoring procedure were typed, coded, and grouped into themes by the primary investigator. Finally, the types of spelling and punctuation errors produced by participants were coded and compared by the primary investigator.

The qualitative analyses of the data followed steps outlined by Creswell (2007), which state that qualitative analysis involves “…preparing and organizing the data…for analysis, then reducing the data into themes through a process of coding and condensing the codes, and finally
representing the data in figures, tables, or a discussion (p.148).” Several steps were taken during each phase of the qualitative analysis process to ensure that this process was followed.

**Data preparation and organization.** To prepare the data for qualitative analysis, all comments provided by raters about reasons for their holistic scores were typed verbatim by the primary investigator. These comments were then broken down into their smallest components (e.g., *spelling and punctuation errors* became *spelling errors* and *punctuation errors*) so that only one code was needed to describe each comment.

**Data coding.** Before coding the data, the primary investigator read through all comments provided by raters to get a sense of the data. General thoughts and impressions were recorded in the margins of the typed comments and in field notes that were accrued throughout the data analysis process. Next, the primary investigator began coding the comments. Each comment received two codes: one that showed whether the comment was positive (+) or negative (-) in nature and one that described the writing feature described in the comment. For example, “many spelling errors” would receive a negative code and a code for *Spelling* (a subtheme of *Mechanics*) (see Appendix L for more examples). A key of all of the codes relating to comment content was kept with the field notes.

After the coding process was complete for the comments, the codes were grouped by themes and subthemes. During this process, two comments were unable to be coded because they were too vague (i.e., “easily fixable” and “very few errors”). The themes and subthemes identified through the coding process included:

1. **Linguistic (Form, Content, and Use):** This theme focused on strengths and weaknesses at the sentence level.
a. *Sentence Structure:* This subtheme included problems with the overall structure of a sentence or several sentences. Some of the issues coded in this subtheme included sentence fragments, run-on sentences, choppy sentences, or disconnected sentences.

b. *Wording Issues:* This subtheme focused on problems with the wording within sentences or the vocabulary used within sentences. Issues such as odd or awkward wording, missing words, or incorrectly used words were included in this subtheme.

c. *Clarity Issues:* Comments in this subtheme focused on clarity at the sentence level. Some comments that were commonly included in this subtheme are “unclear sentences,” “confusing sentences,” and “some sentences were difficult to read.”

d. *Grammar/Syntax:* The Grammar/Syntax subtheme included comments that described the presence or absence of grammatical or syntactic errors. The comments in this category primarily mentioned grammatical errors or errors of syntax.

2. *Mechanics:* This theme included comments that mentioned strengths and weaknesses in the areas of spelling and punctuation.

   a. *Spelling:* This subtheme included comments relating to strengths and weaknesses with spelling.

   b. *Punctuation:* This subtheme included comments that discussed the frequency of punctuation errors.
3. **Prompt Components/Content**: This theme focused on comments relating to the four components that were expected given the nature of the writing prompt: (1) claim, (2) reason(s) for claim, (3) reason(s) against claim, and (4) alternative suggestions or solutions that would please both sides of the issue.

   a. **Presence/Absence of All Prompt Components**: This subtheme included comments that stated if all expected prompt components were present. Examples of comments in this subtheme include “addressed all areas of the prompt,” “included three components,” or “missing content.”

   b. **Level of Support for All Prompt Components**: This subtheme addressed how well all of the prompt components were supported. Examples of comments in this subtheme include “good use of examples,” “could use more supporting details,” and “used personal experiences.”

   c. **Quality of All Prompt Components**: This subtheme included issues relating to the clarity, strength, reasonableness, and persuasiveness of all prompt components. Some comments coded in this subtheme included “persuasive,” “thoughtful arguments,” “repeats prompt as often as states own thoughts,” and “their arguments may not be entirely true.”

   d. **Presence/Absence of Claim/Reasons**: This subtheme included comments that stated whether or not the prompt components of claim, reason(s) for claim, and reason(s) against claim were explicitly stated in the samples. Examples of comments in this subtheme include “states their position,” “addresses other viewpoint,” “does not acknowledge opinions of those who disagree,” and “did not say whether or not he/she agreed with the new policy.”
e. **Level of Support for Claim/Reasons:** This subtheme focused on comments that discussed the level of support provided for the claim, reason(s) for the claim, and reason(s) against the claim. Some sample comments include “supports their position,” “pros and cons given for each view,” “could expand more on why they agree,” and “no strong support for their decision.”

f. **Quality of Claim/Reasons:** The overall quality of the claim, reason(s) for the claim, and reason(s) against the claim were the focus of this subtheme. Some of the comments given this code included “strong opinion,” “thoughtful discussion of both viewpoints,” “view of opposing opinion was poor,” and “other viewpoint was not presented in a very clear way.”

g. **Presence/Absence of Alternative Solutions:** This subtheme addressed whether or not alternative solutions were offered in the writing samples. Examples of these comments include “offered alternatives” and “did not give alternatives.”

h. **Level of Support for Alternative Solutions:** The level of support provided for alternative solutions was the focus of this subtheme. Some examples of comments in this subtheme include “not much support for alternative” and “fully explains ways to reduce printing costs.”

i. **Quality of Alternative Solutions:** This subtheme focused on the overall quality of the provided alternative solutions. Some of the comments in this subtheme included “thoughtful alternatives,” “offered realistic alternatives,” “alternatives were not clearly explained,” and “not a convincing alternative to reduce printing costs.”
4. Overall Quality of Samples: This theme focused on issues relating to the overall quality of the writing samples.
   
a. Clarity/Ease of Reading: This subtheme included comments that mentioned how clear a writing sample was or how easy it was to read a sample. Some of the comments coded in this subtheme include “clearly written,” “easy to read,” “difficult to read,” and “very hard to understand and follow.”
   
b. Overall Quality: Comments in this subtheme addressed the overall quality of the writing samples. Examples include “written well,” “casually written,” “good essay,” and “very basic writing.”
   
c. Length: This subtheme included comments that mentioned the length of the samples, such as “concise” or “short.”
   
d. Overall Structure/Organization: This subtheme addressed issues relating to the overall structure or organization of the writing sample. Some examples of these comments are “organized,” “paragraphs should be combined,” “flows well,” “choppy,” and “good placement of paragraphs.”
   
e. Specific Elements of Structure/Organization: This subtheme included comments that discussed specific features of the structure or organization of a writing sample, such as “no conclusion,” “provides introduction to catch the reader’s attention,” “strong concluding sentence,” and “jumps right into their opinion.”

To be sure that the coding process had been conducted appropriately, several validity checks were used. First, the primary investigators codes were compared to codes given by the raters who provided the comments. During this process, all of the comments were first coded by the primary investigator. Each rater was then given a clean, typed copy of her comments and
asked to code each comment as being positive (+) or negative (-) in nature. The primary investigator then compared her positive and negative codes to those given by the raters and changed any discrepancies to reflect the code given by the rater. This step ensured that the primary investigator had accurately interpreted the intended meanings of the raters’ comments. Overall agreement between the positive and negative codes of the raters and the primary investigator was 99.1%. Additionally, all of the themes and subthemes generated by the primary investigator were reviewed by her advisor to ensure that the comments were appropriately coded and that the created themes and subthemes matched the patterns seen in the data.

To code spelling errors, the primary investigator reviewed the spelling errors made in the writing samples to get a general sense of the data. Spelling were then coded and observations were recorded in field notes. The following codes for spelling were used during this procedure:

1. *Apostrophe*: These types of errors include missing or inappropriately placed apostrophes, such as *its* for *it’s* or *dont* for *don’t*.

2. *Apostrophe/Morphological*: These types of errors included words that had an *Apostrophe* error (described above) as well as the inappropriate use of a bound morpheme, such as *universities* for *university’s*.

3. *Correct*: These words were spelled correctly, but had been coded as misspelled by one of the raters. This code was only awarded to two of the 142 words coded as misspelled in the writing samples.

4. *Homophone*: Words in this category included homophones that were used in the wrong context, such as *aloud* for *allowed*, *some* for *sum*, and *outway* for *outweigh*. 
5. **Orthographic:** In this category, words were spelled using all of the necessary sounds in the word, but not all of these sounds were represented correctly. Examples of orthographical errors include *neccessary* for *necessary* or *sincerly* for *sincerely*.

6. **Phonological:** Words in this category did not include all of the correct sounds, meaning that sounds were added, omitted, substituted, and/or reversed. Some examples of **Phonological** errors include *Dember* for *December*, *thiry* for *thirty*, *unnnessary* for *unnecessary*, and *countiously* for *continuously*.

7. **Phonological/Morphological:** Words in this category had both a **Phonological** error (described above) and issues with bound morphemes, such as *driv* for *driven* or *countious* for *continuously*.

8. **Phonological/Orthographic:** In this category, words had both **Phonological** (described above) and **Orthographic** errors (described above), such as *recyleing* for *recycling* (i.e., **Phonological** error = omitted *c*, **Orthographic** error = failing to remove silent-e to add -ing).

9. **Phonological/Wrong Word:** Words in this category were the wrong word for the context, which may have been due to a **Phonological** spelling error (described above). Examples include *inquire* for *acquire* and *signed* for *sided*.

To code punctuation errors, the primary investigator first identified the writing samples that had ending punctuation errors. Sentences coded as having ending punctuation errors were then reviewed so that a general sense of the data could be gathered. Next, the primary investigator coded all ending punctuation errors as being due to either an inappropriate use of ending punctuation or the presence of a run-on sentence. Once ending punctuation errors had been reviewed, the primary investigator examined the other punctuation errors in the writing
samples. Instances of other punctuation errors due to run-on sentences or sentence fragments were flagged, as other types of other punctuation errors (i.e., primarily comma errors) were frequent and did not appear to differ between typical participants and participants with WLDs. Observations of punctuation errors were recorded in field notes.

**Data representation.** After validity checks for the coded comments were complete, data regarding the number and types of comment codes assigned to each sample were entered into a spreadsheet. The information from the spreadsheet was then used to create tables to represent the qualitative data regarding comment themes and subthemes for each research question. Additionally, a summary of the findings was provided in text. When appropriate, tables and text regarding observations of spelling and punctuation errors were generated.

For the first research question, tables were created to show how often comments of each particular theme were given for a sample by each rater. Inter-rater reliability was qualitatively analyzed by comparing codes between raters, while intra-rater reliability was analyzed by comparing codes within raters. These comparisons were made to show how consistently the raters used various criteria (i.e., Linguistic, Mechanics, Content, and Quality features of writing) to generate holistic scores. A discussion of these results was provided in text.

For the second research question, tables were made to show the average number of comments given to each participant group (i.e., typical participants versus participants with WLDs) in the various themes/subthemes and the types of spelling errors made by each group. Additionally, the four participants with WLDs were matched to four typical participants using gender, age, year in school, and PPVT-4 standard scores (see Table 9 for demographic information). Tables were used to visually represent the similarities and differences within each of these pairs regarding scores on the rating scales, percentages of word- and sentence-level
features (e.g., spelling errors and punctuation errors), and holistic comment themes and subthemes. Additionally, a summary of these findings was provided in text.

For the last research question, comparisons were made between all participants with high and low holistic scores, as well as between matched pairs of participants with high and low holistic scores. To compare all participants with high (i.e., score of 4) versus low (i.e., score of 1) holistic scores, tables were generated to visually represent the average number of comments given to each participant group in the various comment themes/subthemes and the types of spelling errors made between groups. To compare matched pairs, the three participants who received holistic scores of “1” were matched with three participants who received holistic scores of “4” using gender, age, year in school, and PPVT-4 standard scores (see Table 10 for demographic information). When matching participants, priority was given to matching year in school and PPVT-4 standard scores. Tables were created to show the similarities and differences within each of these pairs on the rating scales, percentages of word- and sentence-level features of writing, and comment themes/subthemes. Additionally, a description of these similarities and differences was given in text.
Table 9

**Demographic Information for Matched Pairs of Typical Participants and Participants with WLDs**

<table>
<thead>
<tr>
<th>Pair</th>
<th>Gender</th>
<th>Age (years)</th>
<th>Year in School</th>
<th>Race/Ethnicity</th>
<th>PPVT-4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TYP WLD</td>
<td>TYP WLD</td>
<td>TYP WLD</td>
<td>TYP WLD</td>
<td>TYP WLD</td>
</tr>
<tr>
<td>1</td>
<td>Female Female</td>
<td>18.58 18.67</td>
<td>Freshman Freshman</td>
<td>White/Caucasian</td>
<td>Not Provided</td>
</tr>
<tr>
<td>2</td>
<td>Female Female</td>
<td>20.33 20.33</td>
<td>Sophomore Sophomore</td>
<td>Black/African American</td>
<td>White/Caucasian</td>
</tr>
<tr>
<td>3</td>
<td>Female Female</td>
<td>18.92 18.83</td>
<td>Freshman Freshman</td>
<td>White/Caucasian</td>
<td>Pacific Islander</td>
</tr>
<tr>
<td>4</td>
<td>Male Male</td>
<td>19.75 20.25</td>
<td>Sophomore Freshman</td>
<td>White/Caucasian</td>
<td>White/Caucasian</td>
</tr>
</tbody>
</table>

*Note. TYP = typical participants; WLD = participants with WLDs.*

Table 10

**Demographic Information for Matched Pairs of Participants with High and Low Holistic Scores**

<table>
<thead>
<tr>
<th>Pair</th>
<th>Gender</th>
<th>Age (years)</th>
<th>Year in School</th>
<th>Race/Ethnicity</th>
<th>PPVT-4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High Low</td>
<td>High Low</td>
<td>High Low</td>
<td>High Low</td>
<td>High Low</td>
</tr>
<tr>
<td>1</td>
<td>Female Female</td>
<td>18.25 18.58</td>
<td>Freshman Freshman</td>
<td>White/Caucasian</td>
<td>White/Caucasian</td>
</tr>
<tr>
<td>2</td>
<td>Female Female</td>
<td>20.08 22.17</td>
<td>Sophomore Junior</td>
<td>White/Caucasian</td>
<td>White/Caucasian</td>
</tr>
<tr>
<td>3</td>
<td>Female Male</td>
<td>18.92 20.25</td>
<td>Freshman Freshman</td>
<td>White/Caucasian</td>
<td>White/Caucasian</td>
</tr>
</tbody>
</table>

*Note. High = participants with high holistic scores; Low = participants with low holistic scores.*
CHAPTER IV
RESULTS

Question 1: Reliability of Rating Scales

Quantitative Results

To answer the first research question about how reliably rating scales can be used to measure the persuasive writing abilities of undergraduate students, inter-rater and intra-rater reliability were calculated for both the holistic and analytic rating scales. These procedures and results are discussed in the sections below.

Holistic scores. Inter-rater reliability for holistic scores was calculated for the twelve samples (21% of total samples) that all three raters scored both before and after receiving in-depth training on how to examine features of persuasive writing. Before receiving training, absolute agreement (i.e., both raters gave the same score) between raters ranged from 42% to 58%, and reasonable agreement (i.e., raters’ scores were within one point of each other) was 100% between all pairs of raters. After training, absolute agreement ranged from 50% to 75% and reasonable agreement was 100% between each pair of raters. Agreement between each pair of raters before and after training can be seen in Table 11.

Table 11

<table>
<thead>
<tr>
<th></th>
<th>Rater 1 vs. Rater 2</th>
<th>Rater 2 vs. Rater 3</th>
<th>Rater 1 vs. Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute</td>
<td>Reasonable</td>
<td>Absolute</td>
</tr>
<tr>
<td>Pre-Training</td>
<td>58%</td>
<td>100%</td>
<td>42%</td>
</tr>
<tr>
<td>Post-Training</td>
<td>50%</td>
<td>100%</td>
<td>50%</td>
</tr>
</tbody>
</table>
Intra-rater reliability was also calculated for 22% (6 of 27) of the holistic scores given by each of the three raters both before and after the in-depth training. Prior to the training, absolute intra-rater rater reliability ranged from 33% to 50% and reasonable intra-rater reliability ranged from 83% to 100%. After training, absolute intra-rater reliability increased to a range of 83% to 100% and reasonable agreement for all three raters was 100%. Agreement for each rater before and after training can be seen in Table 12.

Table 12

*Absolute and Reasonable Intra-Rater Reliability for Holistic Scores Before and After In-Depth Training*

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute</td>
<td>Reasonable</td>
<td>Absolute</td>
</tr>
<tr>
<td>Pre-Training</td>
<td>33%</td>
<td>83%</td>
</tr>
<tr>
<td>Post-Training</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Following in-depth training about how to analyze various features of persuasive writing, increases in both inter-rater and intra-rater reliability were seen. On average, inter-rater reliability increased by 5.7% and intra-rater reliability increased by 44.3%.

**Analytic scores.** Inter-rater reliability was calculated for each of the categories of the analytic rating scale using the 12 samples scored by all three raters (i.e., 21% of total samples). When possible, both absolute and reasonable agreement were calculated. Absolute agreement was present when the raters recorded the exact same scores; reasonable agreement was met when the raters recorded scores within one point of each other. For categories that required raters to code individual words or sentences in the samples (i.e., *Sense of Audience, Spelling, Ending Punctuation, Other Punctuation, Sentence Complexity, and Sentence Soundness*), reasonable agreement was unable to be calculated. Absolute inter-rater reliability for each of the categories
ranged from 33% to 100%, and reasonable inter-rater reliability ranged from 75% to 100%.

Inter-rater reliability for each category of the analytic rating scale can be seen in Table 13.

Table 13

*Inter-Rater Reliability for Analytic Rating Scale Scores and Coded Features of Writing*

<table>
<thead>
<tr>
<th>Category</th>
<th>Rater 1 vs. Rater 2</th>
<th>Rater 2 vs. Rater 3</th>
<th>Rater 1 vs. Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute</td>
<td>Reasonable</td>
<td>Absolute</td>
</tr>
<tr>
<td>Presence of Prompt Components</td>
<td>67% 100%</td>
<td>67% 100%</td>
<td>75% 92%</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>33% 83%</td>
<td>50% 100%</td>
<td>33% 83%</td>
</tr>
<tr>
<td>Sense of Audience</td>
<td>99%-----</td>
<td>96%-----</td>
<td>96%-----</td>
</tr>
<tr>
<td>Organization</td>
<td>33% 83%</td>
<td>75% 100%</td>
<td>50% 92%</td>
</tr>
<tr>
<td>Word Choice</td>
<td>42% 92%</td>
<td>42% 92%</td>
<td>42% 75%</td>
</tr>
<tr>
<td>Spelling</td>
<td>61%-----</td>
<td>67%-----</td>
<td>61%-----</td>
</tr>
<tr>
<td>Ending Punctuation</td>
<td>100%-----</td>
<td>99%-----</td>
<td>99%-----</td>
</tr>
<tr>
<td>Other Punctuation</td>
<td>85%-----</td>
<td>86%-----</td>
<td>84%-----</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>86%-----</td>
<td>91%-----</td>
<td>86%-----</td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td>81% 100%</td>
<td>76% 100%</td>
<td>84% 100%</td>
</tr>
</tbody>
</table>

*Note.* Only absolute agreement was calculated for *Sense of Audience, Spelling, Ending Punctuation, Other Punctuation, Sentence Complexity,* and *Sentence Soundness,* as reasonable agreement could not be calculated for the codes given to individual words or sentences.
Table 14

_Intra-Rater Reliability for Analytic Rating Scale Scores and Coded Features of Writing_

<table>
<thead>
<tr>
<th>Category</th>
<th>Rater 1 Absolute</th>
<th>Rater 1 Reasonable</th>
<th>Rater 2 Absolute</th>
<th>Rater 2 Reasonable</th>
<th>Rater 3 Absolute</th>
<th>Rater 3 Reasonable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence of Prompt Components</td>
<td>83%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>83%</td>
<td>100%</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>83%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Sense of Audience</td>
<td>100%</td>
<td>–</td>
<td>100%</td>
<td>–</td>
<td>97%</td>
<td>–</td>
</tr>
<tr>
<td>Organization</td>
<td>67%</td>
<td>100%</td>
<td>83%</td>
<td>100%</td>
<td>67%</td>
<td>100%</td>
</tr>
<tr>
<td>Word Choice</td>
<td>17%</td>
<td>67%</td>
<td>67%</td>
<td>83%</td>
<td>83%</td>
<td>83%</td>
</tr>
<tr>
<td>Spelling</td>
<td>73%</td>
<td>–</td>
<td>67%</td>
<td>–</td>
<td>60%</td>
<td>–</td>
</tr>
<tr>
<td>Ending Punctuation</td>
<td>100%</td>
<td>–</td>
<td>100%</td>
<td>–</td>
<td>100%</td>
<td>–</td>
</tr>
<tr>
<td>Other Punctuation</td>
<td>88%</td>
<td>–</td>
<td>90%</td>
<td>–</td>
<td>93%</td>
<td>–</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>90%</td>
<td>–</td>
<td>91%</td>
<td>–</td>
<td>88%</td>
<td>–</td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td>81%</td>
<td>100%</td>
<td>83%</td>
<td>100%</td>
<td>89%</td>
<td>100%</td>
</tr>
</tbody>
</table>

*Note.* Only absolute agreement was calculated for *Sense of Audience, Spelling, Ending Punctuation, Other Punctuation, Sentence Complexity,* and *Sentence Soundness,* as reasonable agreement could not be calculated for the codes given to individual words or sentences.

Intra-rater reliability was also calculated when possible for each category of the analytic rating scale using six samples (i.e., 22% of 27) that each of the raters scored twice (Table 14).

Again, both absolute and reasonable agreement were calculated when possible. The range of
overall absolute intra-rater reliability for all categories was 17% to 100%, while the range for overall reasonable intra-rater reliability was 67% to 100%.

To examine how differences in the coding of words and sentences by raters impacted the scores given to participants on the analytic rating scale, inter-rater reliability was also calculated using the analytic rating scale scores that resulted from the codes provided by raters. For example, for Sense of Audience, the percentage of sentences that were coded as inappropriate in each sample by each rater was calculated and then converted to the corresponding score on the analytic rating scale (i.e., “4” = 0% inappropriate sentences, “3” = 0.1% to 5.0% inappropriate sentences, “2” = 5.1% to 20.0% inappropriate sentences, “1” = 20.0% inappropriate sentences). Both absolute and reasonable agreement were calculated. The ranges of absolute and reasonable agreement for raters across all categories were 42% to 83% and 67% to 100%, respectively. Overall inter-rater reliability for these derived scores can be seen in Table 15.

Table 15

*Inter-Rater Reliability when Coded Features of Writing were Converted to Analytic Scores*

<table>
<thead>
<tr>
<th>Category</th>
<th>Rater 1 vs. Rater 2</th>
<th></th>
<th>Rater 2 vs. Rater 3</th>
<th></th>
<th>Rater 1 vs. Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute</td>
<td>Reasonable</td>
<td>Absolute</td>
<td>Reasonable</td>
<td>Absolute</td>
</tr>
<tr>
<td>Sense of Audience</td>
<td>83%</td>
<td>83%</td>
<td>67%</td>
<td>75%</td>
<td>67%</td>
</tr>
<tr>
<td>Spelling</td>
<td>83%</td>
<td>100%</td>
<td>75%</td>
<td>100%</td>
<td>58%</td>
</tr>
<tr>
<td>Punctuation</td>
<td>50%</td>
<td>83%</td>
<td>58%</td>
<td>92%</td>
<td>42%</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>50%</td>
<td>92%</td>
<td>83%</td>
<td>100%</td>
<td>50%</td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td>50%</td>
<td>83%</td>
<td>42%</td>
<td>67%</td>
<td>67%</td>
</tr>
</tbody>
</table>
Intra-rater reliability was also calculated using the analytic rating scale scores that resulted from the codes provided by raters for the six samples scored twice by each rater. Again, the total numbers of words or sentences coded for each of the categories of Sense of Audience, Spelling, Punctuation, Sentence Complexity, and Sentence Soundness were converted into percentages for each sample, which were then used to determine the appropriate score on the analytic rating scale. Absolute intra-rater reliability for raters ranged from 50% to 100% and reasonable intra-rater reliability ranged from 67% to 100%. Overall intra-rater reliability for all of these derived scores can be seen for each rater in Table 16.

Table 16

Intra-Rater Reliability when Coded Features of Writing were Converted to Analytic Scores

<table>
<thead>
<tr>
<th>Category</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute</td>
<td>Reasonable</td>
<td>Absolute</td>
</tr>
<tr>
<td>Sense of Audience</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Spelling</td>
<td>83%</td>
<td>100%</td>
<td>50%</td>
</tr>
<tr>
<td>Punctuation</td>
<td>50%</td>
<td>100%</td>
<td>67%</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>83%</td>
<td>100%</td>
<td>50%</td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td>50%</td>
<td>67%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Qualitative Results

Inter-rater reliability. To qualitatively examine inter-rater reliability, the types of comments given by raters during the holistic scoring process were compared. A comparison of the themes of the comments given by raters showed how often they were considering the same features of writing when awarding holistic scores.
Table 17 shows how often each rater provided comments of various themes while scoring the 12 samples scored by all raters during the first holistic scoring procedure. Overall, all raters appeared to primarily rely on the use of Content comments, while the use of other comment themes varied by rater.

