Leveraging Historical Thinking Heuristics as Warrants
in Historical Argumentative Writing

BY

RYAN McCARTY
B.A., Indiana University-Bloomington, 1998
M.Ed., National-Louis University, 2001

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Curriculum and Instruction
in the Graduate College of the
University of Illinois at Chicago, 2016

Chicago, Illinois

Defense Committee:

Cynthia Shanahan, Chair and Advisor
Kimberly Lawless, Department of Educational Psychology
Taffy Raphael
Timothy Shanahan
Anne Britt, University of Northern Illinois
This dissertation is dedicated to my wife Goldie McCarty. She was the first one to encourage me to pursue my PhD and sacrificed tremendously to ensure I completed it. This accomplishment is as much hers as it is mine.
ACKNOWLEDGEMENTS

I would first like to thank my advisor and dissertation committee chair Cynthia Shanahan for her support throughout this process. Your pragmatic advice around study design and data analysis was essential in the successful completion of my research. Your positive attitude and encouragement is greatly appreciated. I would also like to thank my committee members Kimberly Lawless, Taffy Raphael, Timothy Shanahan, and Anne Britt for their expert insight and guidance. Your research serves as an exemplar for my own work. I am honored to have such brilliant folks on my committee and I look forward to collaborating with you in the future.

I would like to thank Michael Manderino, now at Northern Illinois University, for his support in all areas of this research endeavor. Mike, you showed me it was possible to succeed as a PhD student while working full-time and having a family. You continue to lead the way as a model of the impact one can make on the literacy field while following your passions and having the utmost respect for teachers and students. You’ve always been generous with your time when I need your advice or expertise, even if it’s through Twitter at one in the morning. I aspire to support other scholars the way you have supported me. I would also like to express my gratitude to Elise Wilson, the Doctoral Programs Advisor extrodinaire who helped me the navigate the ins and outs of the PhD process while I completed my dissertation from out of state.

I would also like to thank my fellow current and former doctoral students at the University of Illinois at Chicago for your friendship and empathy, especially Timothy Pappageorge, Megan Hughes, and Tiffany Ko. In particular I would like to thank Tim for his insight and practical advice regarding my dissertation defense. Your commitment and caring for our students when we taught literacy courses together at UIC was inspiring. You have an uncommon balance of wisdom, humility and humor and I look forward to working together again soon.

Thank you to my former teaching and coaching colleagues in and around Chicago, Illinois for your tireless commitment to helping all students succeed. I would also like to express my appreciation to the principal, cooperating teacher, and students at my research site for opening your classrooms to me and for your willingness to support literacy research. Thank you to my colleagues in Western Massachusetts and the schools and leaders we support. It is a blessing to work with others that are equally committed to social justice through literacy.

Finally I would like to give my biggest thanks to my family, especially my wife Goldie and daughter Sophie. Goldie, I recognize all the caring and support you have invested to help us accomplish this goal. There were countless hours we could have spent sharing laughs and connecting with each other that I spent in front of a computer instead. I will show you through my actions that my priorities are in order and give you the love you deserve. Sophie, I read journal articles aloud to you while you were still in your mother’s womb. You have been by my side from the beginning of this process and I am so proud of the girl you have become. The other day you told me I had never taken you camping. It made me wonder how many other things we haven’t done together because I’ve been busy. I can’t wait to make up for lost time as our family begins an exciting new stage of our journey together.