Table 17

<table>
<thead>
<tr>
<th>Comment Theme</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic</td>
<td>33.3%</td>
<td>16.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Mechanics</td>
<td>33.3%</td>
<td>0.0%</td>
<td>8.1%</td>
</tr>
<tr>
<td>Content</td>
<td>83.3%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Quality</td>
<td>8.1%</td>
<td>41.7%</td>
<td>58.3%</td>
</tr>
</tbody>
</table>

The results of a comparison of comment themes provided by raters for each sample during the first holistic scoring procedure are shown in Table 18. This comparison revealed that the highest level of agreement between raters was achieved for Content comments, as the average level of agreement between raters was 88.9%. Average levels of agreement for Mechanics, Linguistic, and Quality comments were 58.3%, 44.4%, and 36.1%, respectively. However, most instances of agreement for these last three categories were due to the absence of comments in these themes for both raters in each pair.

Table 18

<table>
<thead>
<tr>
<th>Comment Theme</th>
<th>Rater 1 vs. Rater 2</th>
<th>Rater 2 vs. Rater 3</th>
<th>Rater 1 vs. Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic</td>
<td>83.3%</td>
<td>16.7%</td>
<td>33.3%</td>
</tr>
<tr>
<td>Mechanics</td>
<td>58.3%</td>
<td>58.3%</td>
<td>58.3%</td>
</tr>
<tr>
<td>Content</td>
<td>83.3%</td>
<td>100.0%</td>
<td>83.3%</td>
</tr>
<tr>
<td>Quality</td>
<td>33.3%</td>
<td>41.7%</td>
<td>33.3%</td>
</tr>
</tbody>
</table>

Table 19 shows how often each rater provided comments of various themes while scoring the 12 samples scored by all raters during the second holistic scoring procedure. Again, all of
the raters appeared to primarily rely on the use of *Content* comments, while the use of other comment themes varied by rater.

**Table 19**

*Proportion of Samples Given Specific Comment Themes by Each Rater During Holistic II Scoring*

<table>
<thead>
<tr>
<th>Comment Theme</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Linguistic</em></td>
<td>50.0%</td>
<td>83.3%</td>
<td>41.7%</td>
</tr>
<tr>
<td><em>Mechanics</em></td>
<td>66.7%</td>
<td>0.0%</td>
<td>33.3%</td>
</tr>
<tr>
<td><em>Content</em></td>
<td>83.3%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td><em>Quality</em></td>
<td>0.0%</td>
<td>50.0%</td>
<td>83.3%</td>
</tr>
</tbody>
</table>

Table 20 shows the results of a comparison of comment themes provided by raters for each sample during the second holistic scoring procedure. This comparison revealed that the highest level of agreement between raters was again achieved for *Content* comments, as the average level of agreement between raters remained 88.9%. Average levels of agreement for *Linguistic, Quality,* and *Mechanics* comments were 55.6%, 33.4%, and 33.3%, respectively. Agreement in these last three categories decreased from the first holistic scoring procedure to the second holistic scoring procedure because all raters awarded more comments across themes but usually not to the same samples.

**Table 20**

*Agreement Among Raters for Comment Themes Given During Holistic II Scoring*

<table>
<thead>
<tr>
<th>Comment Theme</th>
<th>Rater 1 vs. Rater 2</th>
<th>Rater 2 vs. Rater 3</th>
<th>Rater 1 vs. Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Linguistic</em></td>
<td>66.7%</td>
<td>58.3%</td>
<td>41.7%</td>
</tr>
<tr>
<td><em>Mechanics</em></td>
<td>25.0%</td>
<td>25.0%</td>
<td>50.0%</td>
</tr>
<tr>
<td><em>Content</em></td>
<td>83.3%</td>
<td>100.0%</td>
<td>83.3%</td>
</tr>
<tr>
<td><em>Quality</em></td>
<td>16.7%</td>
<td>66.7%</td>
<td>16.7%</td>
</tr>
</tbody>
</table>

**Intra-rater reliability.** To examine intra-rater reliability in a qualitative manner, themes of holistic comments were compared within each rater during both the pre-training and post-
training holistic scoring processes. This comparison revealed how consistently each rater considered the same themes while scoring samples using the holistic rating scale.

Table 21 shows how often each rater provided comments within each theme while scoring six samples twice during the first phase of holistic scoring. These results revealed that *Content* comments were provided most often by all three raters, while the use of other comment themes varied by rater.

Table 21

*Proportion of Samples Given Specific Comment Themes by Each Rater During Holistic I Scoring*

<table>
<thead>
<tr>
<th>Comment Theme</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Linguistic</td>
<td>33.3%</td>
<td>50.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Mechanics</td>
<td>33.3%</td>
<td>50.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Content</td>
<td>66.7%</td>
<td>83.3%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Quality</td>
<td>0.0%</td>
<td>50.0%</td>
<td>83.3%</td>
</tr>
</tbody>
</table>

*Note.* 1 = Initial scoring of samples; 2 = Rescoring of samples.

Table 22 shows the results of a comparison of the types of comment themes given by each rater during the initial scoring and rescoring of six samples during the first holistic scoring procedure. Intra-rater agreement for each comment theme was calculated by dividing the number of samples in agreement with regard to the presence or absence of a theme in both the initial and rescored sample by the total number of samples (i.e., six), while total agreement was calculated by dividing the total instances of agreement for all four themes by the total number of opportunities for thematic agreement (i.e., 24). These results revealed that Rater 2 was the most consistent in her use of comment themes during the first holistic scoring procedure, followed by Raters 3 and 1, respectively.
Table 22

*Intra-Rater Agreement for Comment Themes Given During Holistic I Scoring*

<table>
<thead>
<tr>
<th>Comment Theme</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic</td>
<td>50.0%</td>
<td>83.3%</td>
<td>83.3%</td>
</tr>
<tr>
<td>Mechanics</td>
<td>50.0%</td>
<td>100.0%</td>
<td>66.7%</td>
</tr>
<tr>
<td>Content</td>
<td>83.3%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Quality</td>
<td>50.0%</td>
<td>66.7%</td>
<td>50.0%</td>
</tr>
<tr>
<td>All Themes</td>
<td>58.3%</td>
<td>87.5%</td>
<td>75.0%</td>
</tr>
</tbody>
</table>

Table 23 shows how often each rater provided comments within each theme while scoring six samples twice during the second holistic scoring procedure. Consistent with the first phase of the holistic scoring procedure, *Content* comments were again provided most often by all three raters and the use of other comment themes varied by rater.

Table 23

*Proportion of Samples Given Specific Comment Themes by Each Rater During Holistic II Scoring*

<table>
<thead>
<tr>
<th>Comment Theme</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Linguistic</td>
<td>66.7%</td>
<td>33.3%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Mechanics</td>
<td>100.0%</td>
<td>83.3%</td>
<td>50.0%</td>
</tr>
<tr>
<td>Content</td>
<td>50.0%</td>
<td>66.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Quality</td>
<td>16.7%</td>
<td>33.3%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

*Note.* 1 = Initial scoring of samples; 2 = Rescoring of samples.

The results of a comparison of the types of comment themes given during both the initial scoring and rescoring of samples during the second holistic scoring procedure are shown in Table 24. Intra-rater agreement was calculated the same way as for the first holistic scoring procedure (described above). The results of these comparisons revealed that Rater 2 was the most consistent in her use of comment themes during the second holistic scoring procedure, followed by Raters 3 and 1, respectively.
Table 24

Intra-Rater Agreement for Comment Themes Given During Holistic II Scoring

<table>
<thead>
<tr>
<th>Comment Theme</th>
<th>Rater 1</th>
<th>Rater 2</th>
<th>Rater 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic</td>
<td>33.3%</td>
<td>66.7%</td>
<td>50.0%</td>
</tr>
<tr>
<td>Mechanics</td>
<td>83.3%</td>
<td>100.0%</td>
<td>66.7%</td>
</tr>
<tr>
<td>Content</td>
<td>50.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Quality</td>
<td>83.3%</td>
<td>50.0%</td>
<td>50.0%</td>
</tr>
<tr>
<td>All Themes</td>
<td>62.5%</td>
<td>79.2%</td>
<td>66.7%</td>
</tr>
</tbody>
</table>

An analysis of changes within each rater from the first to second holistic scoring procedure showed that intra-rater agreement decreased from pre- to post-training for Linguistic comments in all raters. For Mechanics, agreement increased for Rater 1 and stayed the same for Raters 2 and 3. For Content, agreement decreased for Rater 1 and stayed the same for Raters 2 and 3. Finally, for Quality, agreement increased for Rater 1, decreased for Rater 2, and stayed the same for Rater 3.

Question 2: Comparison of Typical Participants and Participants with WLDs

Quantitative Results

Means and standard deviations of holistic and analytic rating scale scores were calculated for typical participants, participants with WLDs, and all participants. These results, shown in Table 25, indicate that typical participants obtained higher means than participants with WLDs for holistic scores (I and II), Logical Reasoning, Organization, Word Choice, Spelling, Punctuation, Sentence Soundness, and total analytic rating scale scores. Conversely, participants with WLDs had higher means than typical participants for Presence of Prompt Components and Sense of Audience.
Table 25

*Means and Standard Deviations of Holistic and Analytic Rating Scale Scores for Participants*

<table>
<thead>
<tr>
<th>Category</th>
<th>Typical Participants</th>
<th>WLD Participants</th>
<th>All Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>M</em></td>
<td><em>SD</em></td>
<td><em>M</em></td>
</tr>
<tr>
<td>Holistic I</td>
<td>2.88</td>
<td>0.85</td>
<td>2.00</td>
</tr>
<tr>
<td>Holistic II</td>
<td>2.88</td>
<td>0.63</td>
<td>1.75</td>
</tr>
<tr>
<td>Presence of Prompt Components</td>
<td>3.62</td>
<td>0.60</td>
<td>3.75</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>2.88</td>
<td>0.66</td>
<td>2.25</td>
</tr>
<tr>
<td>Sense of Audience</td>
<td>3.74</td>
<td>0.66</td>
<td>4.00</td>
</tr>
<tr>
<td>Organization</td>
<td>2.98</td>
<td>0.55</td>
<td>2.50</td>
</tr>
<tr>
<td>Word Choice</td>
<td>2.76</td>
<td>0.80</td>
<td>1.75</td>
</tr>
<tr>
<td>Spelling</td>
<td>2.90</td>
<td>0.79</td>
<td>1.75</td>
</tr>
<tr>
<td>Punctuation</td>
<td>2.30</td>
<td>1.11</td>
<td>1.25</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>2.22</td>
<td>0.89</td>
<td>2.75</td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td>1.88</td>
<td>1.06</td>
<td>1.25</td>
</tr>
<tr>
<td>Total analytic score</td>
<td>25.28</td>
<td>3.49</td>
<td>21.25</td>
</tr>
</tbody>
</table>

Means and standard deviations were also calculated for the percentages calculated using word- and sentence-level codes given by raters. These results, shown in Table 26, revealed that typical participants produced a higher proportion of inappropriate sentences for the audience of a university president, fewer spelling errors, more ending punctuation errors, fewer other punctuation errors, fewer complex sentences, and more sound sentences than those with WLDs.
Table 26

Means and Standard Deviations of Percentages of Coded Words and Sentences of Participants

<table>
<thead>
<tr>
<th>Category</th>
<th>Typical Participants</th>
<th>WLD Participants</th>
<th>All Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Inappropriate sentences</td>
<td>1.2%</td>
<td>3.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Spelling errors</td>
<td>0.8%</td>
<td>0.8%</td>
<td>1.9%</td>
</tr>
<tr>
<td>Ending punctuation errors</td>
<td>1.2%</td>
<td>2.9%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Other punctuation errors</td>
<td>13.3%</td>
<td>11.2%</td>
<td>41.4%</td>
</tr>
<tr>
<td>Complex sentences</td>
<td>55.4%</td>
<td>15.8%</td>
<td>62.7%</td>
</tr>
<tr>
<td>Sound sentences</td>
<td>66.8%</td>
<td>19.1%</td>
<td>41.0%</td>
</tr>
</tbody>
</table>

To determine whether or not the differences seen between the two participant groups were statistically significant, Fisher’s exact test was used to compare the proportions of typical participants and participants with WLDs receiving high scores (i.e., scores of three or four) versus low scores (i.e., scores of one or two) on the holistic and analytic rating scales. These results are shown in Table 27.

The only statistically significant differences seen with Fisher’s exact test were in the areas of holistic score I \( (p = 0.02) \), holistic score II \( (p = < 0.01) \), and Spelling \( (p = 0.04) \). In all cases, participants with WLDs had a significantly lower proportion of high scores and a significantly higher proportion of low scores than typical participants. The categories of Logical Reasoning and Word Choice also approached significance with the same pattern.
Table 27

Prevalence of High and Low Holistic and Analytic Rating Scale Scores in Typical Participants 
(n = 50) and Participants with WLDs (n = 4)

<table>
<thead>
<tr>
<th>Category</th>
<th>Typical participants</th>
<th></th>
<th>Participants with WLDs</th>
<th></th>
<th></th>
<th></th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
<td>n</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>Holistic I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>33</td>
<td>66.0</td>
<td>17</td>
<td>34.0</td>
<td>4</td>
<td>100.0</td>
<td>0.02</td>
</tr>
<tr>
<td>Low</td>
<td>17</td>
<td>34.0</td>
<td>0</td>
<td>0.0</td>
<td>2</td>
<td>50.0</td>
<td></td>
</tr>
<tr>
<td>Holistic II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>39</td>
<td>78.0</td>
<td>11</td>
<td>22.0</td>
<td>4</td>
<td>100.0</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Low</td>
<td>11</td>
<td>22.0</td>
<td>0</td>
<td>0.0</td>
<td>4</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>Presence of Prompt Components</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>47</td>
<td>94.0</td>
<td>3</td>
<td>6.0</td>
<td>4</td>
<td>100.0</td>
<td>1.00</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>72.0</td>
<td>14</td>
<td>28.0</td>
<td>1</td>
<td>25.0</td>
<td>0.09</td>
</tr>
<tr>
<td>Sense of Audience</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>44</td>
<td>88.0</td>
<td>6</td>
<td>12.0</td>
<td>4</td>
<td>100.0</td>
<td>1.00</td>
</tr>
<tr>
<td>Organization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>42</td>
<td>84.0</td>
<td>8</td>
<td>16.0</td>
<td>2</td>
<td>50.0</td>
<td>0.15</td>
</tr>
<tr>
<td>Word Choice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>70.0</td>
<td>15</td>
<td>30.0</td>
<td>1</td>
<td>25.0</td>
<td>0.10</td>
</tr>
<tr>
<td>Spelling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>80.0</td>
<td>10</td>
<td>20.0</td>
<td>1</td>
<td>25.0</td>
<td>0.04</td>
</tr>
<tr>
<td>Punctuation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>32.0</td>
<td>34</td>
<td>68.0</td>
<td>0</td>
<td>0.0</td>
<td>0.31</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>40.0</td>
<td>30</td>
<td>60.0</td>
<td>3</td>
<td>75.0</td>
<td>0.30</td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>26.0</td>
<td>37</td>
<td>74.0</td>
<td>0</td>
<td>0.0</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Comparisons between the performance of typical participants and participants with WLDs were also made using t-tests (Table 28). The t-tests that were run compared the average percentages derived using raters’ codes of individual words and sentences in the samples, as well as the average total analytic rating scale scores obtained by both groups of participants. These results revealed significant differences in the areas of spelling errors \([t(52) = -2.50, p = 0.02]\), other punctuation errors \([t(52) = -4.58, p < 0.01]\), sound sentences \([t(52) = 2.56, p = 0.01]\), and total analytic scores \([t(52) = 2.28, p = 0.03]\). A comparison of the means between groups in
these areas showed that participants with WLDs made more spelling errors, made more other punctuation errors, produced fewer sound sentences, and received lower total analytic rating scale scores compared to typical participants. Using Cohen’s (1988) recommendations for interpreting effect size, large effect sizes were seen for spelling errors, other punctuation errors, sound sentences, and total analytic scores.

Table 28

*Group Differences for Percentages of Word- and Sentence-Level Features and Total Analytic Rating Scale Scores Between Typical Participants and Participants with WLDs*

<table>
<thead>
<tr>
<th>Category</th>
<th>Typical participants</th>
<th>Participants with WLDs</th>
<th>t(52)</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inappropriate sentences</td>
<td>1.16 3.04</td>
<td>0.00 0.00</td>
<td>0.76</td>
<td>0.45</td>
<td>0.39</td>
</tr>
<tr>
<td>Spelling errors</td>
<td>0.76 0.82</td>
<td>1.85 1.03</td>
<td>-2.50</td>
<td>0.02</td>
<td>1.31</td>
</tr>
<tr>
<td>Ending punctuation errors</td>
<td>1.22 2.94</td>
<td>0.00 0.00</td>
<td>0.82</td>
<td>0.41</td>
<td>0.43</td>
</tr>
<tr>
<td>Other punctuation errors</td>
<td>13.29 11.18</td>
<td>41.36 19.31</td>
<td>-4.58</td>
<td>&lt;0.01</td>
<td>2.38</td>
</tr>
<tr>
<td>Complex sentences</td>
<td>55.40 15.84</td>
<td>62.73 9.99</td>
<td>-0.91</td>
<td>0.37</td>
<td>0.47</td>
</tr>
<tr>
<td>Sound sentences</td>
<td>66.75 19.07</td>
<td>40.97 24.28</td>
<td>2.56</td>
<td>0.01</td>
<td>1.33</td>
</tr>
<tr>
<td>Total analytic score</td>
<td>25.28 3.49</td>
<td>21.25 1.26</td>
<td>2.28</td>
<td>0.03</td>
<td>1.18</td>
</tr>
</tbody>
</table>

*Note.* To interpret Cohen’s $d$, values of 0.20 to 0.49 are small effect sizes, values of 0.50 to 0.79 are medium effect sizes, and values of $\geq 0.80$ are large effect sizes.

**Qualitative Results**

All participants. To qualitatively analyze the differences between the writing samples of all typical participants and all participants with WLDs, comparisons were made between the types of comments given to each group of participants. When all of the comments were considered, 55.6% of the comments given to typical participants and 18.8% of the comments
given to participants with WLDs were positive in nature during the first holistic scoring procedure. When the samples were scored holistically a second time, 57.0% of the comments given to typical participants and 19.4% of the comments given to participants with WLDs were positive in nature.

To more closely analyze the similarities and differences in the comments given to the two participant groups, the average number of positive and negative comments in each theme and subtheme was also calculated for each group (Table 29). A comparison of the average number of positive and negative themes during the first and second holistic scoring procedures revealed that the typical participants received a higher average number of positive comments and a lower average number of negative comments across all themes. The only exception to this pattern was for Mechanics comments, as both typical participants and participants with WLDs received an equal average number of positive comments in this theme during both the first and second holistic scoring procedures, as well as a nearly equal average number of negative comments during the second holistic scoring.

When the subthemes were compared, typical participants usually received a higher or equal average number of positive comments across subthemes compared to participants with WLDs. More specifically, typical participants received a higher average number of positive comments than participants with WLDs during the first phase of holistic scoring in seven of nine Content subthemes and two of five Quality subthemes, as well as an equal average number of positive comments as participants with WLDs for four of four Linguistic subthemes, two of two Mechanics subthemes, one of nine Content subthemes, and three of five Quality subthemes. During the second phase of holistic scoring, typical participants received a higher average number of positive comments than participants with WLDs for two of four Linguistic subthemes,
six of nine Content subthemes, and two of five Quality subthemes, as well as an equal average number of positive comments as participants with WLDs for two of four Linguistic subthemes, two of two Mechanics subthemes, three of nine Content subthemes, and one of five Quality subthemes.

Conversely, participants with WLDs typically received a higher or equal average number of negative comments across subthemes compared to typical participants. More specifically, during the first phase of holistic scoring, participants with WLDs received a higher average number of negative comments than typical participants for three of four Linguistic subthemes, one of two Mechanics subthemes, two of nine Content subthemes, and one of five Quality subthemes, as well as an equal number of negative comments as typical participants for one of four Linguistic subthemes, one of two Mechanics subthemes, five of nine Content subthemes, and four of five Quality subthemes. During the second phase of holistic scoring, participants with WLDs received a higher average number of negative comments than typical participants for three of four Linguistic subthemes, one of two Mechanics subthemes, four of nine Content subthemes, and two of five Quality subthemes, as well as an equal average number of negative comments as typical participants for four of nine Content subthemes and two of five Quality subthemes.
### Table 29

**Average Number of Comments per Theme and Subtheme for Typical Participants (n = 50) and Participants with WLDs (n = 4)**

<table>
<thead>
<tr>
<th>Comment Type</th>
<th>Positive Comments</th>
<th>Negative Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holistic I</td>
<td>Holistic II</td>
</tr>
<tr>
<td></td>
<td>TYP</td>
<td>WLD</td>
</tr>
<tr>
<td><strong>Linguistic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Sentence Structure</em></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><em>Wording</em></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><em>Clarity</em></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><em>Grammar/Syntax</em></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Total Linguistic</strong></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Mechanics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Spelling</em></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><em>Punctuation</em></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Total Mechanics</strong></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Content</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Presence All</em></td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td><em>Support All</em></td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td><em>Strength All</em></td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td><em>Presence Claim/Reasons</em></td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td><em>Support Claim/Reasons</em></td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td><em>Strength Claim/Reasons</em></td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td><em>Presence Suggestions</em></td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td><em>Support Suggestions</em></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><em>Strength Suggestions</em></td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>Total Content</strong></td>
<td>1.5</td>
<td>0.8</td>
</tr>
<tr>
<td><strong>Quality</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Clarity</em></td>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td><em>Overall Quality</em></td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td><em>Length</em></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><em>Organization</em></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><em>Specific Organization</em></td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Total Quality</strong></td>
<td>0.6</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Total All Themes</strong></td>
<td>2.1</td>
<td>0.8</td>
</tr>
</tbody>
</table>

*Note. TYP = typical participants; WLD = participants with WLDs.*

The final qualitative analysis comparing typical participants and participants with WLDs focused on comparing the coded elements within the writing samples of these groups (i.e., inappropriate sentences, misspelled words, ending punctuation errors, and other punctuation elements).
errors). Sentences coded as inappropriate for the Sense of Audience rating scale category were unable to be compared between groups because none of the samples of participants with WLDs had any sentences coded as inappropriate. Additionally, complex sentences and sound sentences were not compared because it was not always clear exactly why raters had coded sentences as complex or sound. For spelling and punctuation errors, it was easy to determine why errors had been coded, as it was clear how the words were misspelled and the raters usually corrected the punctuation errors made in the samples.

Table 30

Proportions of Types of Spelling Errors Made by Typical Participants (n = 50) and Participants with WLDs (n = 4)

<table>
<thead>
<tr>
<th>Spelling Error</th>
<th>TYP</th>
<th>WLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apostrophe</td>
<td>16.5%</td>
<td>4.8%</td>
</tr>
<tr>
<td>Apostrophe/Morphological</td>
<td>0.8%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Correct</td>
<td>1.7%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Homophone</td>
<td>5.8%</td>
<td>14.3%</td>
</tr>
<tr>
<td>Orthographic</td>
<td>51.2%</td>
<td>38.1%</td>
</tr>
<tr>
<td>Phonological</td>
<td>21.5%</td>
<td>23.8%</td>
</tr>
<tr>
<td>Phonological/Morphological</td>
<td>1.7%</td>
<td>4.8%</td>
</tr>
<tr>
<td>Phonological/Orthographic</td>
<td>0.0%</td>
<td>4.8%</td>
</tr>
<tr>
<td>Phonological/Wrong Word</td>
<td>0.8%</td>
<td>9.5%</td>
</tr>
</tbody>
</table>

Note. TYP = typical participants; WLD = participants with WLDs.

When types of spelling errors were compared, it appeared that the two groups made different types of errors (Table 30). Firstly, participants with WLDs made a higher proportion of spelling errors containing multiple types of errors per word than typical participants (i.e., 33.3% versus 9.1%). Additionally, participants with WLDs made a higher proportion of phonological errors (i.e., 38.1% versus 24.0%) and a lower proportion of apostrophe errors (i.e., 4.8% versus 17.4%) than typical participants. However, both groups made an equal proportion of orthographic spelling errors (i.e., 57.1% for group with WLDs versus 57.0% for typical group).
When punctuation errors were compared, it was clear that typical participants made fewer total punctuation errors than participants with WLDs. However, based on the raters’ codes, only typical participants made ending punctuation errors. Closer analysis of the samples revealed that typical participants’ ending punctuation errors were primarily due to the incorrect use of ending punctuation (i.e., for 62.5% of coded sentences, the error involved using a period instead of a question mark, a question mark instead of a period, or accidentally omitting a period) rather than the presence of run-on sentences. However, participants with WLDs frequently produced run-on sentences that ended up being coded as other punctuation errors because they could be corrected with the use of semi-colons and commas. Overall, both groups demonstrated difficulty appropriately using commas, but participants with WLDs produced more punctuation errors due to the presence of sentence fragments and run-on sentences than typical participants.

**Matched pairs.** To further analyze the differences between the writing abilities of typical participants and participants with WLDs, a comparison of the performance of four typical participants and four participants with WLDs matched based on gender, age, year in school, and PPVT-4 standard scores was made. A comparison of their performance on the holistic and analytic rating scales can be seen in Table 31, while Table 32 compares their performance on the various measures initially calculated using percentages of words or sentences (i.e., inappropriate sentences, misspelled words, ending punctuation errors, other punctuation errors, complex sentences, and sound sentences).

Based on a comparison of scores on the holistic and analytic rating scales, typical participants outperformed participants with WLDs across all four pairs in the areas of Holistic I scores, **Spelling**, and total analytic scores. In three pairs, typical participants outperformed participants with WLDs in the areas of Holistic II scores, **Logical Reasoning**, and **Punctuation**.
For *Organization* and *Word Choice*, typical participants outperformed participants with WLDs in two of the pairs. Features of persuasive writing that either did not differ or did not consistently differ between groups in the same direction were *Presence of Prompt Components, Sense of Audience, Sentence Complexity*, and *Sentence Soundness*.

When participants’ percentages on the presence of various features of writing were compared, typical participants produced fewer spelling errors and fewer other punctuation errors than participants with WLDs across all four pairs. Across three pairs, typical participants produced more sound sentences and fewer complex sentences than participants with WLDs. Features that did not consistently differ in similar ways between participant groups were inappropriate sentences and ending punctuation errors.

The types of comments given to the writing samples of the matched pairs of participants during the holistic scoring procedures were also compared. Because a small number of comments were given to each participant, comments for all participants in each group were combined during this analysis. Comments were provided by all three raters for two participants in each group, while the other two participants in each group received comments from only one rater. Since an equal number of participants in each group received comments from all raters, all of these comments were included during this analysis.
Table 31

Comparison of Performance on Holistic and Analytic Rating Scales of Matched Pairs

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Pair 1 TYP</th>
<th>Pair 1 WLD</th>
<th>Pair 2 TYP</th>
<th>Pair 2 WLD</th>
<th>Pair 3 TYP</th>
<th>Pair 3 WLD</th>
<th>Pair 4 TYP</th>
<th>Pair 4 WLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holistic I</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Holistic II</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Presence of Prompt Components</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Sense of Audience</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Organization</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Word Choice</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Spelling</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Punctuation</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total analytic score</td>
<td>36</td>
<td>24</td>
<td>32</td>
<td>27</td>
<td>30</td>
<td>25</td>
<td>28</td>
<td>24</td>
</tr>
</tbody>
</table>

*Note.* TYP = typical participants; WLD = participants with WLDs.
### Table 32

*Comparison of Percentages for Matched Pairs*

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Pair 1</th>
<th></th>
<th>Pair 2</th>
<th></th>
<th>Pair 3</th>
<th></th>
<th>Pair 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TYP</td>
<td>WLD</td>
<td>TYP</td>
<td>WLD</td>
<td>TYP</td>
<td>WLD</td>
<td>TYP</td>
<td>WLD</td>
</tr>
<tr>
<td>Inappropriate sentences</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>11.11%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Misspelled words</td>
<td>0.54%</td>
<td>2.82%</td>
<td>0.34%</td>
<td>1.75%</td>
<td>0.00%</td>
<td>2.39%</td>
<td>0.00%</td>
<td>0.44%</td>
</tr>
<tr>
<td>Ending punctuation errors</td>
<td>8.33%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>5.56%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Other punctuation errors</td>
<td>8.33%</td>
<td>20.00%</td>
<td>20.00%</td>
<td>63.64%</td>
<td>18.18%</td>
<td>31.82%</td>
<td>0.00%</td>
<td>50.00%</td>
</tr>
<tr>
<td>Complex sentences</td>
<td>50.00%</td>
<td>60.00%</td>
<td>93.33%</td>
<td>72.73%</td>
<td>31.82%</td>
<td>68.18%</td>
<td>38.89%</td>
<td>50.00%</td>
</tr>
<tr>
<td>Sound sentences</td>
<td>91.67%</td>
<td>40.00%</td>
<td>53.33%</td>
<td>72.73%</td>
<td>68.18%</td>
<td>13.64%</td>
<td>61.11%</td>
<td>37.50%</td>
</tr>
</tbody>
</table>

*Note.* TYP = typical participants; WLD = participants with WLDs.
The first time that the samples were scored holistically, typical participants received a
total of 25 comments (56.0% positive) and participants with WLDs received 32 total comments
(18.8% positive). When the samples were scored holistically a second time, typical participants
and participants with WLDs were given 34 (44.1% positive) and 31 total comments (19.4%
positive), respectively.

When the total number of comments made in each theme was compared between groups
(Table 33), typical participants usually received more positive comments and fewer negative
comments than participants with WLDs across all themes. More specifically, typical participants
received more positive comments than participants with WLDs in the themes of Linguistic
(Holistic II), Content (Holistic I and II), and Quality (Holistic I and II). Conversely, participants
with WLDs received more negative comments than typical participants in Linguistic (Holistic I
and II), Mechanics (Holistic I), Content (Holistic I and II), and Quality (Holistic I and II) themes.

Table 33

Frequency of Positive and Negative Comments in Each Theme for Matched Pairs of Typical
Participants (n = 4) and Participants with WLDs (n = 4)

<table>
<thead>
<tr>
<th>Comment</th>
<th>Holistic I TYP</th>
<th>Holistic II TYP</th>
<th>Holistic I WLD</th>
<th>Holistic II WLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Mechanics</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Content</td>
<td>11</td>
<td>6</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Quality</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>6</td>
<td>15</td>
<td>25</td>
</tr>
</tbody>
</table>

*Note.* TYP = typical participants; WLD = participants with WLDs.

Comparisons of comments in the Linguistic subthemes revealed that typical participants
usually received more positive comments and fewer negative comments than participants with
WLDs (Table 34). More specifically, while there were no positive comments given to either
group during the first round of holistic scoring for any of the Linguistic subthemes, typical participants received more positive Wording and Clarity comments than participants with WLDs the second time samples were scored holistically. Additionally, participants with WLDs received more negative comments than typical participants for the Sentence Structure (Holistic I and II), Wording (Holistic I), Clarity (Holistic I and II), and Grammar/Syntax (Holistic I) subthemes.

Table 34

**Frequency of Positive and Negative Comments in Linguistic Subthemes for Matched Pairs of Typical Participants (n = 4) and Participants with WLDs (n = 4)**

<table>
<thead>
<tr>
<th>Linguistic Subthemes</th>
<th>Positive Comments</th>
<th>Negative Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holistic I</td>
<td>Holistic II</td>
</tr>
<tr>
<td></td>
<td>TYP</td>
<td>WLD</td>
</tr>
<tr>
<td>Sentence Structure</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Wording</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Clarity</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Grammar/Syntax</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note. TYP = typical participants; WLD = participants with WLDs.*

When comments in the Mechanics subthemes were compared (Table 35), minimal differences were seen between groups because few comments regarding Mechanics were given overall. In fact, no positive comments related to the subthemes Spelling or Punctuation were given to either group. The number of negative comments given in relation to Spelling and Punctuation was either equal or only differed by one comment between groups.

To compare Content subthemes between groups, the nine subthemes in this theme were collapsed into three subthemes because few comments were given in any of the nine subthemes (Table 36). More specifically, all of the subthemes relating to Presence/Absence, Support, and Strength were collapsed into three subthemes based upon the specific prompt component(s) they
referred to, including *All Prompt Components, Claim/Reasons, and Suggestions*. The comparison of comments in these condensed subthemes revealed that typical participants usually received a higher or equal number of positive comments compared to participants with WLDs, but that differences in negative comments were inconsistent.

Table 35

**Frequency of Positive and Negative Comments in Mechanics Subthemes for Matched Pairs of Typical Participants (n = 4) and Participants with WLDs (n = 4)**

<table>
<thead>
<tr>
<th>Mechanics Subthemes</th>
<th>Positive Comments</th>
<th>Negative Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holistic I TYP</td>
<td>Holistic II TYP</td>
</tr>
<tr>
<td>Spelling</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Punctuation</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note.* TYP = typical participants; WLD = participants with WLDs.

Table 36

**Frequency of Positive and Negative Comments in Content Subthemes for Matched Pairs of Typical Participants (n = 4) and Participants with WLDs (n = 4)**

<table>
<thead>
<tr>
<th>Content Subthemes</th>
<th>Positive Comments</th>
<th>Negative Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holistic I TYP</td>
<td>Holistic II TYP</td>
</tr>
<tr>
<td>All</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Claim/Reasons</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Suggestions</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>11</td>
<td>6</td>
</tr>
</tbody>
</table>

*Note.* TYP = typical participants; WLD = participants with WLDs.

When comments made in *Quality* subthemes were compared between groups, few comments were made in many of the *Quality* subthemes (Table 37). However, when comments were present, typical participants always received more positive and fewer negative comments than those with WLDs.
Table 37

*Frequency of Positive and Negative Comments in Quality Subthemes for Matched Pairs of Typical Participants (n = 4) and Participants with WLDs (n = 4)*

<table>
<thead>
<tr>
<th>Quality Subthemes</th>
<th>Positive Comments</th>
<th>Negative Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holistic I TYP</td>
<td>Holistic I WLD</td>
</tr>
<tr>
<td>Clarity</td>
<td>2 0</td>
<td>0 0</td>
</tr>
<tr>
<td>Overall Quality</td>
<td>0 0</td>
<td>1 0</td>
</tr>
<tr>
<td>Length</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>Organization</td>
<td>1 0</td>
<td>0 0</td>
</tr>
<tr>
<td>Specific Organization</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>Total</td>
<td>3 0</td>
<td>1 0</td>
</tr>
</tbody>
</table>

Note. TYP = typical participants; WLD = participants with WLDs.

**Question 3: Features Relating to Overall Writing Quality**

**Quantitative Results**

To determine which features of writing seemed to relate to overall quality, participants were divided into two groups – those receiving high holistic scores (i.e., scores of three or four) and those receiving low holistic scores (i.e., scores of one or two). Then means and standard deviations of analytic rating scale scores were calculated for participants with high and low holistic I scores. These results, shown in Table 38, indicate that participants with high holistic I scores obtained higher means than participants with low holistic I scores across all categories on the analytic rating scale.

Means and standard deviations were also calculated for the percentages calculated using word- and sentence-level codes given by raters. These results, shown in Table 39, revealed that participants with high holistic I scores produced a lower proportion of inappropriate sentences for the audience of a university president, spelling errors, ending punctuation errors, and other
punctuation errors than participants with low holistic I scores, as well as a higher proportion of complex sentences and sound sentences.

Table 38

Means and Standard Deviations of Analytic Rating Scale Scores for Participants with High (n = 33) and Low (n = 21) Holistic I Scores

<table>
<thead>
<tr>
<th>Category</th>
<th>High Holistic I</th>
<th>Low Holistic I</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Presence of Prompt Components</td>
<td>3.79</td>
<td>0.48</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>3.12</td>
<td>0.55</td>
</tr>
<tr>
<td>Sense of Audience</td>
<td>3.82</td>
<td>0.58</td>
</tr>
<tr>
<td>Organization</td>
<td>3.15</td>
<td>0.51</td>
</tr>
<tr>
<td>Word Choice</td>
<td>2.85</td>
<td>0.87</td>
</tr>
<tr>
<td>Spelling</td>
<td>3.06</td>
<td>0.66</td>
</tr>
<tr>
<td>Punctuation</td>
<td>2.36</td>
<td>1.17</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>2.30</td>
<td>0.85</td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td>2.09</td>
<td>1.10</td>
</tr>
<tr>
<td>Total analytic score</td>
<td>26.55</td>
<td>2.72</td>
</tr>
</tbody>
</table>

Means and standard deviations of analytic rating scale scores were calculated for participants with high and low holistic II scores. These results, shown in Table 40, indicate that participants with high holistic II scores obtained higher means than participants with low holistic II scores across all categories on the analytic rating scale except for Sentence Complexity.
Table 39

Means and Standard Deviations of Percentages of Coded Words and Sentences for Participants with High (n = 33) and Low (n = 21) Holistic I Scores

<table>
<thead>
<tr>
<th>Category</th>
<th>High Holistic I</th>
<th>Low Holistic I</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Inappropriate sentences</td>
<td>0.8%</td>
<td>2.7%</td>
</tr>
<tr>
<td>Spelling errors</td>
<td>0.6%</td>
<td>0.5%</td>
</tr>
<tr>
<td>Ending punctuation errors</td>
<td>0.8%</td>
<td>2.3%</td>
</tr>
<tr>
<td>Other punctuation errors</td>
<td>13.0%</td>
<td>11.9%</td>
</tr>
<tr>
<td>Complex sentences</td>
<td>56.8%</td>
<td>16.2%</td>
</tr>
<tr>
<td>Sound sentences</td>
<td>72.8%</td>
<td>15.0%</td>
</tr>
</tbody>
</table>

Means and standard deviations were also calculated for the percentages of word- and sentence-level features coded by raters. These results, shown in Table 41, revealed that participants with high holistic II scores produced a lower proportion of inappropriate sentences for the audience of a university president, spelling errors, ending punctuation errors, other punctuation errors, and complex sentences than participants with low holistic II scores, as well as higher proportion of sound sentences.
Table 40

*Means and Standard Deviations of Analytic Rating Scale Scores for Participants with High (n = 39) and Low (n = 15) Holistic II Scores*

<table>
<thead>
<tr>
<th>Category</th>
<th>High Holistic II</th>
<th>Low Holistic II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Presence of Prompt Components</td>
<td>3.74</td>
<td>0.50</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>3.08</td>
<td>0.58</td>
</tr>
<tr>
<td>Sense of Audience</td>
<td>3.79</td>
<td>0.61</td>
</tr>
<tr>
<td>Organization</td>
<td>3.10</td>
<td>0.50</td>
</tr>
<tr>
<td>Word Choice</td>
<td>2.79</td>
<td>0.83</td>
</tr>
<tr>
<td>Spelling</td>
<td>3.02</td>
<td>0.67</td>
</tr>
<tr>
<td>Punctuation</td>
<td>2.33</td>
<td>1.13</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>2.23</td>
<td>0.81</td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td>2.03</td>
<td>1.06</td>
</tr>
<tr>
<td>Total analytic score</td>
<td>26.13</td>
<td>2.90</td>
</tr>
</tbody>
</table>
Table 41

Means and Standard Deviations of Percentages of Coded Words and Sentences for Participants with High (n = 39) and Low (n = 15) Holistic II Scores

<table>
<thead>
<tr>
<th>Category</th>
<th>High Holistic II</th>
<th>Low Holistic II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Inappropriate sentences</td>
<td>1.0%</td>
<td>3.1%</td>
</tr>
<tr>
<td>Spelling errors</td>
<td>0.6%</td>
<td>0.5%</td>
</tr>
<tr>
<td>Ending punctuation errors</td>
<td>1.1%</td>
<td>2.7%</td>
</tr>
<tr>
<td>Other punctuation errors</td>
<td>12.4%</td>
<td>11.4%</td>
</tr>
<tr>
<td>Complex sentences</td>
<td>55.6%</td>
<td>14.8%</td>
</tr>
<tr>
<td>Sound sentences</td>
<td>71.6%</td>
<td>15.3%</td>
</tr>
</tbody>
</table>

Fisher’s exact test was then used to compare the proportion of participants within each group receiving high (i.e., scores of three or four) versus low (i.e., scores of one or two) low on each category of the analytic rating scale. This test was run twice – once grouping participants using holistic I scores and once using holistic II scores.

Results of Fisher’s exact test using holistic I scores (Table 42) revealed significant differences between groups in the categories of Logical Reasoning ($p < 0.01$), Organization ($p < 0.01$), and Spelling ($p = 0.02$). Results for Sentence Soundness approached significance ($p = 0.06$). In all cases, a higher proportion of participants with high holistic I scores received high scores on the analytic rating scale category and a lower portion received low scores on the analytic rating scale category than participants with low holistic I scores.
Table 42

*Prevalence of High and Low Analytic Rating Scale Scores in Participants with High (n = 33) and Low (n = 21) Holistic I Scores*

<table>
<thead>
<tr>
<th>Category</th>
<th>High Holistic I</th>
<th>Low Holistic I</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
</tr>
<tr>
<td>Presence of Prompt Components</td>
<td>32</td>
<td>97.0</td>
<td>1</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>30</td>
<td>90.9</td>
<td>3</td>
</tr>
<tr>
<td>Sense of Audience</td>
<td>30</td>
<td>90.9</td>
<td>3</td>
</tr>
<tr>
<td>Organization</td>
<td>31</td>
<td>93.9</td>
<td>2</td>
</tr>
<tr>
<td>Word Choice</td>
<td>24</td>
<td>72.7</td>
<td>9</td>
</tr>
<tr>
<td>Spelling</td>
<td>29</td>
<td>87.9</td>
<td>4</td>
</tr>
<tr>
<td>Punctuation</td>
<td>12</td>
<td>36.4</td>
<td>21</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>14</td>
<td>42.4</td>
<td>19</td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td>11</td>
<td>33.1</td>
<td>22</td>
</tr>
</tbody>
</table>

The results of the Fisher’s exact tests comparing participants with high and low holistic II scores (Table 43) revealed significant differences in *Logical Reasoning* ($p < 0.01$), *Organization* ($p < 0.01$), and *Spelling* ($p = 0.03$), while results for *Sentence Soundness* approached significance ($p = 0.08$). For all of these results, the participants with high holistic II scores had a higher proportion of high scores and a lower proportion of low scores than the participants with low holistic II scores.
Table 43

Prevalence of High and Low Analytic Rating Scale Scores in Participants with High (n = 39) and Low (n = 15) Holistic II Scores

<table>
<thead>
<tr>
<th>Category</th>
<th>High Holistic II</th>
<th>Low Holistic II</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
</tr>
<tr>
<td>Presence of Prompt Components</td>
<td>38</td>
<td>97.4</td>
<td>1</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>34</td>
<td>87.2</td>
<td>5</td>
</tr>
<tr>
<td>Sense of Audience</td>
<td>35</td>
<td>89.7</td>
<td>4</td>
</tr>
<tr>
<td>Organization</td>
<td>36</td>
<td>92.3</td>
<td>3</td>
</tr>
<tr>
<td>Word Choice</td>
<td>27</td>
<td>69.2</td>
<td>12</td>
</tr>
<tr>
<td>Spelling</td>
<td>33</td>
<td>84.6</td>
<td>6</td>
</tr>
<tr>
<td>Punctuation</td>
<td>13</td>
<td>33.3</td>
<td>26</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>16</td>
<td>41.0</td>
<td>23</td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td>12</td>
<td>30.8</td>
<td>27</td>
</tr>
</tbody>
</table>

In addition to Fisher’s exact tests, t-tests were run to compare the performance of participants with high versus low holistic scores on the features of writing calculated using percentages of total words or sentences (i.e., inappropriate sentences, spelling errors, ending punctuation errors, other punctuation errors, complex sentences, and sound sentences), as well as total analytic rating scale scores (Table 44). The results of t-tests comparing participants with high versus low holistic I scores revealed significant differences between groups for spelling errors \( t(52) = 2.56, p = 0.02 \), sound sentences \( t(52) = -3.81, p < 0.01 \), and total analytic scores \( t(52) = -4.87, p < 0.01 \). An analysis of group means revealed that participants with high holistic I scores had fewer spelling errors, more sound sentences, and higher total analytic rating
scale scores than participants with low holistic I scores. Using Cohen’s (1988) recommendations for interpreting effect size, large effect sizes were seen for all three of these measures.

Table 44

*Group Differences for Word- and Sentence-Level Percentages and Total Analytic Rating Scale Scores Between Participants with High and Low Holistic I Scores*

<table>
<thead>
<tr>
<th>Category</th>
<th>High holistic I scores</th>
<th>Low holistic I scores</th>
<th>t(52)</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>t</td>
</tr>
<tr>
<td>Inappropriate sentences</td>
<td>0.85</td>
<td>2.75</td>
<td>1.44</td>
<td>3.24</td>
<td>0.72</td>
</tr>
<tr>
<td>Spelling errors</td>
<td>0.58</td>
<td>0.52</td>
<td>1.27</td>
<td>1.14</td>
<td>2.56</td>
</tr>
<tr>
<td>Ending punctuation errors</td>
<td>0.81</td>
<td>2.26</td>
<td>1.64</td>
<td>3.58</td>
<td>0.96</td>
</tr>
<tr>
<td>Other punctuation errors</td>
<td>13.03</td>
<td>11.87</td>
<td>10.04</td>
<td>16.11</td>
<td>1.58</td>
</tr>
<tr>
<td>Complex sentences</td>
<td>56.83</td>
<td>16.16</td>
<td>54.55</td>
<td>14.77</td>
<td>-0.52</td>
</tr>
<tr>
<td>Sound sentences</td>
<td>72.85</td>
<td>14.99</td>
<td>52.25</td>
<td>21.69</td>
<td>-3.81</td>
</tr>
<tr>
<td>Total analytic score</td>
<td>26.55</td>
<td>2.72</td>
<td>22.52</td>
<td>3.31</td>
<td>-4.87</td>
</tr>
</tbody>
</table>

*Note.* To interpret Cohen’s $d$, values of 0.20 to 0.49 are small effect sizes, values of 0.50 to 0.79 are medium effect sizes, and values of $>0.80$ are large effect sizes.

T-test results comparing participants with high versus low holistic II scores, shown in Table 45, revealed significant differences between groups in other punctuation errors [$t(52) = 2.66, p = 0.01$], sound sentences [$t(52) = -3.98, p < 0.01$], and total analytic scores [$t(52) = -4.48, p < 0.01$]. Additionally, group differences for spelling errors approached significance. A comparison of group means revealed that participants with high holistic II scores had fewer spelling errors, fewer other punctuation errors, more sound sentences, and higher total analytic rating scale scores than participants with low holistic two scores. Using Cohen’s (1988)
recommending for interpreting effect sizes, large effect sizes were seen for other punctuation errors, sentence soundness, and total analytic scores.

Table 45

*Group Differences for Word- and Sentence-Level Percentages and Total Analytic Rating Scale Scores Between Participants with High and Low Holistic II Scores*

<table>
<thead>
<tr>
<th>Category</th>
<th>High holistic II scores</th>
<th>Low holistic II scores</th>
<th>t(52)</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inappropriate sentences</td>
<td>1.07 ± 3.09</td>
<td>1.23 ± 2.57</td>
<td>0.23</td>
<td>0.82</td>
<td>0.05</td>
</tr>
<tr>
<td>Spelling errors</td>
<td>0.64 ± 0.52</td>
<td>1.37 ± 1.32</td>
<td>2.07</td>
<td>0.06</td>
<td>0.89</td>
</tr>
<tr>
<td>Ending punctuation errors</td>
<td>1.12 ± 2.75</td>
<td>1.16 ± 3.19</td>
<td>0.04</td>
<td>0.97</td>
<td>0.01</td>
</tr>
<tr>
<td>Other punctuation errors</td>
<td>12.42 ± 11.37</td>
<td>23.03 ± 16.96</td>
<td>2.66</td>
<td>0.01</td>
<td>0.81</td>
</tr>
<tr>
<td>Complex sentences</td>
<td>55.57 ± 14.78</td>
<td>56.91 ± 17.87</td>
<td>0.28</td>
<td>0.78</td>
<td>0.09</td>
</tr>
<tr>
<td>Sound sentences</td>
<td>71.64 ± 15.26</td>
<td>47.16 ± 21.87</td>
<td>-3.98</td>
<td>&lt;0.01</td>
<td>1.42</td>
</tr>
<tr>
<td>Total analytic score</td>
<td>26.13 ± 2.90</td>
<td>22.00 ± 3.36</td>
<td>-4.48</td>
<td>&lt;0.01</td>
<td>1.36</td>
</tr>
</tbody>
</table>

*Note.* To interpret Cohen’s d, values of 0.20 to 0.49 are small effect sizes, values of 0.50 to 0.79 are medium effect sizes, and values of ≥0.80 are large effect sizes.

**Qualitative Results**

**All participants.** To qualitatively analyze the differences between the writing samples of participants with high and low holistic scores, comparisons were made between the types of comments given to each group of participants. When all of the comments were considered, 89.2% of the comments given to participants with high holistic scores and 0% of the comments given to participants with low holistic scores were positive in nature during the first holistic scoring procedure. When the samples were scored holistically a second time, 94.3% of the
comments given to participants with high holistic scores and 9.5% of the comments given to participants with low holistic scores were positive in nature.

To more closely analyze the similarities and differences in the comments given to the two participant groups, the average number of positive and negative comments in each theme and subtheme was also calculated for each group (Table 46). A comparison of the average number of positive and negative themes during the first and second holistic scoring procedures revealed that participants with high holistic scores received a higher average number of positive comments and a lower average number of negative comments across all themes than participants with low holistic scores. The only exception to this pattern was the fact that both groups received an equal average number of positive comments for Mechanics during both the first and second holistic scoring procedures.

When the subthemes were compared, participants with high holistic scores typically received a higher or equal average number of positive comments than participants with low holistic scores across subthemes. More specifically, during the first phase of holistic scoring, participants with high holistic scores received a higher average number of positive comments than participants with low holistic scores for one of four Linguistic subthemes, seven of nine Content subthemes, and five of five Quality subthemes, as well as an equal average number of positive comments for three of four Linguistic subthemes, two of two Mechanics subthemes, and two of nine Content subthemes. During the second phase of holistic scoring, participants with high holistic scores received a higher average number of positive comments than participants with low holistic scores for three of four Linguistic subthemes, five of nine Content subthemes, and five of five Quality subthemes, as well as an equal average number of positive comments as
participants with low holistic scores for one of four *Linguistic* subthemes, two of two *Mechanics* subthemes, and three of nine *Content* subthemes.

Table 46

**Average Number of Comments per Theme and Subtheme for Participants with High And Low Holistic Scores**

<table>
<thead>
<tr>
<th>Comment Type</th>
<th>Positive Comments</th>
<th>Negative Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Low&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>Linguistic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Structure</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Wording</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Clarity</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Grammar/Syntax</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Total Linguistic</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Mechanics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spelling</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Punctuation</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Total Mechanics</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Content</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presence All</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Support All</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Strength All</td>
<td>0.3</td>
<td>0.0</td>
</tr>
<tr>
<td>Presence Claim/Reasons</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>Support Claim/Reasons</td>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Strength Claim/Reasons</td>
<td>0.3</td>
<td>0.0</td>
</tr>
<tr>
<td>Presence Suggestions</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>Support Suggestions</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Strength Suggestions</td>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Total Content</td>
<td>2.1</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Quality</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clarity</td>
<td>0.8</td>
<td>0.0</td>
</tr>
<tr>
<td>Overall Quality</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>Length</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Organization</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>Specific Organization</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>Total Quality</td>
<td>1.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Total All Themes</td>
<td>3.7</td>
<td>0.0</td>
</tr>
</tbody>
</table>

*Note.* High = participants with holistic score of 4; Low = participants with holistic score of 1.

<sup>a</sup> n = 14.  <sup>b</sup> n = 2.  <sup>c</sup> n = 7.  <sup>d</sup> n = 2
Conversely, participants with low holistic scores typically received a higher or equal average number of negative comments compared to participants with high holistic scores. More specifically, during the first phase of holistic scoring, participants with low holistic scores received a higher average number of negative comments than participants with high holistic scores for two of four Linguistic subthemes, one of two Mechanics subthemes, and three of nine Content subthemes, as well as an equal average number of negative comments as participants with high holistic scores for two of four Linguistic subthemes, one of two Mechanics subthemes, five of nine Content subthemes, and five of five Quality subthemes. During the second phase of holistic scoring, participants with low holistic scores received a higher average number of negative comments than participants with high holistic scores for two of four Linguistic subthemes, two of two Mechanics subthemes, four of nine Content subthemes, and three of five Quality subthemes, as well as an equal average number of negative comments as participants with high holistic scores for two of four Linguistic subthemes, five of nine Content subthemes, and two of five Quality subthemes.

The final qualitative analysis comparing all participants with high and low holistic scores focused on comparing the coded elements within the writing samples of these groups (i.e., inappropriate sentences, misspelled words, ending punctuation errors, and other punctuation errors). Sentences coded as inappropriate for the Sense of Audience rating scale category were not compared between groups because none of participants with low holistic scores had any inappropriate sentences in their samples. Additionally, complex sentences and sound sentences were not compared because it was not always clear exactly why raters had coded sentences as complex or sound. However, spelling and punctuation errors were compared because it was
clear how words were misspelled and what types of punctuation errors were made, as the raters
typically corrected punctuation errors in the samples.

When spelling errors types were compared, it appeared that the two groups made
different types of errors (Table 47). First of all, participants with high holistic scores made
primarily Orthographic errors, while participants with low holistic scores made no errors that
were purely Orthographic in nature. Conversely, participants with low holistic scores made
primarily Phonological errors. Participants with high holistic score made Phonological errors,
but a substantially smaller proportion than participants with low holistic scores. Participants
with low holistic scores also made a higher proportion of Apostrophe and Homophone errors
than participants with high holistic scores. Differences in other types of spelling errors were
relatively small between participant groups.

Table 47

<table>
<thead>
<tr>
<th>Spelling Error</th>
<th>Holistic I High</th>
<th>Holistic I Low</th>
<th>Holistic II High</th>
<th>Holistic II Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apostrophe</td>
<td>5.9%</td>
<td>25.0%</td>
<td>0.0%</td>
<td>25.0%</td>
</tr>
<tr>
<td>Apostrophe/Morphological</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Correct</td>
<td>2.9%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Homophone</td>
<td>5.9%</td>
<td>25.0%</td>
<td>9.1%</td>
<td>25.0%</td>
</tr>
<tr>
<td>Orthographic</td>
<td>64.7%</td>
<td>0.0%</td>
<td>72.7%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Phonological</td>
<td>20.6%</td>
<td>50.0%</td>
<td>9.1%</td>
<td>50.0%</td>
</tr>
<tr>
<td>Phonological/Morphological</td>
<td>0.0%</td>
<td>0.0%</td>
<td>9.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Phonological/Orthographic</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Phonological/Wrong Word</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Note. High = participants with holistic score of 4; Low = participants with holistic score of 1.

When punctuation errors were compared, there was much variability in the frequency of
punctuation errors for both groups. Both groups had participants who made ending punctuation
errors. However, the ending punctuation errors made by participants with high holistic scores were due to the incorrect use of ending punctuation, while the errors made by participants with low holistic scores were due to run-on sentences. The frequency of other punctuation errors ranged from 0% to 38.5% of sentences for participants with high holistic scores and 0% to 50.0% for participants with low holistic scores. In both groups, other punctuation errors primarily stemmed from missing or inappropriately placed commas. Additionally, each group had one participant who received one other punctuation error code for a run-on sentence. Overall, both groups tended to make similar other punctuation errors, but the participants with low holistic scores tended to use more run-on sentences in their writing than participants with high holistic scores.

**Matched pairs.** To further analyze the differences between the writing samples of participants with high versus low holistic scores, three participants with high holistic scores and three participants with low holistic scores were matched as well as possible for gender, age, year in school, and *PPVT-4* standard scores. The performance of these matched pairs on the rating scales and features measured by deriving percentages of total words or sentences (i.e., inappropriate sentences, misspelled words, ending punctuation errors, other punctuation errors, complex sentences, and sound sentences) was compared to examine similarities and differences between these groups (Tables 48 and 49).

A comparison of scores on the analytic rating scale revealed that participants with high holistic scores consistently received higher scores than participants with low holistic scores for *Presence of Prompt Components* and total analytic rating scale scores. For *Logical Reasoning, Organization, Punctuation,* and *Sentence Complexity,* participants with high holistic scores received higher scores than participants with low holistic scores in two of the three matched
pairs. Scores for participants with high holistic scores were equal to scores for participants with low holistic scores for Sense of Audience, Word Choice, Spelling, and Sentence Soundness in at least two of the three matched pairs.

When the percentages of word- and sentence-level features of writing were compared between groups, it was revealed that participants with high holistic scores always had a lower proportion of misspelled words and a higher proportion of sound sentences. For other punctuation errors and complex sentences, participants with high holistic scores produced fewer other punctuation errors and more complex sentences than participants with low holistic scores in two of the three matched pairs. Participants with high holistic scores and low holistic scores produced an equal proportion of ending punctuation errors and inappropriate sentences in two and three of the three matched pairs, respectively.

The types of comments given to the writing samples of the matched pairs of participants during the holistic scoring procedure were also compared. Because few comments were given to each participant, comments for all participants in each group were combined during this analysis. Comments were provided by all three raters for one participant in the low holistic score group, but all other participants in both groups received comments from only one rater. Therefore, comments from only one rater were selected for the participants with comments from three raters so that the number of comments given to each participant would be fairly balanced.

The first time that the samples were scored holistically, participants with high holistic scores received a total of 10 comments (100% positive) and participants with low holistic scores received 13 total comments (7.7% positive). When the samples were scored holistically a second time, participants with high and low holistic scores were given 13 (69.2% positive) and 17 total comments (11.8% positive) respectively.
Table 48

Comparison of Performance on Holistic and Analytic Rating Scales for Matched Pairs of High and Low Holistic Scores

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Pair 1</th>
<th></th>
<th></th>
<th>Pair 2</th>
<th></th>
<th></th>
<th>Pair 3</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>Holistic I</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Holistic II</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presence of Prompt Components</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sense of Audience</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organization</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word Choice</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spelling</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Punctuation</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total analytic score</td>
<td>36</td>
<td>27</td>
<td>30</td>
<td>20</td>
<td>34</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 49

Comparison of Percentages for Matched Pairs of High and Low Holistic Scores

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Pair 1</th>
<th></th>
<th></th>
<th>Pair 2</th>
<th></th>
<th></th>
<th>Pair 3</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>Inappropriate sentences</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td></td>
</tr>
<tr>
<td>Misspelled words</td>
<td>0.5%</td>
<td>0.6%</td>
<td>2.6%</td>
<td>4.7%</td>
<td>0.0%</td>
<td>0.4%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ending punctuation errors</td>
<td>0.0%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>11.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other punctuation errors</td>
<td>22.2%</td>
<td>0.0%</td>
<td>17.4%</td>
<td>22.2%</td>
<td>0.0%</td>
<td>50.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complex sentences</td>
<td>77.8%</td>
<td>35.7%</td>
<td>60.9%</td>
<td>66.7%</td>
<td>73.3%</td>
<td>50.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sound sentences</td>
<td>83.3%</td>
<td>64.3%</td>
<td>65.2%</td>
<td>22.2%</td>
<td>60.0%</td>
<td>37.5%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
When the total number of comments made in each theme were compared between groups (Table 50), participants with high holistic scores usually received more positive comments and fewer negative comments than participants with low holistic scores. More specifically, participants with high holistic scores received more positive comments than participants with low holistic scores in the themes of Linguistic (Holistic II), Content (Holistic I and II), and Quality (Holistic I and II). Conversely, participants with low holistic scores received more negative comments than participants with high holistic scores in Linguistic (Holistic I), Mechanics (Holistic I and II), Content (Holistic I and II), and Quality (Holistic II) themes.

Table 50

<table>
<thead>
<tr>
<th>Comment</th>
<th>Positive Comments</th>
<th>Negative Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holistic I</td>
<td>Holistic II</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Linguistic</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mechanics</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Content</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Quality</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

*Note. High = participants with holistic score of 4; Low = participants with holistic score of 1.*

Comparisons of comments in the Linguistic subthemes revealed that participants with high holistic scores usually received an equal or higher number of positive comments and an equal or lower number of negative comments than participants with low holistic scores (Table 51). More specifically, while there were no positive comments given to either group during the first round of holistic scoring for any of the Linguistic subthemes, participants with high holistic scores received more positive Clarity comments than participants with low holistic scores the second time samples were scored holistically. Additionally, participants with low holistic scores
received more negative comments than participants with high holistic scores in the subthemes of Sentence Structure (Holistic I), Clarity (Holistic I), and Grammar/Syntax (Holistic I).

Table 51

Frequency of Positive and Negative Comments in Linguistic Subthemes for Matched Pairs of Participants with High (n = 3) and Low Holistic Scores (n = 3)

<table>
<thead>
<tr>
<th>Linguistic Subthemes</th>
<th>Positive Comments</th>
<th>Negative Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holistic I High</td>
<td>Holistic I Low</td>
</tr>
<tr>
<td>Sentence Structure</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Wording</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Clarity</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Grammar/Syntax</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. High = participants with holistic score of 4; Low = participants with holistic score of 1.

When comments in the Mechanics subthemes were compared (Table 52), minimal differences were seen between groups because few comments regarding Mechanics were given overall. In fact, no positive comments related to the subthemes of Spelling or Punctuation were given to either group. The number of negative comments given in relation to Spelling and Punctuation was either equal or only differed by one comment between groups.

Table 52

Frequency of Positive and Negative Comments in Mechanics Subthemes for Matched Pairs of Participants with High (n = 3) and Low Holistic Scores (n = 3)

<table>
<thead>
<tr>
<th>Mechanics Subthemes</th>
<th>Positive Comments</th>
<th>Negative Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holistic I High</td>
<td>Holistic I Low</td>
</tr>
<tr>
<td>Spelling</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Punctuation</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. High = participants with holistic score of 4; Low = participants with holistic score of 1.
To compare Content subthemes between groups, the nine subthemes under this theme were collapsed into three subthemes because few comments were given in each of the nine subthemes (Table 53). All of the subthemes relating to Presence/Absence, Support, and Strength were collapsed into three subthemes based upon the specific prompt component(s) they referred to, including All Prompt Components, Claim/Reasons, and Suggestions. The comparison of comments in these condensed subthemes revealed that participants with high holistic scores usually received a higher number of positive comments and a lower number of negative comments compared to participants with low holistic scores.

Table 53

*Frequency of Positive and Negative Comments in Content Subthemes for Matched Pairs of Participants with High (n = 3) and Low Holistic Scores (n = 3)*

<table>
<thead>
<tr>
<th>Content Subthemes</th>
<th>Positive Comments</th>
<th>Negative Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holistic I</td>
<td>Holistic II</td>
</tr>
<tr>
<td>All</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Claim/Reasons</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Suggestions</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

*Note.* High = participants with holistic score of 4; Low = participants with holistic score of 1.

When comments made in Quality subthemes were compared between groups, few comments were made in many of the Quality subthemes (Table 54). However, when comments were present, participants with high holistic scores always received more positive and fewer negative comments than those with low holistic scores.
Table 54

*Frequency of Positive and Negative Comments in Quality Subthemes for Matched Pairs of Participants with High (n = 3) and Low Holistic Scores (n = 3)*

<table>
<thead>
<tr>
<th>Content Subthemes</th>
<th>Positive Comments</th>
<th></th>
<th></th>
<th>Negative Comments</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holistic I High</td>
<td>Low</td>
<td>Holistic II High</td>
<td>Low</td>
<td>Holistic I High</td>
<td>Low</td>
</tr>
<tr>
<td>Clarity</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Overall Quality</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Length</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Organization</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Specific Organization</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note.* High = participants with holistic score of 4; Low = participants with holistic score of 1.
CHAPTER V

DISCUSSION

Summary of Major Results

The primary purpose of this study was to examine the persuasive writing abilities of undergraduate students. To examine this phenomenon, the current study focused on analyzing the reliability of rating scales used to measure persuasive writing in undergraduates, similarities and differences between the persuasive writing abilities of typical undergraduates and undergraduates with WLDs, and the features of persuasive writing that related most to overall writing quality.

To look at these areas, both quantitative and qualitative methods were used. Quantitative methods involved scoring the persuasive writing samples of undergraduate students using holistic and analytic rating scales. Comparisons of these scores were then made within and among raters, as well as between different groups of participants (i.e., those with and without WLDs; those with high and low holistic scores). The qualitative methods used in this study involved collecting comments from raters about their reasons for assigning each sample a particular holistic score. These comments were coded and grouped into themes and subthemes. Comparisons of these themes were made between and among raters, as well as between different participant groups (i.e., those with and without WLDs; those with high and low holistic scores). Additionally, some of the coded features of the writing samples were analyzed qualitatively (i.e., spelling and punctuation errors) and compared between participants with and without WLDs, as well as participants with high and low holistic scores.

The following sections will discuss the results from the analyses used to answer each of the three research questions. This discussion will include what the results suggest about the
reliability of rating scales, the similarities and differences between undergraduates with and without WLDs, and the features of writing that seem to relate most to overall writing quality. Additionally, the implications that these findings have for the assessment and treatment of undergraduates with WLDs will be discussed. Limitations of the current study and implications for future research will also be presented.

**Reliability of Holistic and Analytic Rating Scales**

One of the purposes of this study was to determine how reliably rating scales could be used to examine the writing skills of undergraduate students. To answer this question, both inter-rater and intra-rater reliability were calculated for the holistic and analytic rating scales.

**Holistic rating scale.** For the holistic rating scale, inter-rater reliability was low before the in-depth training about how to analyze specific features of persuasive writing, as absolute inter-rater reliability ranged from 42% to 58% within the three pairs of raters. This finding was not surprising, as the raters were only given one-word descriptions of each of the point values they used to holistically score the samples (i.e., 4 = excellent, 3 = good, 2 = fair, and 1 = poor). With limited guidance about how to determine which point value should be awarded to each sample, it was expected that there were be some variability among the methods used to award holistic scores by raters. Therefore, it was anticipated that there would be low inter-rater reliability for the holistic scores collected before the in-depth training.

Qualitative results regarding inter-rater reliability before the in-depth training revealed that all of the raters consistently used features of *Content* to award holistic scores, but that there was little agreement among raters in regard to the use of *Linguistic, Mechanics,* and *Quality* features while awarding holistic scores. Again, this was not surprising, as the raters had been given so little guidance about what to look for when determining which holistic score to award
for each sample. The fact that they primarily relied on features of *Content* made sense, as the raters were required to read the persuasive writing prompt before scoring samples in order to get an idea of what participants were expected to include in their writing samples.

Following the in-depth training, absolute inter-rater reliability for holistic scores increased slightly (i.e., from 42% – 58% to 50% – 75% agreement). Given the intensity of the training and the level of detail given in the instructions for how to analyze the specific features of writing, it was anticipated that the second round of holistic ratings would reflect a notable increase in inter-rater reliability. However, the training did not result in satisfactory levels of inter-rater reliability (i.e., 80% or higher).

Based on the results of the qualitative analysis, it appeared that raters continued to primarily focus on *Content* features while scoring samples. This was not surprising, as they continued to have little guidance during this task and were required to review the writing prompt that contained a detailed outline of expected content before scoring. While there was an increase in the consideration of *Linguistic*, *Mechanics*, and *Quality* features, there was still limited consistency in the consideration of these features when determining holistic scores for samples. Therefore, it appears that in addition to the in-depth training, more detailed descriptions of each of the point values would be necessary to increase inter-rater reliability on the holistic rating scale. More specifically, if each point value contained detailed descriptions about the features of writing that should have been considered and what they should have looked like at each level, raters would have to focus on the same features of writing as they provided holistic scores.

Although the training did not result in a desirable level of change for inter-rater reliability for the holistic rating scale, it did produce a satisfactory level of intra-rater reliability. Before training, absolute intra-rater reliability ranged from 33% to 50%. As with inter-rater reliability,
these results were not surprising because limited guidance was provided about how to scores the samples and the raters had little prior experience scoring writing samples. However, following the in-depth training, absolute intra-rater reliability increased to 83% to 100%.

The qualitative analysis for intra-rater reliability showed that raters attended to a wider range of features of writing when providing holistic scores following the in-depth training. For example, Rater 2’s provision of Linguistic and Mechanics comments increased from pre- to post-training, while her use of Content and Quality comments remained relatively the same. Therefore, she was providing a wider range of comments for samples following training than prior to training. However, the raters’ consistency in using these features did not substantially change from pre- to post-training, as levels of intra-rater agreement from pre- to post-training were relatively similar. Therefore, while the in-depth training generally resulted in each rater considering a wider range of features while providing holistic scores, it did not result in increased consistency in the use of comment themes for any of the raters. Therefore, it is likely that detailed descriptions of each point value would be needed to increase intra-rater consistency in terms of the features of writing used to award holistic scores, as information about the types and quality of features of writing would force each rater to consider the same features every time they provided a holistic score.

**Analytic rating scale.** In regard to the analytic rating scale, inter-rater reliability reached acceptable levels (i.e., 80% or higher) in the areas of Sense of Audience, Ending Punctuation, Other Punctuation, Sentence Complexity, and Sentence Soundness. All of these categories were coded sentence by sentence. Additionally, these categories were either objective in nature (i.e., Ending Punctuation, Other Punctuation, and Sentence Complexity) or were subjectively coded using fewer options than the other subjectively scored categories on the analytic rating scale (i.e.,
Sense of Audience and Sentence Soundness. More specifically, when coding for Sense of Audience, raters only had two choices – appropriate or inappropriate. For Sentence Soundness, a three-point scale was used rather than a four-point scale (i.e., it had been previously found that adequate levels of inter-rater reliability could not be achieved with a four-point scale when expert clinicians coded samples). Therefore, it appears that agreement among raters was best for objective measures or for subjective measures that had only two or three options (rather than four). This was not surprising, as researchers have reported that reliability tends to increase when scales provide concrete or objective descriptions of the levels for each criteria and when fewer point values are available for each criteria (e.g., Knoch, 2011). Furthermore, it is often easier for raters to use scales with an odd number of point values, as they can default to the middle value rather than having to pick a value that is closer to one extreme or another (e.g., Knoch, 2007). However, when raters default to the midpoint value, valuable information is lost because it becomes more difficult to identify writers’ strengths and weaknesses when their scores fall at the midpoint of a scale. Another reason why inter-rater reliability for the objective measures was likely higher than the subjective measures was due to the way in which inter-rater reliability was calculated. More specifically, for all of the features of writing that reached acceptable levels of inter-rater reliability, inter-rater reliability was calculated based on the total number of sentences per sample, rather than a single score per sample. This likely inflated inter-rater reliability, as there were 202 opportunities for agreement with these measures and only 12 opportunities for agreement with all of the other measures except Spelling.

Conversely, acceptable levels of inter-rater reliability were not met for Presence of Prompt Components, Logical Reasoning, Organization, Word Choice, and Spelling. Most of these categories were scored subjectively using the four-point analytic rating scale, meaning that
raters were unable to default to a midpoint value while scoring. Additionally, due to the subjective nature of most of these categories, raters were unable to count features of writing to make a determination about which point value to award a sample. The only exception (besides Spelling) was Presence of Prompt Components, as raters were required to count the number of prompt components that were present to score this category on the rating scale. However, raters still had some difficulty with this category, as they were often unsure of whether or not a feature was explicit enough to be considered fully present. Therefore, it seems that the in-depth training and descriptions given for how to score each of these categories were not enough to help multiple raters award consistent scores across samples in these areas. However, Spelling was an objective measure that required raters to code individually misspelled words in samples. It is likely that this category had low inter-rater reliability because the raters submitted their counts of misspelled words after reading through the samples only once. While resolving discrepancies in scores, all of the raters reported that they continued to find spelling errors as they coded other categories, but did not code these words or report them to the primary investigator because they had already submitted their data sheets for spelling errors. Had the raters been encouraged to continue coding and reporting spelling errors throughout the scoring process, it is likely that inter-rater reliability for Spelling would have been higher.

Although there was high agreement among raters for the coding of inappropriate sentences (i.e., Sense of Audience), ending punctuation errors (i.e., Punctuation), other punctuation errors (i.e., Punctuation), complex sentences (i.e., Sentence Complexity), and sound sentences (i.e., Sentence Soundness), there was enough disagreement to result in substantial differences when these percentages were converted to analytic rating scale scores. More specifically, inter-rater reliability was low when scores on the analytic rating scale were derived
from the percentage of coded sentences. In fact, acceptable levels of inter-rater reliability were not reached by all three rater pairs for any of these categories when converted to scores on the analytic rating scale. Therefore, even slight differences in the coding of sentences resulted in different scores on the analytic rating scale.

Also examined were differences in reliability across the three iterations of the analytic rating scale. Recall, the first scale was a four-point scale that resulted in low reliability scores and actually prompted the development of a three-point scale in its place. While this second iteration (i.e., the three-point scale) yielded higher reliability calculations, there was concern that raters were defaulting to the middle scores. When inter-rater reliability for the current study was compared to inter-rater reliability measures from previous iterations of the analytic rating scale, several improvements were seen. Inter-rater reliability improved in six categories from the original four-point scale to the current scale (i.e., Sense of Audience, Spelling, Ending Punctuation, Other Punctuation, Sentence Complexity, and Sentence Soundness), remained relatively the same in three of the categories (i.e., Presence of Prompt Components, Organization, and Word Choice), and decreased slightly in one category (i.e., Spelling). Compared to the three-point scale (i.e., the second iteration of the analytic rating scale), inter-rater reliability on the current scale increased in five categories (i.e., Presence of Prompt Components, Sense of Audience, Spelling, Other Punctuation, and Ending Punctuation), remained relatively the same for one category (i.e., Sentence Soundness), and decreased slightly for four categories (i.e., Logical Reasoning, Organization, Word Choice, and Sentence Complexity). It was expected that improvements in inter-rater reliability would be seen in relation to the original four-point rating scale, as many revisions had been made to the descriptions of each point value in response to issues with inter-rater reliability. However, it was
expected that inter-rater reliability would decrease or remain the same in most categories with the shift from a three-point to a four-point scale, as reliability tends to decrease when more point values are present (Knoch, 2011). The fact that many improvements in inter-rater reliability were seen from both of the previous two iterations of the analytic rating scale shows promise that reliability can improve when changes are made to the training of raters and the rating scale itself.

When intra-rater reliability was examined for the analytic rating scale used in the current study, it was found that levels of intra-rater reliability were higher than or equal to inter-rater reliability level across nearly all raters and all categories. More specifically, acceptable levels (i.e., 80% or higher) of intra-rater reliability were achieved in the areas of Presence of Prompt Components, Logical Reasoning, Sense of Audience, Ending Punctuation, Other Punctuation, Sentence Complexity, and Sentence Soundness. For Organization and Word Choice, one of the three raters achieved acceptable levels of intra-rater reliability. The only area that was below an acceptable level of intra-rater reliability for all raters was Spelling. As stated previously, low levels of reliability in Spelling were likely due to the fact that the raters submitted counts of spelling errors after one reading of each sample rather than continuing to code and report spelling errors as they read the samples multiple times. Overall, even though scores did not always agree among raters, it appeared that each of the raters was fairly consistent, internally, in awarding scores using the analytic rating scale.

Summary of findings. The findings for inter-rater and intra-rater reliability on the holistic and analytic rating scales indicate that improvements to the scales are necessary to achieve adequate levels of inter-rater and intra-rater agreement. However, the findings were consistent with other research, as higher levels of inter-rater and intra-rater reliability were typically found on the analytic rating scale than the holistic rating scale (see Weigle, 2002).
Furthermore, the fact that raters tended to achieve higher levels of reliability with the objective measures on the rating scale was consistent with Knoch (2011), who reported that rating scales with more objective criteria tend to result in higher levels of reliability than those with more subjective descriptors. Overall, the analytic rating scale appears to have more potential for high levels of reliability than the holistic rating scale.

**Clinical and research implications.** The findings of the current study confirm the fact that there are inherent problems with the reliability of rating scales used to assess writing. The analytic rating scale used in this study was carefully constructed, with much consideration given to the features of writing included on the scale and how they were measured. Additionally, prior to the current study, the scale had been tested several times using writing samples from college students. The issues that arose during these processes resulted in revisions to the rating scale to improve reliability. Yet, despite all of the time and consideration that went into the development of the analytic rating scale, problems with both inter-rater and intra-rater reliability remained. Other researchers have also reported difficulty establishing adequate levels of reliability when using rating scales (e.g., Koutsoftas & Gray, 2012). However, as mentioned previously, the results of the current study show that gains in reliability can be made when appropriate changes are made to the rating scales and training procedures. Therefore, while it may be difficult to achieve acceptable levels of reliability, it seems possible that continued adjustments in these areas could result in a more reliable rating scale that could be used to assess the writing skills of college students who may present with writing difficulties.

Additionally, other writing assessments using rating scales have reported high levels of reliability. More specifically, the writing tests of the *ACT*, *SAT*, and *GRE* all use holistic rating scales to score the writing samples produced by examinees and have reported high levels of
inter-rater reliability (ACT, 2009; College Board, 2012; Schaeffer, Briel, & Fowles, 2001). The holistic rating scales used in these measures provide detailed descriptions of the specific features of writing expected to be seen at each point value, while the holistic rating scale used in the current study only used one-word descriptions of each point value. Therefore, it is possible that providing descriptions of each point value on the holistic rating scale used in this study could improve inter-rater reliability. However, even with high levels of reliability, holistic rating scales provided limited information about the specific strengths and weaknesses of the writing samples they assess. This could make it difficult for SLPs to develop individualized treatment goals and procedures for their clients with writing difficulties solely using information from a holistic rating scale.

Although low levels of inter-rater reliability were seen for the subjective holistic rating scale and the subjective categories of the analytic rating scale, adequate levels of reliability were able to be achieved in some of the more objective measures on the analytic rating scale (i.e., Sense of Audience, Ending Punctuation, Other Punctuation, Sentence Complexity, and Sentence Soundness). Therefore, it would likely be beneficial to continue exploring different methods of objectively assessing various features of writing in the writing samples of college students to determine which seem to be most reliable and identify writing problems in this population. Furthermore, it would be beneficial to continue to analyze persuasive writing samples rather than writing samples composed in other genres, as higher cognitive and linguistic demands are placed upon writers as they compose persuasive texts than other texts in other genres (see Berman & Katzenberger, 2004; Nippold, 2000). Therefore, it is possible that some of the breakdowns seen in the writing samples of undergraduates with WLDs may not be present or as obvious in other genres. In the current study, the frequency and type of spelling errors and punctuation errors,
and well as the frequency of run-on sentences, sentence fragments, and grammatical errors/awkwardness, showed the most promise as features that might distinguish between undergraduate students with and without writing disorders.

While the use of these more objective measures may increase reliability, there are many features of writing that are difficult to assess in an objective way, such as organization, logical reasoning, or other elements of macrostructure. Therefore, subjective measures, such as rating scales, may be necessary to measure some of these features of writing.

Overall, because rating scales are much quicker and easier to use than more objective measures of writing, it would be valuable to continue to research the use of rating scales to assess the persuasive writing abilities of college students who may struggle with writing. More specifically, methods of increasing the reliability of rating scales should be explored. Additionally, attention should be given to determining the most efficient and reliable methods of measuring various features of writing, whether subjective or objective in nature.

Similarities and Differences Between Typical Participants and Participants with WLDs

A second purpose of this study was to determine which features of persuasive writing seem to best distinguish between typical undergraduate students and undergraduate students with WLDs. Both quantitative and qualitative analyses were used to answer this question.

Similarities and differences between participant groups. The results of the current study suggest that there are several features of persuasive writing that differentiate between undergraduate students with and without WLDs. More specifically, results of Fisher’s exact test and t-tests showed that typical participants outperformed participants with WLDs on the holistic rating scale and several categories of the analytic scale, including Spelling/percentage of misspelled words, percentage of sentences with other punctuation errors, percentage of sentences
that are sound, and total analytic rating scale scores. Furthermore, results for *Logical Reasoning*, *Word Choice*, and *Organization* approached significance in their ability to differentiate between participant groups.

The qualitative findings supported these results, too. More specifically, participants with WLDs received fewer positive comments and more negative comments than typical participants, suggesting that the overall quality of their writing samples was weaker than typical participants. Additionally, the types of comments given to the different groups of participants suggested that those with WLDs performed more poorly than typical participants in the areas of *Linguistic*, *Mechanics*, *Content*, and *Quality*. Furthermore, the matched-pair comparisons revealed that participants with WLDs typically performed more poorly than their typical peers on the holistic rating scale and several areas of the analytic rating scale, including *Logical Reasoning*, *Spelling/percentage of misspelled words*, *Punctuation/percentage of sentences with other punctuation errors*, and *percentages of sentences that were sound*.

Some of the features of writing that did not differentiate between participants with and without WLDs were *Presence of Prompt Components*, *Sense of Audience*, *percentage of sentences with ending punctuation errors*, and *Sentence Complexity/percentage of complex sentences*. It is possible that the *Presence of Prompt Components* category failed to differentiate groups due to the nature of the writing prompt, which specifically outlined the content that participants were supposed to address in their writing samples. Had this outline not been provided, it is possible that the results may have been different. More specifically, there is a chance that the participants with WLDs in the current study would have struggled to include all of the necessary macrostructural features expected in persuasive writing samples without the
outline provided in the prompt, as researchers have found that individuals with LLDs do not include all necessary macrostructural elements in their writing (e.g., Hall-Mills & Apel, 2012).

The lack of differences between groups seen for Sense of Audience was likely due to the fact that it was rare for participants in either group to include sentences in their writing samples that were inappropriate for the audience of a university president. More specifically, none of the participants with WLDs and few of the typical participants produced inappropriate sentences in their writing. This feature was included on the analytic rating scale due to other writing samples that have been collected from college students with WLDs by the primary investigator and her advisor that have demonstrated the frequent use of inappropriate language for the given audience. It is possible that more differences would have been seen in this area if a larger number of participants with WLDs had been included in the study.

Finally, Sentence Complexity/percentage of sentences that were complex did significantly differ between typical participants and participants with WLDs. Additionally, in three of the four matched pairs, typical participants produced a lower proportion of complex sentences than participants with WLDs. This suggests that sentences that are complex are not necessarily well written. In fact, many of the complex sentences written by participants with WLDs were awkwardly worded and contained errors of morphology and/or syntax, such as:

If during a project they [college students] need sources, they should take time to carefully read and choose them when choosing the first to pop up and wastes thirteen pages.

Conversely, complex sentences written by typical participants usually contained few errors and were easy to read, such as:
In order to save money and cut costs, the university has decided to make students pay for printing.

Therefore, it appears that Sentence Complexity does not reflect the actual quality of the sentences in a sample. Furthermore, Sentence Soundness might a better measure to use to differentiate between the writing samples of undergraduates with and without WLDs, as it takes into consideration the quality and clarity of a sentence.

During data analysis, there were a few provocative findings. First of all, the qualitative analysis of spelling errors revealed that typical participants produced a much higher proportion of Apostrophe spelling errors than participants with WLDs. A closer analysis of the data revealed that one participant did not use any apostrophes in his or her writing sample, which accounted for 30% of the apostrophe errors made by typical participants. However, even with those Apostrophe errors removed, typical participants still produced a larger proportion of Apostrophe errors than participants with WLDs. A closer analysis of the Apostrophe spelling errors made by each participant group revealed that typical participants had an average of 2.36 opportunities to produce Apostrophe errors per sample, while participants with WLDs had an average of 2.25 opportunities to produce Apostrophe errors per sample. However, typical participants produced Apostrophe errors on 28% of these opportunities and participants with WLDs only made errors on 11% of these opportunities. Therefore, both participant groups had similar opportunities to produce Apostrophe errors, but typical participants made these errors more frequently than participants with WLDs. It is possible that the frequency of errors may have been more balanced between groups if there had been an equal number of participants in each group. More specifically, there is a chance that other individuals with WLDs would have
made *Apostrophe* errors more frequently than the four participants used in the current study, which would have balanced out the frequency of errors between participant groups.

Another provocative finding was related to the types of comments given by raters. More specifically, raters tended to provide primarily negative comments regarding microstructural elements of writing (i.e., *Linguistic* and *Mechanics* comments), but primarily positive comments regarding macrostructural elements of writing (i.e., *Content* and *Quality* comments). These findings suggest that problems with microstructure may be more obvious to readers than problems with macrostructure. This makes sense, as writers would be expected to have strong, solid microstructural skills at the college level. Therefore, problems in these areas would be especially noticeable to readers, while issues with macrostructure would be less so. Conversely, readers are impressed by and comment on strong macrostructural elements in writing samples because they do not have the same level of expectations for these skills as they do for microstructural elements. They do not usually notice or comment on strong microstructural skills because these skills are expected to be strong in college students.

**Connection to previous research.** Many of the findings of the current study were consistent with previous research. For example, the fact that typical participants received higher holistic scores than participants with WLDs was consistent with the findings of Gregg et al. (2002), who found that college students with LDs received lower scores on holistic rating scales for expository writing than those without disabilities. Furthermore, researchers have found that adolescents and adults with LLDs have difficulty with organization, idea development, and theme development (similar to *Organization* and *Logical Reasoning*) (Dockrell et al., 2009; Harrison & Beres, 2007). Additionally, many researchers have found that adolescents and adults with LLDs tend to produce a high number of grammatical errors in their writing (e.g., Dockrell
et al., 2009; Duquès, 1989; Smith-Lock et al., 2009; Suddarth et al., 2012), which was consistent with the finding of differences between participant groups in the area of Sentence Soundness/percentage of sentences that were sound. However, it is important to note that the measure of Sentence Soundness on the rating scale used in the current study went beyond the grammatical measures used in other studies, as it examined instances of awkwardness and issues of clarity in sentences in addition to grammatical errors. Therefore, it was able to capture some of the features of writing described by Shaughnessy (1977) that may not necessarily be considered grammatically incorrect, but are awkwardly worded or difficult for readers to fully understand. Other findings consistent with previous research were the increased number of spelling and punctuation errors and reduced total analytic rating scale scores in participants with WLDs, as other researchers have found differences between adolescents and adults with and without LLDs in regards to spelling errors, punctuation errors, and overall writing quality (e.g., Duquès, 1989; Gregg et al., 2002; Harrison & Beres, 2007; Smith-Lock et al., 2009; Suddarth et al., 2012). Finally, the results for Word Choice in the current study were unable to be compared to the findings of other studies, as this measure examined lexical sophistication and other studies focused on lexical diversity.

Despite all of the similarities to the findings of previous research, some differences were also seen between the results of the current study and other studies. First of all, the fact that differences in Presence of Prompt Components were not found between participants with and without WLDs went against the findings of Hall-Mills and Apel (2012), who found that adolescents with LLDs included only half of the expected macrostructural elements in their narrative and expository writing samples. However, the content outlined in the writing prompts used in their study differed from the macrostructural elements that they were counting. For
example, the expository writing prompt asked participants to write a paper describing a career that they would like to have after finishing high school and why they would like to have that particular career. The macrostructural elements that were counted included (1) assignment (i.e., addressed all parts of the assignment), (2) logical sequence (i.e., ideas were organized logically and clearly), (3) introduction (i.e., provided a clear and complete thesis statement), (4) body (i.e., provided adequate support for their thesis statement), and (5) conclusion (i.e., provided adequate ending that extended the topic). Therefore, only the first macrostructural element was clearly outlined in the writing prompt. However, the results indicated that only two of the 12 participants adequately addressed the assignment macrostructural element. This indicates that their findings truly were different than those of the current study, as participants in this study typically included the assignment elements outlined in the writing prompt.

Additionally, many other researchers have found reduced grammatical complexity abilities in adolescents and adults with LLDs (e.g., Morris & Crump, 1982; Smith-Lock et al., 2009; Hall-Mills & Apel, 2012), while the current study did not find significant differences in this area (i.e., Sentence Complexity/percentages of sentences that were complex). However, the current study examined grammatical complexity by calculating the percentage of sentences that were complex, while other studies tended to measure grammatical complexity in other ways (e.g., mean length of T-unit or number of complex correct sentences). Therefore, it is difficult to directly compare the results of the current study to those of other studies that used different measures of grammatical complexity.

**Clinical and research implications.** First of all, the results of the current study suggest that some of the features on the existing analytic rating scale either need to be removed from the scale or measured in different ways to strengthen raters’ abilities to differentiate between
undergraduate students with and without WLDs (i.e., Presence of Prompt Components, Sense of Audience/percentage of sentences that were inappropriate, Punctuation/percentage of sentences with ending punctuation errors, Sentence Complexity/percentage of sentences that were complex, and Sentence Soundness). For example, although the Sentence Complexity category on the rating scale did not differentiate between participants with and without WLDs, qualitative analyses revealed that participants with WLDs tended to produce more run-on sentences and fragments than typical participants. Therefore, rather than measuring the percentage of complex sentences, it may be more useful to measure the percentage of sentences that are run-ons or fragments.

Additionally, while Punctuation and percentage of sentences with ending punctuation errors did not differ between groups, the percentages of sentences with other punctuation errors did. Thus, it may be better to examine the percentage of sentences with any type of punctuation error instead of analyzing ending and other punctuation errors separately. Overall, it would be beneficial to further examine different methods for measuring the various features of writing to try to identify those that are clinician-friendly and capable of identifying college students with writing problems.

Furthermore, the results of statistical analyses revealed that the percentages of word- and sentence-level features did a better job of distinguishing between participant groups than scores on the analytic rating scale. For example, the percentage of sentences with other punctuation errors differentiated between participant groups, but the Punctuation category on the rating scale did not distinguish between groups. This suggests that the percentages may be better measures than the rating scale, or that the criteria on the rating scale need to be adjusted to better reflect differences between groups.
In addition to categories on the analytic scale, it was also found that scores on the holistic rating scale differentiated between participants with and without WLDs. Because the holistic rating scale was faster and easier to use than the analytic rating scale and differentiated between undergraduates with and without WLDs, it shows promise as a means of identifying college students who struggle with writing. More specifically, a holistic rating scale could be used as a screening measure for college students who may present with writing problems, as it appears this type of measure would be able to quickly identify students with weak writing skills. If students perform poorly on the holistic rating scale, SLPs could further evaluate their writing abilities using an analytic rating scale. The use of the holistic scale would be quick and easy at identifying problems, while the analytic rating scale would be able to pinpoint specific strengths and weaknesses to help SLPs plan individualized treatment goals and methods for their clients.

Additionally, the fact that participants with WLDs performed more poorly than typical participants in nearly all areas suggests the need for intensive intervention for undergraduates with writing difficulties. Furthermore, because all of the participants with WLDs demonstrated some differences in their individual strengths and weaknesses, it will also be important for SLPs to provide individualized treatment to undergraduates with writing difficulties. For example, the findings from the qualitative analysis of spelling errors revealed both similarities and differences in the types of spelling errors produced by individual participants with WLDs. Treatment methods for spelling problems (and other writing difficulties) differ depending upon the specific needs and difficulties of the client. Therefore, because the types and levels of writing difficulties are different for each college student with a WLD, it is important for SLPs to closely analyze the individual strengths and weaknesses of their college clients to determine the appropriate goals and treatment methods to use for each client.
Finally, the results of the current study revealed that there are some features of writing that are weak in both typical undergraduates and undergraduates with WLDs. More specifically, both groups struggled to appropriately use commas and produce sound sentences. This suggests that many of today’s undergraduate students may possess some of the characteristics of the basic writers described by Shaughnessy (1977). Therefore, it is likely that most undergraduate students with and without WLDs would benefit from additional instruction and practice in these areas (perhaps through the undergraduate composition courses required at most universities).

Overall, results for the second research question suggest that undergraduates with WLDs perform more poorly than their typical peers in a number of areas, which is consistent with the existing literature on this topic. Furthermore, the differences between participant groups suggested that the writing difficulties in undergraduates with WLDs likely stem from underlying linguistic weaknesses, as analyses of spelling errors and sentence structure revealed phonological difficulties and problems generating syntactically sound sentences. Thus, it appears to be possible to distinguish between undergraduates with and without WLDs, as both the holistic and analytic rating scales show promise in distinguishing between these two populations. Features on the analytic scale that were best able to identify writing problems included Logical Reasoning, Organization, Word Choice, Spelling, Punctuation, and Sentence Soundness. Therefore, diagnostic writing measures for undergraduate students should focus on examining their persuasive writing skills in these areas to help determine whether or not writing problems exist.
Features of Persuasive Writing Relating to Overall Writing Quality

The final research question sought to determine the features of writing that relate to overall writing quality. Both quantitative and qualitative methods were used to answer this question.

**Similarities and differences between participant groups.** The quantitative results revealed that participants with high holistic scores received higher scores in the areas of *Logical Reasoning, Organization, Spelling,* and *Sentence Soundness* than participants with low holistic scores. Additionally, those with high holistic scores made fewer spelling errors, produced more sound sentences, made fewer other punctuation errors, and had higher total scores on the analytic rating scale than those with low holistic scores. These findings suggest that these writing features distinguish between those with high versus low quality writing samples.

The qualitative results supported these findings. More specifically, those with high holistic scores received more positive comments and fewer negative comments than those with low holistic scores across nearly all themes and subthemes. Additionally, in the matched pairs, participants with high holistic scores always or usually had higher scores than participants with low holistic scores in the areas of *Presence of Prompt Components, Logical Reasoning, Organization, Punctuation, Sentence Complexity,* and total analytic rating scale scores. Participants with high holistic scores also always or usually made fewer spelling errors, had higher proportions of sound sentences, made fewer other punctuation errors, and produced more complex sentences than participants with low holistic scores.

**Clinical and research implications.** Overall, these findings suggest that certain features of persuasive writing have a stronger impact on the overall writing quality of undergraduate students than other features. More specifically, the areas of *Logical Reasoning, Organization,*
Spelling/percentage of word that were misspelled, percentages of sentences with other punctuation errors, and Sentence Soundness/percentages of sentences that were sound seemed to best distinguish between participants with high and low holistic scores. Therefore, when providing therapy to undergraduate students with writing difficulties, SLPs will likely find that the greatest gains in overall writing quality are made when weaknesses in these areas are targeted.

Additionally, as mentioned in the discussion of the second research question, raters tended to provide more negative comments for microstructural elements than macrostructural elements and more positive comments for macrostructural elements than microstructural elements. Therefore, it appears that raters may be more distracted by problems with microstructure than macrostructure. This was consistent with Shaughnessy’s (1977) suggestion that readers can tolerate a certain number of grammatical errors (e.g., problems with verbs, nouns, pronouns, and subject-verb agreement) while reading writing samples, but that an excessive number of these errors makes it difficult for readers to focus on anything else.

Participants with low holistic scores tended to produce a high number of microstructural errors (e.g., spelling, grammatical, and other punctuation errors) that interfered with the flow and clarity of their writing, which likely contributed to the raters’ perception of these samples as having low overall quality. Therefore, SLPs may find it beneficial to target these microstructural skills in therapy for undergraduates who struggle with writing, as it is likely that the professors will also find these errors to be distracting and penalize students for these errors.

The analyses for this research question also showed that the holistic rating scale was able to distinguish between participants with strong and weak writing skills, as those with high holistic scores performed better across most areas of the analytic rating scale than those with low
holistic scores. This finding provides further evidence that holistic rating scales can be useful in identifying undergraduate students with weak writing skills. However, while all of the participants with diagnosed WLDs in the current study received low holistic scores, there were also a number of typical participants who received low holistic scores. These participants did not report a history of language, literacy, or learning problems, so it unclear whether or not the writing difficulties seen in the writing samples of these participants would justify the diagnosis of a WLD. It is possible that these participants might represent the basic writers described by Shaughnessy (1977), who have difficulties with writing due to a lack of education or experience rather than underlying language or learning issues. Alternatively, this finding could also indicate that the holistic rating scale may have strong sensitivity (i.e., correctly identified participants with WLDs as having writing problems) but poor specificity (i.e., several typical participants were identified as having writing problems). Either way, further assessment would be necessary to determine whether or not intervention was warranted for these students.

Even though the holistic rating scale identified both typical undergraduates with weak writing skills and undergraduates with WLDs stemming from linguistic weaknesses, this may not be an issue. If typical undergraduates demonstrate problems with writing that would benefit from intervention, they should be identified. While it is currently unclear whether or not the writing problems of typical undergraduates with weak writing skills are similar to those experienced by undergraduates with WLDs, these students would likely still struggle to complete and perform well on writing assignments at the college level. Therefore, it seems likely that they would benefit from receiving some form of intervention. However, it is not yet clear whether their writing difficulties would justify services through a speech and hearing clinic or a less formal setting, such as a university writing center or tutoring service. In either case, use of a
holistic scale could serve to identify college students with weak writing skills, and an analytic scale could reveal whether or not their difficulties required the attention of an SLP or another professional.

Finally, although the holistic rating scale may be able to identify undergraduates with writing difficulties, it does not provide enough information about the strengths and weaknesses of these students in order to develop an appropriate treatment plan. Therefore, SLPs would need to conduct a more careful analysis of individual strengths and weaknesses than what a holistic rating scale can provide, such as through the use of an analytic rating scale. Therefore, as mentioned previously, it seems that holistic rating scales may have value as screening tools, while analytic rating scales could be used to help plan specific therapy goals and methods.

**Limitations of the Study**

While the results of this study do provide insight into the writing skills of undergraduate students and the feasibility of using rating scales to evaluate the writing abilities of this population, there were some limitations to the current study.

One of the limitations of the study related to the selection and sample size of participants. All of the typical participants were recruited through classes or groups in the field of Communication Sciences and Disorders. While not all of the typical participants were majors in this field, they did not adequately represent all fields of study on the campus. Because the writing instruction and expectations vary from one field to another, it is possible that the inclusion of participants from all fields of study on campus would have resulted in different results.

Additionally, this study included only 54 participants (i.e., 50 typical undergraduates and four undergraduates with WLDs). More than 20,000 students are enrolled at the university
where participants were recruited. Therefore, the sample size of the current study does not sufficiently represent the students at the studied university. Furthermore, the results of this study cannot be applied to undergraduate students at other colleges and universities across the country.

Another issue relating to sample size was the fact that data were only collected from four undergraduates with WLDs. This limited the types of statistical analyses that were able to be used to compare participants, and it also limits the trustworthiness of the results. Writing samples would need to be collected from much larger samples of undergraduate students with and without WLDs across the country than in this study for the results to be applied to all undergraduate students in the U.S.

The methods of data collection are also another potential limitation. While the method of collecting writing samples for the current study was much more authentic than many of the tasks using on other measures of writing, it was different than the types of writing tasks typically expected of college students in the classroom. More specifically, college students are typically allowed to type their writing samples (rather than handwrite) and are given several days or weeks to produce a writing sample (rather than 20 minutes). Therefore, the results of this study do not necessarily represent the full writing abilities of undergraduate students, as the writing products produced for the current study were likely different than the products they would produce for their classes.

Another limitation of the study was the fact that both of the rating scales demonstrated issues with inter-rater and intra-rater reliability. That said, the first aim of this study was to examine the reliability of these kinds of writing scales, so the problems with reliability were actually an important finding. Moreover, although inter-rater reliability scores were lower than ideal, all disagreements were discussed and consensus reached for the samples that were scored
by all three raters. Therefore, the problems with reliability do not impact the trustworthiness of
the findings of the current study, as the agreed-upon scores were used for the quantitative and
qualitative analyzed conducted to answer the second and third research questions.

A final potential limitation of this study lies in the data analysis process. First-year
graduate RAs were used to analyze the writing samples using the holistic and analytic rating
scales. While they appeared to work hard and give their best effort during these tasks, it is
possible that the results of this study would be different if practicing SLPs had been used to score
and code the writing samples. As first-year graduate students, the RAs had not completed all of
their graduate-level coursework and had limited clinical experiences. These factors could have
influenced their scoring and coding of the samples. However, when previous versions of the
rating scale were being tested, problems with reliability were found when the primary
investigator (a doctoral student and clinical fellow at the time) and her advisor (i.e., a clinical
expert and professor in communication disorders) scored writing samples using the rating scale.
Therefore, it is likely that differences in scoring and coding were due to the inherent reliability
problems seen in rating scales rather than the educational and clinical experiences of the raters
used in the current study.

**Implications for Future Research**

While the findings of the current study provided much insight into the writing abilities of
undergraduate students and the feasibility of using rating scales to measure the writing skills of
this population, they also leave many questions unanswered and provide ideas for future studies.
First of all, it appears that several changes are necessary to improve the reliability of the existing
holistic and analytic rating scales. Inter-rater and intra-rater reliability were low in areas that
were subjective in nature, such as the holistic scale or *Logical Reasoning* and *Organization* on
the analytic scale. To improve reliability, the criteria for each point value of these areas will likely need to be revised. These revisions could include more detailed descriptions and an increased use of examples than what are currently provided with the existing analytic rating scale. These types of revisions were made from the previous two versions of the analytic rating scale to the version used in the current study and improvements in inter-rater reliability were seen. Additionally, during the process of resolving discrepancies in scores among the raters, several issues of confusion were mentioned that could be addressed to make the analytic rating scale clearer and easier to use than it is in its current form. These types of changes would likely improve reliability. However, because problems with reliability are characteristic of rating scales, it is possible that acceptable levels of reliability may not be achieved even with continued revisions and improvements.

Another possible way of attempting to improve reliability would be to create a more rigorous training procedure. For example, the training procedure could be adapted to be more like the Writing Process Test, which requires examiners to practice scoring writing samples using five-point analytic rating scales until adequate levels of agreement are reached between the examiner and pre-scored samples provided in the test manual (Warden & Hutchinson, 1992). During the training procedure for the current study, all three of the raters worked together while scoring six samples and were provided feedback about their scoring procedures. However, they were not required to reach a certain level of agreement before they started scoring samples independently. Furthermore, by being able to work together, they were not forced to make decisions on their own or spend as much time thinking about how to score and code the samples. Instead, they were able to rely on the responses of the other raters. Therefore, requiring raters to score samples independently until they reach an acceptable level of agreement with pre-scored
samples before they start scoring actual writing samples could potentially increase both inter-rater and intra-rater reliability.

Other changes could also be made to the existing analytic rating scale to improve its ability to differentiate between undergraduate students with and without WLDs. More specifically, some of the categories on the existing analytic rating scale did not appear to differentiate between participants with and without WLDs. Therefore, other methods of measuring these same features of writing could be tested to determine whether or not they provide a better means of identifying undergraduates with WLDs. For example, *Sentence Complexity* did not appear to differ between typical participants and participants with WLDs. However, an in-depth analysis of punctuation errors revealed that participants with WLDs tended to produce more run-on sentences and sentence fragments than typical participants. Therefore, it is possible that a measure focusing on calculating the percentage of sentences that are run-ons or fragments would be better at differentiating between groups than the percentage of sentences that are complex. Overall, examining other methods of measuring the writing features that did not differentiate between participants with and without WLDs in the current study could potentially reveal other measures that would be better suited to identify writing disorders in undergraduate students.

Overall, both the holistic and analytic rating scales appeared to differentiate between participants with and without WLDs. However, because the holistic rating scale was quicker and easier to use than the analytic rating scale, it would be beneficial to further examine the potential of holistic rating scales to identify undergraduate students who may present with writing difficulties. If holistic rating scales are able to accurately identify undergraduates with writing problems, they would provide a fast and easy way for SLPs to screen the writing skills of college
students who may present with writing disorders. Additionally, because holistic rating scales provide limited information about the specific strengths and weaknesses of a writer, it would also be beneficial to continue to work toward developing an analytic rating scale that could be used to assess the writing skills of undergraduate students. Ideally, this analytic rating scale would provide detailed information about a writer’s specific strengths and weaknesses so that SLPs could develop appropriate individualized intervention plans and use the rating scale to monitor their clients’ progress. Overall, more work is needed to determine if reliable and valid rating scales can be developed that accurately identify writing problems in undergraduate students and provide enough information for SLPs to develop appropriate treatment plans for their college-aged clients.

Another possible direction for future studies would be to more closely examine similarities and differences in the use of the writing process between undergraduates with and without WLDs. The existing standards for college students place a heavy emphasis on the need to be able to flexibly use all phases of the writing process, so it would be important to know how well undergraduates with WLDs are able to carry out all phases of the writing process. Given that there are few existing writing measures that consider the writing process, the development of a tool that could examine these skills in undergraduate students would be beneficial. However, an examination of the writing process is not necessary to identify writing difficulties in undergraduate students. Instead, the assessment of the writing process could be used for students who have been identified as having a writing disorder and are receiving treatment for their writing difficulties. Individuals with WLDs may struggle to carry out all phases of the writing process, or may use inefficient or inappropriate strategies while engaged in the writing process.
Therefore, this type of assessment could be used to identify these problems so that they could be addressed in therapy.

Additionally, there were certain features of writing that both typical participants and participants with WLDs seemed to struggle with in this study. More specifically, both groups received relatively low scores in the areas of Punctuation and Sentence Soundness. It would be beneficial for future studies to closely examine the weaknesses of undergraduates in these areas and determine the similarities and differences seen between the abilities of undergraduates with and without WLDs. These analyses could provide college-level writing instructors with information about the types of writing skills that should be addressed in their courses with typical undergraduate students. Additionally, they could provide more information about how to best differentiate between the writing samples of undergraduates with and without WLDs.

Finally, the existing literature on the writing abilities of undergraduate students is not representative of all undergraduate students in the country, as the studies include small sample sizes of participants from limited regions in the U.S. Therefore, studies using larger sample sizes and a variety of participants from different colleges and universities across the country are needed to develop a stronger understanding of the writing abilities of undergraduate students with and without WLDs.

**Overall Significance of the Study**

The primary goals of this study were to learn more about the writing abilities of undergraduate students and determine the feasibility of using rating scales to assess the writing of this population. Overall, the results of this study suggest that rating scales can be effective tools for measuring the writing skills of undergraduate students, as inter-rater and intra-rater reliability were relatively high for certain aspects of the rating scales and the rating scales were
able to differentiate between undergraduate students with and without WLDs. However, inter-rater and intra-reliability did not reach acceptable levels across all categories of the rating scales, so continued revisions will need to be made to attempt to increase reliability to satisfactory levels.

Additionally, the fact that differences were seen in the performance of typical participants and participants with WLDs on the holistic and analytic rating scales indicates that it is possible to differentiate between the writing samples of these populations. This is an important finding, as it suggests that it is possible to identify undergraduate students with writing problems when the appropriate measure is used. It may be possible that both types of rating scales can be used in the diagnostic process, with a holistic rating scale used for initial screening purposes and an analytic rating scale used to identify features of writing that should be targeted during intervention. The features of writing that differed between the two groups show promise as features that could be included on an analytic rating scale used to develop a treatment plan, as undergraduate students with WLDs are likely to have difficulties in those areas. Additionally, the features of persuasive writing on the analytic rating scale that did not differ between groups may be best measured in different ways to discriminate between undergraduates with and without WLDs, or may not need to be included on measures developed to identify writing problems in this population.

Finally, the fact that there appeared to be certain features of writing that differentiated between participants with high and low holistic scores suggests that these are the skills that most relate to overall writing quality. This finding is clinically relevant, at it suggests that these are the skills that would result in the greatest improvements in the overall writing abilities of undergraduate students with writing difficulties when targeted in therapy.
Overall, the findings of the current study are clinically relevant for both the assessment and treatment of writing problems in undergraduate students. Continued research in this area can help contribute to the development of strong writing assessment and intervention tools to be used with this population. Given the high number of students enrolled in colleges and universities that are likely to have writing difficulties and the limited number of resources available to provide writing assessment and treatment for this population, the findings of the current and future studies in this area will be beneficial for SLPs and their clients alike.
REFERENCES


College Board. (2013a). *8 things to know about how colleges use admissions tests.* Retrieved from https://bigfuture.collegeboard.org/get-in/testing/8-things-to-know-about-how-colleges-use-admission-tests

College Board. (2013c). *Big future: College search*. Retrieved from
https://bigfuture.collegeboard.org/college-search


Dockrell, J. E., Lindsay, G., & Connelly, V. (2009). The impact of specific language impairment on adolescents’ written text. *Exceptional Children, 75* (4), 427-446.


dyslexics. *Journal of Adolescent Health Care, 6*, 31-34.

language of learning disabled and non-learning disabled students at four age levels. *Learning
Disability Quarterly, 5* (2), 163-172.

narrative texts written by adults with a history of language impairment. *Reading and
Writing: An Interdisciplinary Journal, 22*, 735-752.

National Governors Association Center for Best Practices, Council of Chief State School
Governors Association Center for Best Practices, Council of Chief State School.

Nippold, M. A. (2000). Language development during the adolescent years: Aspects of

Nippold, M. A. (2010). It’s not too late to help adolescents succeed in school. *Language,
Speech, and Hearing Services in Schools, 41*, 137-138.

adolescents, and adults: A study of syntactic, semantic, and pragmatic development.

Success in Postsecondary Writing. *College English, 74* (6), 520-533.

institutional model for college writing assessment. *College Composition and
Communication, 60* (2), 285-320.


APPENDIX A

Summaries of Existing Writing Measures Normed for College Students

Kaufman Test of Educational Achievement – 2nd Edition (KTEA-2)

The Comprehensive Form of the KTEA-2 (Kaufman & Kaufman, 2004) is normed for individuals between the ages of 4;6 to 25;11, or in grades kindergarten through 12, and contains two parallel forms. Subtests that measure writing on the Comprehensive Form include Written Expression and Spelling, which can be combined to generate a Written Language composite. Total administration time for both of these subtests is approximately 25 to 30 minutes. A Brief Form of the KTEA-2 is also available, which is normed for individuals between the ages of 4;6 to 90;11. On this form, the Written Expression subtest examines written language and spelling skills using fewer items than on the Comprehensive Form. The Brief Form is primarily used for screening purposes.

The Written Expression subtest on the Comprehensive Form is normed for individuals between the ages of 4;6 and 25;11. Examinees in kindergarten and pre-kindergarten trace and copy letters, as well as write dictated letters. Examinees in grade 1 and above complete a variety of writing tasks in a storybook context. The test includes three different storybooks that are used with different age levels. Some of the tasks required of examinees on this subtest include filling in missing words, punctuation and capitalization to short passages, generating sentences using a provided word or phrase, generating sentences to finish a paragraph, and writing an essay based on the story the examinees’ helped complete in the storybook. Items are scored in the areas of task (i.e., generates sentences that include all required ideas and make sense), structure (i.e., writes complete sentences without errors), word form (i.e., writes grammatically correct sentences), capitalization, and punctuation. Both age grade (up to 12) norms are available for
this subtest. Additionally, an error analysis form is available for this subtest with norms up to grade 12.

The Spelling subtest on the Comprehensive Form is normed for individuals from grade 1 to age 25;11. To complete this subtest, examinees must write regular and irregular words of increasing complexity that are dictated by the examiner. Age and grade (up to 12) norms are available for this subtest. An error analysis form is also available for this subtest with norms up to grade 12.


The OWLS-2 (Carrow-Woolfolk, 2011) is normed for individuals between the ages of 3;0 and 21;11. The Written Expression Scale (normed for 5;0 to 21;11) of this measure can be administered individually or in groups and takes approximately 10 to 30 minutes to administer. This scale examines writing at the word, sentence, and paragraph level. To administer the Written Expression Scale, examiners present verbal, written, and/or visual stimuli and examinees write their responses in a provided Response Booklet. Some of the tasks required of examinees include writing dictated words or sentences, editing errors of punctuation and capitalization, combining sentences, and completing a story. The items on this measure are scored in the areas of conventions (i.e., letter formation, word formation, spelling, capitalization, punctuation, note conventions, and general conventions), lexical/semantic, syntactic (i.e., grammatical morphemes, sentence structure, and general syntax), pragmatic, and text structure (i.e., text organization, use of detail, and cohesion). Raw scores can be converted into standard scores, percentile ranks, age equivalents, and grade equivalents.
Peabody Individual Achievement Test – Revised/Normative Update (PIAT-R/NU)

The PIAT-R/NU (Markwardt, 1998) is normed for individuals between the ages of 5;0 to 22;11 and in grades kindergarten through 12. Although norms are not given for college students, this test could be useful for freshmen who have just completed high school or used as a criterion-referenced measure for college students beyond the freshmen level. The Spelling and Written Expression subtests measure writing skills and can be administered individually or in combination with other subtests in this measure. If both the Spelling and Written Language subtests are administered, a Written Language composite score can be derived.

On the Spelling subtest, examinees are shown multiple choice arrays of letters or words and are expected to identify the correct response based on a letter sound, letter name, or word spoken by the examiner. This subtest is untimed and raw scores can be converted using either age or grade norms.

The Written Expression subtest examines pre-writing skills in kindergarten and first grade examinees and story writing skills in second to twelfth grade examinees. The older examinees are given a picture prompt and expected to compose a story within 20 minutes. Scores on this subtest are based on handwriting legibility, sentence structure, punctuation, capitalization, coherence, grammaticality, and various elements of story structure. Raw scores on this subtest can be converted to grade-based or grade-based scaled scores, or raw scores can be compared to the entire standardization sample.

Spelling Performance Evaluation for Language and Literacy – 2nd Edition (SPELL-2)

SPELL-2 (Masterson et al., 2006) is a standardized measure that can be used with individuals from age 7 to adult, or grade 2 through adult. This measure takes approximately 30 to 60 minutes to administer and provides an in-depth analysis of examinees’ spelling skills.
SPELL-2 is administered using a computer with minimal supervision, but examinees can write their responses on paper if they would prefer. The computer program automatically adjusts to the examinee’s abilities and conducts a thorough analysis of examinees’ spelling errors. This analysis provides insight into the examinees’ skills in the areas of phonological awareness, orthographic knowledge, morphological knowledge, semantic relationships, and mental orthographic images. SPELL-2 automatically generates results and recommendations reports, which provide information about how well examinees’ performed in 60 different categories and suggestions for how to improve their spelling and word reading skills.

Test of Adolescent and Adult Language – Fourth Edition (TOAL-4)

The TOAL-4 (Hammill et al., 1994) is normed for individuals between the ages of 12;0 to 24;11. The Writing/Vocabulary and Writing/Grammar subtests measure writing at the sentence level. The Writing/Vocabulary subtest takes 10 to 25 minutes to administer and requires examinees to read a stimulus word and compose a sentence that appropriately uses the word (e.g., “happy”, “He was happy that he won the race.”). In order to receive credit for an item on this subtest, the examinee must use the word correctly in the sentence and generate a sentence that demonstrates appropriate meaning overall. Examinees are not penalized for errors of spelling, punctuation, or capitalization. Additionally, nonstandard English usage within sentences is not penalized. The Writing/Grammar subtest takes 15 to 35 minutes to administer and requires examinees to combine two to six simple sentences into a single simple, compound, or complex sentence (e.g., “I lost my phone. I lost it on Thursday.” = “I lost my phone on Thursday.”). To receive credit for each item, examinees must create a grammatically correct sentence that maintains the meanings of the combined sentences. Errors of spelling, punctuation, and capitalization are not penalized. Raw scores on each subtest can be converted into
percentiles and standard scores. Additionally, several composite scores can be calculated by combining scores on various subtests.

**Test of Written Spelling – Fifth Edition (TWS-5)**

The TWS-5 (Larsen, Hammill, & Moats, 2013) is normed for individuals between the ages of 6;0 and 18;11, or in grades 1 to 12. While this measure was not normed with college students, it can still be used as a criterion-referenced measure for adolescents and adults beyond high school. The TWS-5 measures spelling skills by having examinees spell words dictated by the examiner. Administration time is approximately 20 minutes. Raw scores can be converted to percentiles, standard scores, age equivalents, and grade equivalents.

**Wide Range Achievement Test – Fourth Edition (WRAT-4)**

The WRAT-4 (Wilkinson & Robertson, 2006) is normed for individuals aged 5;0 to 94;11 and has two equivalent forms. The Spelling subtest of this measure examines writing skills at the word level. More specifically, on this subtest, spelling skills are measured by having examinees write dictated letters (ages 7 years and younger) or spell dictated words (ages 8+ years). The Spelling subtest can be administered individually or in a group setting. Raw scores on this measure can be converted to standard scores, percentile ranks, normal curve equivalents, stanines, and grade equivalents.

**Woodcock-Johnson III Normative Update Tests of Achievement (WJ III NU)**

The WJ III NU (Woodcock, McGrew, et al., 2007; Woodcock, Shrank, et al., 2007) is normed for individuals from age 2;0 to 90+ years. Three forms of this test are available, including Forms A and B of the Standard and Extended Batteries, as well as Form C of the Brief Battery. The subtests that measure writing skills on the Standard and Extended Batteries include Spelling (Test 7), Writing Fluency (Test 8), Writing Samples (Test 11), Editing (Test 16),
Spelling of Sounds (Test 20), and Punctuation and Capitalization (Test 22). Additionally, the
Handwriting Legibility Scale and Writing Evaluation Scale are available to examine samples of
writing collected during testing or from other sources (e.g., classroom writing assignment).
Several composite scores can be generated by combining scores on various writing subtests,
including Broad Written Language (Spelling, Writing Fluency, and Writing Samples), Basic
Writing Skills (Spelling and Editing), and Written Expression (Writing Fluency and Writing
Samples). On the WJ III NU Brief Battery, the Spelling (Test 3), Writing Samples (Test 6), and
Writing Fluency (Test 9) subtests examine writing skills.

To complete the Spelling subtest, examinees are required to draw lines, trace letters,
produce uppercase and lowercase letters, and spell dictated words. The skills measured depend
upon the examinees’ age and scores are based upon correctness of letter formations and/or
spelling.

On the Writing Fluency subtest, examinees must generate and write as many simple
sentences as possible in seven minutes when given a stimulus picture and three words to include
in the sentences. Issues with spelling, punctuation, capitalization, and/or handwriting are ignored
when scoring this subtest. Scores are based upon the number of syntactically-correct simple
sentences generated in the given time period.

The Writing Samples subtest requires examinees to write words or sentences when given
visual and/or verbal stimuli. Each item on this subtest is scored using a holistic scale that
includes point values of 0 (inadequate response), 0.5 (borderline response – does not clearly fit
descriptions of inadequate or standard responses, but falls in between), 1 (standard response), 1.5
(borderline response – does not clearly fit descriptions of superior or standard responses, but falls
in between), and 2 (superior response). Items on this subtest are scored based on content
adequacy, adherence to task requirements, and appropriateness of sentence structure. On most items, problems with spelling, punctuation, capitalization, and handwriting are ignored unless these issues severely impact the intelligibility of the response.

The Editing subtest requires examinees to read short written passages and then identify and correct errors of capitalization, punctuation, word usage, or spelling. However, these corrections are made verbally rather than in writing. Items are scored based on whether or not examinees accurately identify and correct the errors.

To complete the Spelling of Sounds subtest, examinees must write letters based upon sounds provided by the examiner. Additionally, examinees listen to an audio recording of nonwords and low-frequency words that contain regular English spelling patterns and spell the words that they hear. Items are scored based on whether or not the words were spelled using appropriate letter patterns. Results from this subtest provide insight into examinees’ phonological and orthographic coding skills.

The Punctuation and Capitalization subtest requires examinees to perform various tasks that ask them to provide examples of their punctuation and capitalization skills. Some of the items require examinees to demonstrate that they know how to produce a certain punctuation mark, uppercase letter, or lowercase letter. More challenging items require examinees to write responses using correct capitalization or punctuation, or place punctuation marks in the appropriate locations in given words, phrases, or sentences.

Handwriting can be evaluated using the Handwriting Legibility Scale, which provides a standardized method of evaluating examinees’ handwriting, or an informal evaluation of handwriting elements. Examinees’ writing from various subtests of the WJ III NU can be analyzed or samples can be collected from other sources. The Writing Evaluation Scale (Mather
& Woodcock, 1997) allows for the thorough examination of examinees’ writing skills by providing a method of analyzing longer writing samples, such as classroom writing assignments. This scale allows examiners to take a closer look at examinees’ handwriting, spelling, punctuation, capitalization, vocabulary, syntax/usage, narrative text structure, expository test structure, sense of audience, and affective skills. This measure is informal, so norms are not available.

For all normed subtests of the *WJ III NU*, raw scores can be converted into age equivalents, grade equivalents, percentile ranks, standard scores, W scores, T scores, normal curve equivalents, z scores, stanines, Cognitive/Academic Language Proficiency (CALP) levels, instructional zones, or Relative Proficiency Indices (RPI).

*Writing Process Test (WPT)*

Although not normed for college students, the *WPT* (Warden & Hutchinson, 1992) can be used with older students and adults as a criterion-referenced measure. The *WPT* is normed for students in grades 2 through 12, or ages 8 to 19 years, and can be administered either individually or in groups. This assessment includes two parallel forms that examine both the writing process and writing product by having examinees plan, draft, and revise a writing sample. The writing prompt is given verbally by the examiner and asks examinees to compose an article for a school newspaper about one of two topics (i.e., the most exciting or boring thing that you have ever done, or how you would spend one million dollars). Given the nature of the prompt, it is possible that examinees could compose a sample that is narrative and/or expository in nature. To examine the writing process, the examiner asks the examinee questions about how often he or she used various strategies while planning, composing, and revising his or her writing sample. The writing product is analyzed by both the examinee and examiner in the areas of
purpose/focus, audience, vocabulary, style/tone, support/development, organization/coherence, sentence structure/variety, grammar/usage, capitalization/punctuation, and spelling. Both grade- and age-based norms are available.
APPENDIX B

Script for Personal and/or Email Recruitment of BGSU NSSLHA Members for Pilot Study
(to be read by and/or emailed by NSSLHA Vice President)

Hi Fellow NSSLHA Members,

As you might know, Dr. Katz runs a specialty clinic, called the ROWing Clinic, which provides assessment and treatment services for school-age students, college students, and adults who struggle with reading, writing, and organizational skills.

With Dr. Katz and several other undergraduate and graduate research assistants, I am involved in developing and conducting a research study aimed learning more about the writing skills of both typical college students as well as college students with reading and/or writing disorders.

Currently, there are very few normed assessments available for college-age students (or adults), so many of the measures that Dr. Katz uses in the ROWing Clinic are informal. Unfortunately, it is sometimes difficult to know whether a college-student client’s performance on an informal measure is truly weak or whether it falls within normal limits for typical college students of the same age/year in school. Dr. Katz is beginning to run a series of studies aimed at learning more about what the ‘typical’ college student’s skills are in various areas of oral and written language. We conducted one of these studies last spring in the area of syntactic comprehension, and now, she and her research team are going to look at writing.

This is where you come in! We are working to develop a solid rubric to use for assessing writing, but we need a few ‘typical’ college students’ writing samples to help us fine-tune the rubric so that it is reliable. We are looking for 5-6 volunteers to participate in a pilot study aimed at fine-tuning this rubric. Participation should take no more than 45 minutes, and your name will not appear on any of the measures given. Additionally, none of your peers (including me) will know if you have volunteered to participate or actually participated. Dr. Katz has two research assistants who are new to BGSU: 1) a doctoral student, Stephanie Richards, who will be scheduling participants as well as testing them, and 2) a master’s student, Lauren Best, who will be testing students. After the data have been collected, identifying information will be removed (i.e., names as well as handwriting – the writing samples will be typed). Then, all of Dr. Katz’s research assistants – Stephanie Richards, Lauren Best, Tayler Finsel, Heather Langenkamp, Becky Spencer and I will be involved in reviewing the writing samples. Tayler, Heather, Becky, and I will not have any knowledge of who participated in the study or who has written the samples.

If you are interested in participating in this important study, please contact Stephanie by email at richasa@bgsu.edu. We can only take as many as 6 people, so if you want to participate, email her soon!

If you have any questions, feel free to talk with me, Stephanie, and/or Dr. Katz – she is happy to take emails or calls.

Thanks so much for your help with this project!
APPENDIX C

Persuasive Writing Prompt

Directions:

You will be given 20 minutes to compose a writing sample using the prompt below. Please do everything you know how to do as a writer to complete this assignment. If you would like, you may use the provided scratch paper to plan your sample before you begin. You will need to handwrite your response; if you make a mistake, just scratch through it. I cannot provide you with any assistance on this task. We want to get a good picture of your writing without assistance.

You should plan to take a position or side on the topic presented and argue your position clearly and completely.

In order to save money, the university is now making students pay for printing. They feel that by doing so, students will be more likely to be mindful of unnecessary printing.

Write a letter to the university president.

1) State whether or not you agree with this change.

   a. If you do agree with this change, explain why you think this change is good, but also acknowledge the opinions of students who might disagree with this change. In addition, offer alternative(s) that would help the university reduce printing costs as well as be more accepted by students who disagree.

   b. If you disagree with this change, explain why you think the change is not good, but also acknowledge the opinions of students who might agree with this change. In addition, offer alternative(s) that would help the university reduce printing costs as well as be more accepted by students.

Do you have any questions before we begin?

I will set a timer for 20 minutes, and when you have 5 minutes remaining, I will give you a warning.
APPENDIX D

Persuasive Analytic Rating Scale

<table>
<thead>
<tr>
<th>Criteria</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence of Prompt Components</td>
<td>4 components present: claim, reason(s) for, reason(s) against, alternative solutions</td>
<td>3 of 4 components present</td>
<td>2 of 4 components present</td>
<td>0 or 1 of 4 components present</td>
</tr>
<tr>
<td>Logical Reasoning</td>
<td>The logic of arguments is clear and persuasive</td>
<td>Can see attempt at logical presentation, but could be pushed a bit further; good attempt; arguments are mostly clear and persuasive</td>
<td>Can see attempt at logical presentation, but needs a lot more to make it a good argument; lacking in clarity and/or persuasiveness</td>
<td>No attempt at or idea how to present argument logically</td>
</tr>
<tr>
<td>Sense of Audience</td>
<td>0.0% sentences inappropriate for audience</td>
<td>0.1% to 5.0% inappropriate sentences</td>
<td>5.1% to 20.0% inappropriate sentences</td>
<td>&gt; 20.0% inappropriate sentences</td>
</tr>
<tr>
<td>Organization</td>
<td>Sample is perfectly organized and easy to follow; nothing could be done to improve organization</td>
<td>Organization is acceptable, making sample easy to follow; some minor improvements could be made</td>
<td>Some sections of the sample are misplaced and/or absent; organization is generally lacking, but may still be able to follow the sample</td>
<td>Most of sample or entire sample is difficult to follow due to problems with organization</td>
</tr>
<tr>
<td>Word Choice</td>
<td>Frequent use of sophisticated words that are used correctly</td>
<td>Some use of sophisticated words that are used correctly; no sophisticated words are used incorrectly</td>
<td>Some use of sophisticated words that are used correctly, but some sophisticated words are used incorrectly</td>
<td>No use of sophisticated words or sophisticated words are used incorrectly</td>
</tr>
<tr>
<td>Spelling</td>
<td>0.00% misspelled words (#misspelled words/# total words)</td>
<td>0.01% to 1.00% misspelled words</td>
<td>1.01% to 2.00% misspelled words</td>
<td>&gt; 2.00% misspelled words</td>
</tr>
<tr>
<td>Punctuation</td>
<td>0.0% ending punctuation errors and 0.0% other punctuation errors</td>
<td>0.0% ending punctuation errors and 0.1% to 10.0% other punctuation errors</td>
<td>0.0% to 10.0% ending errors and 10.1% to 20.0% other errors; OR 0.1% to 10.0% ending errors and 0.0% to 20.0% other errors</td>
<td>≥ 0.0% ending errors and &gt; 20.0% other errors; OR &gt;10.0% ending errors and ≥ 0.0% other errors</td>
</tr>
<tr>
<td>Sentence Complexity</td>
<td>≥ 80% complex sentences (# complex sentences/# total sentences)</td>
<td>60% to 79% complex sentences</td>
<td>40% to 59% complex sentences</td>
<td>&lt; 40% complex sentences</td>
</tr>
<tr>
<td>Sentence Soundness</td>
<td>≥ 90% sentences coded as 3</td>
<td>80% to 89% sentences coded as 3</td>
<td>70% to 79% sentences coded as 3</td>
<td>&lt; 70% sentences coded as 3</td>
</tr>
</tbody>
</table>
APPENDIX E

Demographic Questionnaire for Participants with WLDs
Katz (2009)

The ROWing Clinic
Student Survey Questionnaire

BACKGROUND INFORMATION

1. I was born in _________________________, _________________________.
   (month) (year)

2. I am a
   a. ___________ Male
   b. ___________ Female

3. Race/ethnicity
   a. ___________ American Indian or Alaskan Native
   b. ___________ Asian
   c. ___________ Black or African American
   d. ___________ Hispanic or Latino
   e. ___________ Native Hawaiian or Other Pacific Islander
   f. ___________ White or Caucasian

4. I am currently a
   a. ___________ Freshman
   b. ___________ Sophomore
   c. ___________ Junior
   d. ___________ Senior
   e. ___________ Non-degree student
   f. ___________ Master’s student
   g. ___________ Doctoral student
   h. ___________ Other ________________________________

5. I am currently working a paid job   Yes □   No □
   If “yes,” how many hours per week? ___________

6. I (check all that apply)
   a. ___________ am in the Honors program
   b. ___________ was in the Honors program
   c. ___________ am a UPAS student
   d. ___________ was a UPAS student
   e. ___________ None of the above
7. My major is currently  

______________________________________________________________

8. Before coming to BGSU, I (check all that apply)
   a. ___________ Attended another 4-year college or university
   b. ___________ Attended a community college
   c. ___________ Earned a Bachelor’s degree
   d. ___________ Earned an Associate’s degree/certificate/diploma
   e. ___________ Worked/had a job
   f. ___________ Completed high school
   g. ___________ Did not complete high school, but took my GED
   h. ___________ Other ____________________________________________

9. I am seeking an evaluation in this clinic because (check all that apply)
   a. ___________ I would like to obtain academic accommodations to help me in college
   b. ___________ I want to know if I have a disorder or disability
   c. ___________ I want to know what can be done to improve my skills
   d. ___________ I want to receive therapy services from this clinic
   e. ___________ Other ____________________________________________

HISTORY OF EDUCATIONAL DIFFICULTIES

10. I have previously been diagnosed with (check all that apply)
    a. ___________ Learning Disability
    b. ___________ Attention Deficit (with or without Hyperactivity) Disorder
    c. ___________ Language Disorder
    d. ___________ Dyslexia
    e. ___________ Reading Disorder
    f. ___________ Writing Disorder
    g. ___________ Math Disorder
    h. ___________ Autism Spectrum Disorder
    i. ___________ Pervasive Developmental Disorder (PDD)
    j. ___________ Other ____________________________________________
    k. ___________ None of the above

11. I received special education services during (check all that apply)
    a. ___________ Preschool
    b. ___________ Elementary school
    c. ___________ Middle school
    d. ___________ Junior high school
    e. ___________ High school
    f. ___________ Other __________________________________________________________________________
12. I received speech and/or language therapy during (check all that apply)
   a. _________ Preschool
   b. _________ Elementary school
   c. _________ Middle school
   d. _________ Junior high school
   e. _________ High school
   f. _________ Other _________________________________
   g. _________ None of the above

13. If you checked any responses for question #11, please respond to the following question:

   I received speech-language therapy for (check all that apply)
   a. _________ Articulation difficulties
   b. _________ Stuttering
   c. _________ Voice
   d. _________ Understanding language (e.g., vocabulary, grammar, directions)
   e. _________ Using language (e.g., word finding, grammar, producing sentences)
   f. _________ Communication in social situations (i.e., pragmatics)
   g. _________ Learning sound-letter correspondence
   h. _________ Learning how to read words
   i. _________ Learning how to spell
   j. _________ Reading comprehension
   k. _________ Writing
   l. _________ Other _________________________________

14. I received occupational therapy during (check all that apply)
   a. _________ Preschool
   b. _________ Elementary school
   c. _________ Middle school
   d. _________ Junior high school
   e. _________ High school
   f. _________ Other _________________________________
   g. _________ None of the above

15. Before coming to BGSU, I received professional academic help with
   a. _________ Reading
   b. _________ Writing
   c. _________ Language
   d. _________ Math
   e. _________ Other _________________________________
   f. _________ I did not receive professional academic help before coming to college
16. I am currently receiving accommodations through Disability Student Services (DSS) at BGSU.
   a. __________ Yes, currently
   b. __________ Not currently, but in the past
   c. __________ Not ever

17. If you check a or b for question # 15, please indicate the accommodations you did or do receive:
   _______________________________________________________________________

18. I tried to get academic help at BGSU from (check all that apply)
   a. __________ The Writing Center
   b. __________ Study Skills Center
   c. __________ Math and Statistics Center
   d. __________ A tutor ____________________________________________
   e. __________ Other ____________________________________________

19. I have taken
   a. __________ ACEN 1000 (or ACEN 100)
   b. __________ UNIV 1200 (or UNIV 120)
   c. __________ Neither of these classes

20. I tell my professors about my academic difficulties
   a. __________ All of the time
   b. __________ Most of the time
   c. __________ Some of the time
   d. __________ Rarely
   e. __________ Never

LISTENING, SPEAKING, READING, AND WRITING

21. I enjoy reading books that I choose (for fun)
   a. __________ All of the time
   b. __________ Most of the time
   c. __________ Some of the time
   d. __________ Rarely
   e. __________ Never

22. I enjoy my reading assignments for class
   a. __________ All of the time
   b. __________ Most of the time
   c. __________ Some of the time
   d. __________ Rarely
   e. __________ Never
23. I have trouble reading short words
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

24. I have trouble reading long words
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

25. I have trouble sounding out words
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

26. I have trouble understanding what I read
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

27. I have trouble with spelling
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

28. I have trouble writing papers for school
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never
29. I have trouble writing emails to friends
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

30. I have trouble writing emails to professors (or other more formal emails)
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

31. I have trouble organizing my ideas for writing
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

32. I have trouble with capitalization and/or punctuation
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

33. I have trouble with grammar (talking or writing)
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

34. Compared to other people my age, I have a good vocabulary
   a. _________ Strongly agree
   b. _________ Agree
   c. _________ Neither agree nor disagree
   d. _________ Disagree
   e. _________ Strongly disagree
35. The use of technology for my classes (e.g., Blackboard) is
   a. _________ Great for me
   b. _________ Good for me
   c. _________ Not good or bad for me
   d. _________ Somewhat difficult for me
   e. _________ Difficult for me

ATTENTION AND ORGANIZATION

36. I have trouble keeping myself organized
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

37. I have trouble paying attention during class
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

38. I have trouble paying attention while I am reading
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

39. I feel I have trouble interacting with my peers
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never

40. Compared to other people my age, I think I read slowly
   a. _________ All of the time
   b. _________ Most of the time
   c. _________ Some of the time
   d. _________ Rarely
   e. _________ Never
PERSONAL BELIEFS

41. I believe that it is too late for me to really benefit from services focused on improving my reading skills
   a. __________ Strongly agree
   b. __________ Agree
   c. __________ Neither agree nor disagree
   d. __________ Disagree
   e. __________ Strongly disagree

42. I believe that it is too late for me to really benefit from services focused on improving my writing skills
   a. __________ Strongly agree
   b. __________ Agree
   c. __________ Neither agree nor disagree
   d. __________ Disagree
   e. __________ Strongly disagree

43. I decided to go to college because [check all that apply]
   a. __________ my parents made me
   b. __________ for employment opportunities/opportunities for better money
   c. __________ as a natural next step
   d. __________ to improve my reading and/or writing skills
   e. __________ for social reasons (e.g., to make friends and have fun)
   f. __________ to pursue a professional career
   g. __________ I do not know why I decided to go to college
   h. __________ Other _________________________________________
   i. __________ I did not want to go to college

44. I would be willing to take a for-credit course at BGSU that would be focused on improving my reading comprehension skills.
   a. __________ Strongly agree
   b. __________ Agree
   c. __________ Neither agree nor disagree
   d. __________ Disagree
   e. __________ Strongly disagree

45. I would be willing to take a for-credit course at BGSU that would be focused on improving my writing skills.
   a. __________ Strongly agree
   b. __________ Agree
   c. __________ Neither agree nor disagree
   d. __________ Disagree
   e. __________ Strongly disagree
46. I would be willing to take a for-credit course at BGSU that would be focused on improving my organizational skills.
   a. _________ Strongly agree
   b. _________ Agree
   c. _________ Neither agree nor disagree
   d. _________ Disagree
   e. _________ Strongly disagree

47. I would be willing to participate in free weekly therapy/intervention sessions on campus that would be focused on improving my reading comprehension skills.
   a. _________ Strongly agree
   b. _________ Agree
   c. _________ Neither agree nor disagree
   d. _________ Disagree
   e. _________ Strongly disagree

48. I would be willing to participate in free weekly therapy/intervention sessions on campus that would be focused on improving my writing skills.
   a. _________ Strongly agree
   b. _________ Agree
   c. _________ Neither agree nor disagree
   d. _________ Disagree
   e. _________ Strongly disagree

49. I would be willing to participate in free weekly therapy/intervention sessions on campus that would be focused on improving my organizational skills.
   a. _________ Strongly agree
   b. _________ Agree
   c. _________ Neither agree nor disagree
   d. _________ Disagree
   e. _________ Strongly disagree

50. I would be willing to pay for weekly therapy/intervention sessions on campus that would improve my reading, writing, and/or organizational skills.
   a. _________ Strongly agree
   b. _________ Agree
   c. _________ Neither agree nor disagree
   d. _________ Disagree
   e. _________ Strongly disagree
51. If you agree or strongly agree to the above question(s),

I would be willing to participate in weekly therapy/intervention sessions on campus that would be focused on improving my reading comprehension, writing, and/or organizational skills if (check all that apply)

a. __________ the charge for an hourly session (1-2 times per week) were $50-$60
b. __________ the charge for an hourly session (1-2 times per week) were $30-$49
c. __________ the charge for an hourly session (1-2 times per week) were $15-$29
d. __________ the charge for an hourly session (1-2 times per week) were $10-$14
e. __________ the charge for an hourly session (1-2 times per week) were $5-$10
f. __________ the charge for an hourly session (1-2 times per week) were less than $5
### APPENDIX F

**Demographic Questionnaire for Typical Participants**

<table>
<thead>
<tr>
<th>Subject # __________________________</th>
</tr>
</thead>
</table>

| 1. Gender: | 2. Age: ______ |
| __ Male | __ Female |

| 3. Year in School: |
| __ Freshman | __ Junior |
| __ Sophomore | __ Senior |
| __ Non-degree Student | __ Other: ____________________ |

<table>
<thead>
<tr>
<th>4. Major: __________________________</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>5. Current Overall GPA: ______</th>
</tr>
</thead>
</table>

| 6. ACT/SAT Score: |
| ACT: ______ Overall SAT: ______ |
| Other: ______ |

| 7. I (check all that apply) |
| __ am in the Honors program |
| __ was in the Honors program |
| __ am a UPAS student |
| __ was a UPAS student |
| __ None of the above |

| 8. Race/Ethnicity: |
| __ American Indian/Native American |
| __ Asian |
| __ Black/African American |
| __ Hispanic/Latino |
| __ White/Caucasian |
| __ Pacific Islander |
| __ Other: ____________________ |
| __ Prefer not to disclose |

| 9. I have been previously diagnosed with: (check all that apply) |
| __ Learning Disability |
| __ Attention Deficit Disorder or Attention Deficit/Hyperactivity Disorder |
| __ Language Disorder |
| __ Dyslexia |
| __ Reading Disorder |
| __ Writing Disorder |
| __ Math Disorder |
| __ Autism |
| __ Asperger Syndrome |
| __ Pervasive Developmental Disorder (PDD) |
| __ Other: ____________________ |
| __ None of the above |

| 10. I have received accommodations through Disability Student Services (DSS) at BGSU: |
| __ Yes, currently |
| __ Not currently, but I have in the past |
| __ Not ever |

| 11. I received special education services during: (check all that apply) |
| __ Preschool |
| __ Elementary school |
| __ Middle school |
| __ Junior high school |
| __ High school |
| __ Other: ____________________ |
| __ None of the above |
12. Before coming to BGSU, I received professional academic help with:
___ Reading  
___ Writing  
___ Language  
___ Math  
___ Other: ____________________
___ I did not receive professional academic help before coming to college

13. I enjoy reading books that I choose (for fun):
___ All of the time  
___ Most of the time  
___ Some of the time  
___ Rarely  
___ Never

14. I enjoy reading assignments for class:
___ All of the time  
___ Most of the time  
___ Some of the time  
___ Rarely  
___ Never

15. I have difficulty understanding what I read
___ All of the time  
___ Most of the time  
___ Some of the time  
___ Rarely  
___ Never

16. I enjoy writing (for fun):
___ All of the time  
___ Most of the time  
___ Some of the time  
___ Rarely  
___ Never

17. I enjoy writing assignments for class:
___ All of the time  
___ Most of the time  
___ Some of the time  
___ Rarely  
___ Never

18. I am a(n) _____ writer:
___ Excellent  
___ Good  
___ Adequate  
___ Weak  
___ Very Poor

19. I have taken the following course (check all that apply):
___ GSW 1100  
___ GSW 1110  
___ GSW 1120  
___ None of the above

20. I decided to go to college because:
(check all that apply)
___ My parents made me  
___ For employment opportunities/ opportunities for more money  
___ As a natural next step  
___ To improve my reading and/or writing skills  
___ For social reasons (e.g., to make friends and have fun)  
___ To pursue a professional career  
___ I do not know why I decided to go to college  
___ Other: ____________________  
___ I did not want to go to college
APPENDIX G

Informed Consent

“Profiles of Undergraduate and Graduate Students with Reading, Writing, and/or Organizational Difficulties”

My name is Lauren Katz, and I am an Assistant Professor in the Department of Communication Sciences and Disorders at Bowling Green State University (BGSU). I am also Director of the ROWing Clinic, which was developed to help college students who are struggling with reading, writing, and/or organizational skills. You are here today because you have indicated interest in participating in this research study. Specifically, we are interested in learning more about the writing skills of typical college students. For our purposes, this is important because we are trying to help college students who might be struggling with learning disabilities, language disorders, reading disorders, and/or writing disorders, and we are not sure if what they know is or is not comparable to what their typical peers know. We hope that your participation in this study will help us better identify difficulties in and provide help to students who are struggling with writing.

If you decide to participate in this study, my research assistants and/or I will be administering three separate measures, which will take approximately 45 minutes in total. One measure is a short questionnaire that asks you to respond to demographic questions and questions about your educational background. Another measure will require that you point to pictures that best represent words spoken by the examiner, and the third measure will ask you to write a persuasive letter in response to a prompt.

I anticipate no risk to you as a result of your participation in this study. There are also no costs to you for participating in this study. Also, your decision to participate (or not to participate) will have no impact on your relationship to BGSU or your relationship with the examiners involved in this assessment. You may receive credit for participating in this study as part of your class; however, your participation in this study is voluntary. You may choose to not participate or withdraw from the study at any time.

We hope that your participation will eventually benefit other undergraduate and graduate students in college and university settings, and also the fields of speech-language pathology, special education, general education, and higher education.

The information that you provide on the survey questionnaire, as well as your testing results, will be anonymous. In fact, your name will not appear on any of the tests or documents used today, so we will have no record of your identity to connect with your results. A participant code will be assigned in place of your name to keep your data anonymous and private. Also, your name will not be used in any publications related to this research. All data will be kept in a filing cabinet in my locked research laboratory at BGSU and will not be available to anyone who is not directly involved in this study.
Once the study is completed, if you are interested, I would be happy to share our results with you. If you have any questions at any point prior to, during or following the study, please contact me:

Lauren A. Katz  
250 Health Center Building  
Bowling Green State University  
Bowling Green, OH 43403  
419-372-7165  
katzla@bgsu.edu

If you have any concerns of questions about your treatment as a participant in this study, please call or write:

Chair, Human Subjects Review Board  
University Hall 309A  
Bowling Green State University  
Bowling Green, OH 43403  
419-372-7716  
email: hsrb@bgnet.bgsu.edu
APPENDIX H

Instructions for Holistic Scoring

Today, you will be reading writing samples collected from both typical college students and college students with written language disorders. Your job while reading will be to attend to the overall quality of each writing sample. More specifically, you will be trying to decide whether you feel each sample is of excellent, good, fair, or poor quality. During this task, you should focus on the particular features of each sample that you feel make it strong or weak and use your gut instinct to determine a score that you feel best reflects the overall quality of each sample.

To complete this task, you will need to:

1. Read the persuasive writing prompt that the participants were asked to respond to for this study. This will give you an idea of what the expectations were for this task so that you know what you should be looking for as you score the samples.

2. Read through the first twelve writing samples in your stack without scoring them. This will give you the opportunity to see the variability present in the samples so that you can better determine how you should score them.

3. Reread each of the first twelve samples individually, using the provided highlighters to highlight sections of each sample that you find to be especially strong (yellow) or weak (pink).

4. After reading each sample, immediately write your initials, the sample number, and your score on one of the provided data sheets using the following rubric:

<table>
<thead>
<tr>
<th>Excellent</th>
<th>Good</th>
<th>Fair</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

5. Once you provide a score for a sample, provide comments in the space given on the data sheet to explain why you selected that particular score for that sample.

6. Continue this process with the remaining fifteen samples. Be sure to highlight, score, and comment for each sample.

While scoring these samples, please keep in mind that you are to use your gut. Therefore, do not spend too much time on each sample. At the same time, do not rush through them, but instead make sure that you read them carefully enough to be able to provide an accurate score. In addition, after the first twelve samples, you each have fifteen different writing samples in your stack, meaning that you should be scoring at a different pace. In other words, do not worry about trying to match the pace of the others in the room, as the samples they are scoring may be
significantly shorter or longer than yours. Finally, please refrain from talking with the other raters about the samples or how you are scoring the samples until all analyses have been completed on the writing samples. This will ensure that you each continue to score the samples in the same manner throughout the entire analysis process.
APPENDIX I
Instructions for Analytic Rating Scale Scoring

For this task, you will be reading the same writing samples that you scored during the holistic scoring process. This time, instead of looking at overall quality, you will be examining each writing sample for several specific features of writing. During this task, you can refer to the rating scale, rating scale guidelines, and examples of scored samples to help you make a decision about the proper scores to give each sample. If necessary, you can also refer to outside resources (e.g., a dictionary) to help you through this process.

To complete this task, you will need to:

1) Re-read the persuasive writing prompt to remind you of the expectations for this task.
2) Score all 27 of your samples using one rating scale category at a time (e.g., score all samples for Presence of Prompt Components, then score all samples for Logical Reasoning...).
3) Try to score all samples for a single rating scale category during a short time span (e.g., during one sitting or at least during the same day).
4) Read through each sample as many times you need to provide an accurate score, but be sure to read through each sample at least once for each score that you give.
5) If you are required to record codes on the samples for the rating scale category you are working on (i.e., Sense of Audience, Spelling, Punctuation, Sentence Complexity, and Sentence Soundness), be sure to do so appropriately. Even when you are not required to write codes on the samples, feel free to write on them as much as you need to help you through the process.
6) Once you have completed scoring or coding a sample, record your score or frequency count on the provided data sheet.
7) On the rating scale, keep track of the date that you score for each component of the rating scale.
8) Keep track of anything that you find to be confusing or frustrating during the scoring and coding process. You can record your frustrations on the back of the rating scale or on a separate sheet of paper.

While scoring these writing samples, please keep in mind that you are trying to provide scores and codes as accurately as you can given the guidelines that you have been provided. Therefore, be sure that you refer back to the rating scale, rating scale guidelines, and scoring examples as much as you need to help ensure that you are scoring and coding the samples correctly. At the same time, be sure that you do not rush through the scoring process, but instead take as much time as you need with each sample to provide accurate scores and codes. Finally, please refrain from talking with the other raters about the samples or how you are scoring the samples until
after all analyses have been completed on the writing samples. This will ensure that you each continue to score the samples in the same manner throughout the entire analysis process.
APPENDIX J

Guidelines for Using Analytic Rating Scale

**Presence of Prompt Components (use unsegmented samples)**

To score *Presence of Prompt Components*, you will be counting how many of the following prompt components are present:

- **Claim.** The “claim” states whether or not the writer agrees or disagrees with the change in the printing policy. In order to receive credit for writing a claim, the writer must have explicitly stated his/her position. Implicit statements of the claim do not receive credit.

- **Reason(s) for.** The “reason(s) for” provide reasons for the writer’s claim.
  - If the writer is for the printing policy, you will look for him/her to provide reasons why he/she thinks that the printing policy is a good idea.
  - If the writer is against the printing policy, you will look for him/her to provide reasons why he/she thinks that the printing policy is a bad idea. The writer can restate the reason(s) given in the prompt (i.e., “to help the university save money” or “to make students more mindful of unnecessary printing”) and/or provide reason(s) of his/her own (e.g., to be more environmentally-friendly) to receive credit for “reason(s) for”.

- **Reason(s) against.** The “reason(s) against” provide reasons against the writer’s claim, or acknowledges the feelings of the writer’s opponents.
  - If the writer is for the printing policy, you would look for him/her to provide reasons why he/she thinks his/her opposition would be against the printing policy. This could include a simple acknowledgement of the other side’s feelings and/or a detailed description of reasons why his/her opponents would be against the printing policy.
  - If the writer is against the printing policy, you would look for him/her to provide reasons why he/she thinks his/her opposition would be for the printing policy. This could include a simple acknowledgement of the other side’s feelings and/or a detailed description of reasons why his/her opponents would be for the printing policy. The writer can restate the reason(s) given in the prompt (i.e., “to help the university save money” or “to make students more mindful of unnecessary printing”) and/or provide reason(s) of his/her own (e.g., to be more environmentally-friendly) to receive credit for “reason(s) against”.

- **Solution(s)/Suggestion(s).** The “solution(s)/suggestion(s)” require the writer to offer at least one alternative solution to the printing policy that would both help the university reduce printing costs and be more accepted by students. This solution must both help the university save money and prevent students from paying an excessive amount of money to print. Therefore, solutions that are offered for other reasons (e.g., to be more environmentally-friendly) would not receive credit.
**Logical Reasoning (use unsegmented samples)**

To score for *Logical Reasoning*, you need to determine whether or not the arguments presented by the writer are clear and/or persuasive. To judge clarity, you might ask yourself the following questions:

- Do I understand what the writer is trying to say?
- Do I have to read sections of the text multiple times to understand what the writer is trying to say?
- Do the arguments that the writer provides make sense given the context of the prompt?

When considering persuasiveness, you might ask:

- Does the writer convince me to take his/her side?
- Does the writer provide enough support to convince me to take his/her side?
- Would the solution(s) that the writer suggests truly save the university money and make students happy?

**Sense of Audience (use segmented samples)**

For *Sense of Audience*, you will be required to identify the number of sentences in each sample that are inappropriate for an audience of a university president. You will flag each sentence that you find to be inappropriate. While coding, you should not penalize writers’ for the use of contractions in their writing. However, some features that would deem a sentence inappropriate include:

- Use of slang
- Conversational/oral nature
- Rudeness

**Scoring example (you do not need to explain your reasoning):**

The change is a terrible idea, Mrs. President and I tell you why. *Inappropriate* (rude; telling the president that her idea is “terrible” is insulting and calls her ability to do her job into question; more appropriate statement would be “I disagree with the change...”)

Not only did you raise our tuition this year, you also make us pay more money for intermural sports. *Inappropriate* (the use of “not only did you...” is slightly rude and suggests that the president is the only person who makes decisions about increasing prices on campus)

But you seem to believe that buying Huge Flat screen TVs for the sub line and Steak Escape was needed. *Inappropriate* (rude; “seem to believe” suggests that the president’s beliefs are incorrect and seems to question the president’s ability to successfully do her job)
Even though 1 white board (which would be thousand of dollars cheaper) could provide the same effect.

Don’t lie to us and say that us paying for our prints would make us “mindful” of what we print. **Inappropriate** (rude; accusing university president of lying to students)

The fact is you’re just money hungry. **Inappropriate** (rude; insulting the university president)

Step down. **Inappropriate** (rude; it would be inappropriate for a student to say this to a university president)

**Organization (use unsegmented samples)**

When scoring for Organization, you will need to consider the flow and organization within sentences, paragraphs, and the entire sample. The sample does **not** need to be written in paragraphs or include a greeting and salutation to receive full credit for Organization. However, you could penalize samples for the following:

- Lack of clear introductory and/or concluding elements (Example 2)
- Lack of flow within sentences, between sentences, and/or between paragraphs (Example 2)
- When information would be better presented in a different part of the text (Example 2)
  - Could include problems at the sentence, paragraph, and/or sample level
- Inappropriate breaks in paragraphs (Examples 1 and 3)
- Lack of appropriate paragraph structure (i.e., topic and supporting sentences) (Example 3)
  - Could include missing topic and/or supporting sentences in the sample or instances where the topic sentence does not accurately represent the information presented in the paragraph

While scoring for Organization, it will be important to note that scores of “4” and “1” are likely to be rare, as it is unlikely that you will find many samples with perfect or disastrous organization.

**Example 1:**

Thank you President for listening to my viewpoint and considering my alternatives to saving money and our planet.

If you have any questions pleas feel free to contact me. I would be honored to assist in this transition.

**Example 1 shows the last two paragraphs of a sample, which would be better combined into a single paragraph.**
Example 2:

If we pay for printing as students, we would refuse to print off our work. We use the printer everyday to print off worksheets and other forms of homework. We need the money already to buy our food, books, and other school supplies. I do not think you really would save money, and more students would become broke during the semester.

Not every student on campus prints unnecessary copies, so It would not be fair to those who do need the printers for homework, and schedules. This would not be a good change for the University if students are made to pay for copies. This would turn out to be a catrostrophe. People need to print off work frequently for classes, so it would not be fair for Students

Example 2 shows an entire sample that is missing clear introductory and concluding elements. Additionally, some of the information would be better presented in a different part of the text. For example, this writer did not clearly state his/her claim until the end of the sample, and it would have made more sense to have the claim at the beginning of the sample. Finally, there seems to be a lack of smooth flow between sentences and paragraphs, as few transition words and phrases are used to connect sentences and paragraphs. The lack of transitions gives this sample a choppy feel.

Example 3:

I have three alternatives to offer that would help the university reduce printing costs. My three alternatives are that for every page a student pays for they get one free. I call this BOGO buy one get one. This alternative may help students feel that the university did not put all the costs on the student.

Second, I would in printing costs in the student general fees. In turn a printing stipend would be available for each student. Each student would have an allotment for up to five years during their tenured.

Lastly printing should be made available at a fraction of the costs. This alternative will aid students in financing their printing while maintaining the resources they need to be successful student. And the university would still save money.

Example 3 shows three of the body paragraphs from a writing sample that have problems with appropriate paragraph breaks and paragraph structure. In the first paragraph, the topic sentence suggests that three alternatives will be presented in the paragraph. However, only one alternative is presented in the first paragraph. Therefore, all three of these paragraphs should be combined to fulfill the promise of the topic sentence of the first paragraph. Another option would be to change the topic sentence of the first paragraph to reflect the discussion of the first alternative (e.g., “The first alternative suggestion that I have that would help the
university reduce printing costs is…”). This example could also be considered to have problems with inappropriate paragraph structure because the topic sentence of the first paragraph does not match up to the information presenting in the supporting sentences of that paragraph.

**Word Choice (use unsegmented samples)**

To score for **Word Choice**, you should look for sophisticated words that are used correctly. Sophisticated words might include:

- Words that are morphologically complex, long, and/or multisyllabic (e.g., “informational” or “environmentally”)
- Low frequency words (e.g., “bleak” or “guise”)
- Words that are more advanced than typical college vocabulary

It is important to remember that not all words that are morphologically complex, long, and/or multisyllabic are sophisticated, and that some words that are not morphologically complex, long, and/or multisyllabic are sophisticated. Therefore, you will need to use your best judgment to determine whether or not you believe a word should be considered sophisticated. While looking for sophisticated words, you should **not** count the following words as sophisticated because they were presented to the writers in the writing prompt:

- Mindful
- Unnecessary
- Alternative or alternatives
- Disagree
- Acknowledge

Additionally, credit should **not** be given to morphologically complex words that include morphemes that change tense (e.g., present progressive –ing or past tense –ed), number (e.g., plural-s), or possession (e.g., possessive ‘s). Any sophisticated words that are used incorrectly should **not** receive credit (e.g., “Thank you for your concentration[instead of consideration]”). Finally, sophisticated words that are used multiple times in the same sample should **not** receive high credit for sophistication, as the overuse of sophisticated words makes them less sophisticated.

**Spelling (use segmented samples)**

To code for **Spelling**, you will read through each sample and code any words that are misspelled. To code spelling errors, you will circle them and write “SPL” above them to indicate that a spelling error is present. Words that should be coded include:

- Misspelled words (e.g., “presedent” for “president”)
- Homophones that are used incorrectly (e.g., “their” for “there”, or “too” for “two” or “to”
• Contractions that are missing an apostrophe or demonstrate incorrect apostrophe usage (e.g., “its” for “it’s”); apostrophe errors of possession should not be counted as spelling errors.

However, words that contain morphological errors (e.g., “print” for “prints” or “printing”), capitalization errors (e.g., “if you Have any Questions, please feel free to contact me.”), spacing errors (e.g., “over all”), or the incorrect usage of hyphens (e.g., “re-sources” or “visually impaired”) should not be coded as misspelled.

Coding example:

SPL

The change is a [errable] idea…

**Punctuation (use unsegmented samples)**

To score for *Punctuation*, you will be required to read through each sample and identify errors of ending punctuation or other punctuation in each sentence. You should code each sample that has an error of punctuation. Ending punctuation errors include:

• Incorrect ending punctuation (e.g., use of a period for a question mark)
• Run-on sentences that require more than the addition of a semi-colon to be fixed
  • Example: “The printing policy would be bad for the university because if you make students pay too much to print a lot of them could spend their money on buying their own printers and paper instead of spending it on printing at the university which might result in more paper being used and cost the university more money because the electricity bill would be higher from all of the students using printers in their dorm rooms.”

Other punctuation errors could include:

• Omitted or misused apostrophes in instances of possession (e.g., “the student’s of BGSU” or “When peoples finances are involved…”)
  • Apostrophe errors with contractions should not be counted as punctuation errors
• Omitted or misused commas
• Omitted or misused semicolons
• Omitted or misused quotation marks
• Omitted or misused periods in abbreviations (e.g., “Dr” for “Dr.”)
• Run-on sentences that can be fixed with the use of a semi-colon
  • Example: “The new printing policy is a good idea because it will save the university money it might even produce enough extra money to help the university buy other supplies for the school.”
Sentence fragments should not be penalized for ending punctuation, and should only be penalized for other punctuation errors if the issue is something other than the fact that it is a sentence fragment. Additionally, writers should not be penalized for using or not using punctuation in non-obligatory contexts, such as the following:

- **Comma to separate the final item in a series of three or more items**
  - Acceptable to include or omit final comma
  - **Example:** both “printer, paper, and ink” and “printer, paper and ink” are acceptable

- **Comma to separate two independent clauses**
  - Acceptable to omit comma if both clauses are short and balanced
  - **Example:** both “Printing is necessary, so it should be free.” and “Printing is necessary so it should be free.” are acceptable

- **Comma to set off introductory elements**
  - Acceptable to omit a comma after brief introductory element if this does not result in confusion for readers
  - **Example:** both “Last year we didn’t have to pay for printing.” and “Last year, we didn’t have to pay for printing.” are acceptable

- **Commas to set off a parenthetical element (i.e., part of a sentence that could be removed without changing the essential meaning of the sentence)**
  - Acceptable to omit commas when the parenthetical element is closely related to the surrounding text and the omission of the commas does not confuse readers
  - **Example:** both “My psychology professor Dr. Black makes us print materials every week.” and “My psychology professor, Dr. Black, makes us print materials every week.” are both acceptable

- **Comma to separate coordinate adjectives (e.g., “cheap, effective plan” or “new, energy-efficient printers”)**
  - Commas should be placed between adjectives that you could put “and” in between (e.g., “cheap and environmentally-friendly idea”), but do not need to be used between adjectives that you could not put “and” in between (e.g., “small personal printers”)
  - **Example:** “cheap environmentally-friendly idea” and “small, personal printers” would be incorrect

**Sentence Complexity (use segmented samples)**

To score for Sentence Complexity, you will be required to code each sentence that is either complex or compound-complex as complex. Sentences will be defined by the writer’s punctuation. The following should not be coded as complex sentences:

- Run-on sentences
- Use of infinitives
  - “Trees are chopped down to provide materials to make paper.” would not be complex
Trees are chopped down to provide materials to make paper, which is harmful to the environment.” would be complex

Scoring example:

A third reason that I disagree with the university’s change is because students rely on printed materials to do well in their classes. CPX

I have never been able to read journal articles or book chapters off of the computer and understand what I am reading. CPX

Instead, I need a hard copy so that I can highlight and jot down notes to enhance my understanding of the material. CPX

Therefore, requiring me to pay money to print materials for class would be detrimental to my academic success.

Sentence Soundness (use segmented samples)

To score for Sentence Soundness, you will be scoring each sentence in the sample using the following scale:

<table>
<thead>
<tr>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Sentence is completely sound or acceptable as written</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Sentence may not be written how you would write it, but changes are not necessary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• You are completely able to understand the point of the sentence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Some errors or awkwardness may be present, but sentence is easily fixable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Despite errors/awkwardness, you are still able to understand the point of the sentence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Frequent errors or instances of awkwardness are present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Sentence needs lots of work to be fixed (this would include a run-on sentence that requires many changes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Sentence is very confusing and you are unable to grasp the point or you must read the sentence several times to grasp the point</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Errors that you would mark down for include:

• Sentence fragments (e.g., “If you have any questions.” or “Which would help the university reduce printing costs.”)

• Run-on sentences (e.g., “On a less positive note, if the university makes students pay to print, if it is expensive enough, a lot more paper could be used with personal printers, which might use more paper overall on campus, and with all those personal printers running it may
increase the electricity bill enough that the whole goal of saving money would be cancelled out.

- Multiple redundancies (e.g., “If you have any interest in any of my ideas or any concerns about any of my opinions, please feel free to contact me.” – “any” was used too many times)
- Syntactic errors (e.g., “I writing to you in regards of the new printing policy.” – should be “I am writing…”)
- Semantic errors/incorrect use of words (e.g., “Students will be made that their being charged to print.” or “Thanks for taking my idea into concentration.” – “their” should be “they’re” and “concentration” should be “consideration”)
- Morphological errors (e.g., “I will offer you three alternative to help reduce printing costs and make both side happy.” – both “alternative” and “side” should be plural)
- Flawed logic (e.g., “Printing costs are expensive to our pockets but mostly to our environment.” Printing costs can be expensive to pockets, but not to the environment. In fact, printing costs are not even harmful to our environment; if they are high enough, they may even be beneficial to the environment. However, printing can be harmful to the environment.)
- Weirdness or awkwardness (e.g., “I am writing to you in regards to the printing policy.” – “in regards to” is awkward and would be better written as “in regard to”; “A time period could be inplaced to allow BGSU to gauge the paper output.” – “inplaced” does not make sense and should be “put in place”; “If I must pay to print, then I might be inclined to only extract materials directly from my computer screen.” – “extract materials” is awkward and might make more sense as “read materials”)
- Issues of sentence clarity (e.g., “The new printing policy might even produce enough extra money to help fund other and supplies that the school uses.” – It is unclear what is meant by “other and supplies” in this sentence.)

While scoring sentences, you should consider how the sentence makes sense given the previous and subsequent sentence. However, you should not penalize writers for starting a sentence with a conjunction if the use of a conjunction is appropriate in the given context. Additionally, you should ignore any errors of punctuation (e.g., “I call this idea BOGO buy one get one.” – missing hyphen or colon between “BOGO” and “buy one get one”) and spelling while scoring for Sentence Soundness, as these errors will be captured by the Punctuation and Spelling scores.
APPENDIX K

Guidelines for Identifying Complex Sentences

**Complex Sentence**: sentence containing one independent clause and at least one dependent clause

**Independent Clause**: the main clause in a sentence that can stand alone as a sentence and does not depend on any other clause in the sentence for full meaning; independent clauses can be connected within a single sentence by a coordinating conjunction

**Dependent Clause**: the subordinate clause in a sentence that contains a subject and a predicate, but cannot stand alone; must be attached to an independent clause with either a subordinating conjunction or relative pronoun; adds information to an independent clause, but can be removed without changing the meaning of the independent clause

**Coordinating Conjunctions**:

- And
- But
- Or
- So (when *so* can be read as *so that*, it is **NOT** considered a coordinating conjunction)

**Subordinating Conjunctions**:

- After
- Although
- As
- As if
- As long as
- As much as
- As soon as
- As though
- As well as
- Because
- Before
- By the time
- Even
- Even if
- Even though
- Every time
- How
- If
- If only
- If when
- If then
- Inasmuch
- In case
- In order that
- In the event that
- Just as
- Just in case
- Lest
- Now
- Now since
- Now that
- Now when
- Once
- Only if
- Provided
- Provided that
- Rather than
- Since
- So that
- Supposing
- Than
- That
- The first time
- Though
- Til
- Unless
- Until
- When
- Whenever
- Where
- Whereas
- Where if
- Wherever
- Whether
- (whether or not)
- Which
- While
- Who
- Whoever
- Why
Tips for Identifying Complex Sentences:

1. Look for subordinating conjunctions
   a. The presence of a subordinating conjunction usually indicates that a sentence is complex
   b. Be sure to check for the implied “that” (e.g., “I don’t think [that] you should change the printing policy…” – Even though the “that” is omitted from this sentence, it is still complex because of the implied “that”)

I suggest that you stick to step 1 when trying to identify complex sentences. However, if you still are not sure whether or not a sentence is complex after step 1, you can try step 2:

2. Locate verbs or verb units (i.e., group of verbs that works together) and determine whether or not they are in separate clauses
   a. If multiple verbs/verb units, determine whether or not they are in separate clauses
   b. If in separate clauses, determine the types of clauses that they are in – dependent and/or independent
   c. If at least one independent and one dependent clause is present, then it is a complex sentence

Examples:

I believe that the new printing policy is a bad idea.

1. Subordinating conjunction = “that”
2. # of verbs/verb units = 2 (“believe” and “is”)
   a. 2 separate clauses
   b. “I believe” is an independent clause; “that the new printing policy is a bad idea” is a dependent clause
   c. 1 independent clause + 1 dependent clause = complex sentence

If students are forced to pay for printing, they will refuse to print off their work.

1. Subordinating conjunction = “if”
2. # of verbs/verb units units = 2 (“are forced to pay” and “will refuse to print off”)
   a. 2 separate clauses
   b. “they will refuse to print off their work” is an independent clause; “If students are forced to pay for printing” is a dependent clause
   c. 1 independent clause + 1 dependent clause = complex sentence
I do not think you really would save money with the new printing policy, and more students would become broke during the semester.

1. Subordinating conjunction = implied “that” (i.e., “I do not think [that]…”)
2. # of verbs/verb units = 3 (“think”, “would save”, and “would become”)
   a. 3 separate clauses
   b. “I do not think” and “more students become broke during the semester” are independent clauses; “[that] you would really save money with the new printing policy” is a dependent clause
   c. 2 independent clauses + 1 dependent clause = complex sentence
# APPENDIX L

Examples of Comment Themes and Subthemes

<table>
<thead>
<tr>
<th>Theme/Subtheme</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linguistic</strong></td>
<td></td>
</tr>
</tbody>
</table>
| Sentence Structure  | -Incomplete sentences  
                    -Many run-on sentences  
                    -Choppy sentences     |
| Wording Issues      | -Oddly worded  
                    -Some confusing wording  
                    -Missing words        |
| Clarity Issues      | +Reader gets point even with awkward sentences  
                    -Some unclear sentences  
                    -Sentences were confusing  
                    -Unclear sentences makes it difficult to read |
| Grammar/Syntax      | +Free of grammatical errors  
                    +Few grammatical errors  
                    -Grammar problems  
                    -Some grammatical errors  
                    -Errors of syntax    |
| **Mechanics**       |                                                                          |
| Spelling            | -Spelling errors  
                    -Spelling errors makes it difficult to read  
                    -Few spelling errors |
| Punctuation         | -Insufficient use of punctuation  
                    -Needs commas, etc.  
                    -Punctuation errors |
| **Prompt Components/Content** |                                                    |
| Presence/Absence of All Prompt Components | +Addressed all areas  
                      +Covers all areas of the prompt  
                      +4 prompt components  
                      -Missing content  
                      -3 prompt components |
| Level of Support for All Prompt Components | +Good use of examples  
                      +Used personal examples  
                      -Could use more supporting details |
| Quality of All Prompt Components     | +Persuasive  
                      +Good arguments  
                      +Addresses each point well  
                      -Positions may not be entirely true  
                      -Not a very strong argument  
                      -Repeated prompt as often as stating own thoughts |
| **Presence/Absence of Claim/Reasons** | +Addresses opposite viewpoint  
+Stated opinion  
+Offered both sides of the opinion  
-Did not address other opinion  
-Not sure if the agree or disagree with the change |
| **Level of Support for Claim/Reasons** | +Personal experience to support opinion  
+Supported their positions  
+Addresses both opinions with good examples  
-Could expand more on why they agree  
-Didn’t give details on opposing opinion |
| **Quality of Claim/Reasons** | +Strong opinion  
+Thoughtful discussion of both viewpoints  
+Discusses both viewpoints with much thought  
-Discussion of opposing viewpoint not explained clearly  
-Does not provide thoughtful reasoning why students may be for the change |
| **Presence/Absence Alternative Solutions** | +Alternatives were given  
+Offered alternatives  
-Did not give alternatives  
-Did not provide alternatives that would help the university reduce printing costs |
| **Level of Support for Alternative Solutions** | +Some explanations of alternatives  
-Should have described the alternative more |
| **Quality of Alternative Solutions** | +Good and reasonable alternatives  
+Gives thoughtful alternatives  
+Great alternative was addressed  
-Alternatives were not clearly explained  
-Not a convincing alternative to reducing costs |
| **Overall Quality of Samples** |  |
| **Clarity/Ease of Reading** | +Clearly written  
+Easy to follow  
+Easy to read  
-Found it difficult to read  
-Very hard to understand and follow |
| **Overall Quality** | +Well written  
+Good essay  
+Enjoyable to read  
-Very basic  
-Casually written  
-Not great writing |
<table>
<thead>
<tr>
<th></th>
<th>+Concise</th>
<th>-Short</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Length</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Overall Structure/Organization</strong></td>
<td>+Organized</td>
<td>- Doesn’t flow well/poor transitions</td>
</tr>
<tr>
<td></td>
<td>+Very well organized</td>
<td>- Very choppy and not cohesive</td>
</tr>
<tr>
<td><strong>Specific Elements of Structure/Organization</strong></td>
<td>+Strong concluding sentence</td>
<td>- First paragraph not integrated well</td>
</tr>
<tr>
<td></td>
<td>+Great intro, got my attention</td>
<td></td>
</tr>
</tbody>
</table>
Thank you for your submission of Continuing Review/Progress Report materials for this project. The Bowling Green State University Human Subjects Review Board has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a project design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

The final approved version of the consent document(s) is available as a published Board Document in the Review Details page. You must use the approved version of the consent document when obtaining consent from participants. Informed consent must continue throughout the project via a dialogue between the researcher and research participant. Federal regulations require that each participant receives a copy of the consent document.

Please note that you are responsible to conduct the study as approved by the HSRB. If you seek to make any changes in your project activities or procedures, those modifications must be approved by this committee prior to initiation. Please use the modification request form for this procedure.

You have been approved to enroll 210 participants. If you wish to enroll additional participants you must seek approval from the HSRB.

All UNANTICIPATED PROBLEMS involving risks to subjects or others and SERIOUS and UNEXPECTED adverse events must be reported promptly to this office. All NON-COMPLIANCE issues or COMPLAINTS regarding this project must also be reported promptly to this office.

This approval expires on August 18, 2013. You will receive a continuing review notice before your project expires. If you wish to continue your work after the expiration date, your documentation for continuing review must be received with sufficient time for review and continued approval before the expiration date.

Good luck with your work. If you have any questions, please contact the Office of Research Compliance at 419-372-7716 or hsrb@bgsu.edu. Please include your project title and reference number in all correspondence regarding this project.