RPM
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>a. Theoretical Framework</td>
<td>4</td>
</tr>
<tr>
<td>b. Statement of the Problem</td>
<td>10</td>
</tr>
<tr>
<td>c. Research Questions</td>
<td>17</td>
</tr>
<tr>
<td>d. Definition of Key Terms</td>
<td>20</td>
</tr>
<tr>
<td>II. REVIEW OF THE LITERATURE</td>
<td>24</td>
</tr>
<tr>
<td>a. Sociocultural Framework</td>
<td>24</td>
</tr>
<tr>
<td>b. A Model of Literacy Progression</td>
<td>27</td>
</tr>
<tr>
<td>i. Stage 1: Basic literacy</td>
<td>29</td>
</tr>
<tr>
<td>ii. Stage 2: Intermediate literacy</td>
<td>29</td>
</tr>
<tr>
<td>iii. Stage 3: Disciplinary literacy</td>
<td>31</td>
</tr>
<tr>
<td>c. Expert/Novice Studies</td>
<td>31</td>
</tr>
<tr>
<td>d. Discourse Perspective</td>
<td>34</td>
</tr>
<tr>
<td>e. Reading and Writing Connection</td>
<td>37</td>
</tr>
<tr>
<td>f. Knowledge Telling and Knowledge Transforming</td>
<td>40</td>
</tr>
<tr>
<td>g. Multiple Text Comprehension</td>
<td>42</td>
</tr>
<tr>
<td>i. Foundational research comprehending single texts- Situation model</td>
<td>42</td>
</tr>
<tr>
<td>ii. From single to multiple texts</td>
<td>44</td>
</tr>
<tr>
<td>iii. The Documents model</td>
<td>46</td>
</tr>
<tr>
<td>iv. Students’ engagement with multiple historical texts</td>
<td>48</td>
</tr>
<tr>
<td>h. Heuristic Use and Students</td>
<td>50</td>
</tr>
<tr>
<td>i. Close Reading</td>
<td>52</td>
</tr>
<tr>
<td>j. Why Study Argumentation?</td>
<td>53</td>
</tr>
<tr>
<td>i. Why apply Toulmin’s Argument Pattern?</td>
<td>56</td>
</tr>
<tr>
<td>1. Why emphasize warrants?</td>
<td>57</td>
</tr>
<tr>
<td>2. Argumentation in the disciplines</td>
<td>58</td>
</tr>
<tr>
<td>3. Students writing warranted historical arguments</td>
<td>61</td>
</tr>
<tr>
<td>4. Warrants as the “home” of disciplinary thinking</td>
<td>62</td>
</tr>
<tr>
<td>5. Examples of warrant generating questions</td>
<td>65</td>
</tr>
<tr>
<td>k. Summary</td>
<td>66</td>
</tr>
<tr>
<td>III. METHODS</td>
<td>68</td>
</tr>
<tr>
<td>a. Participants and Setting</td>
<td>68</td>
</tr>
<tr>
<td>b. Materials</td>
<td>73</td>
</tr>
<tr>
<td>i. Texts for instrumentation</td>
<td>73</td>
</tr>
<tr>
<td>ii. Texts for instruction</td>
<td>76</td>
</tr>
<tr>
<td>iii. Models of writing</td>
<td>76</td>
</tr>
<tr>
<td>iv. Instrumentation for main study: Types of claims</td>
<td>76</td>
</tr>
<tr>
<td>v. Pretest</td>
<td>78</td>
</tr>
<tr>
<td>vi. Warrant selection task</td>
<td>79</td>
</tr>
<tr>
<td>vii. Warrant subtypes</td>
<td>80</td>
</tr>
<tr>
<td>viii. Foils</td>
<td>84</td>
</tr>
<tr>
<td>ix. Warrant ranking task</td>
<td>84</td>
</tr>
<tr>
<td>x. Warrant writing task</td>
<td>85</td>
</tr>
<tr>
<td>c. Overview of Validity and Fidelity</td>
<td>85</td>
</tr>
<tr>
<td>i. Assignment to condition</td>
<td>85</td>
</tr>
<tr>
<td>TABLE OF CONTENTS (continued)</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td></td>
</tr>
<tr>
<td>ii. Pilot testing..........................</td>
<td>86</td>
</tr>
<tr>
<td>iii. Internal validity......................</td>
<td>86</td>
</tr>
<tr>
<td>d. Overview of Procedures..................</td>
<td>87</td>
</tr>
<tr>
<td>i. Overview of comparison treatment........</td>
<td>87</td>
</tr>
<tr>
<td>ii. Overview of treatment condition.........</td>
<td>88</td>
</tr>
<tr>
<td>1. Day 1: Pretest and intro to argument (all students together)</td>
<td>89</td>
</tr>
<tr>
<td>2. Day 2 for comparison treatment............</td>
<td>90</td>
</tr>
<tr>
<td>3. Day 2 for treatment condition.............</td>
<td>91</td>
</tr>
<tr>
<td>4. Day 3 for comparison treatment...............</td>
<td>92</td>
</tr>
<tr>
<td>5. Day 3 for treatment condition.............</td>
<td>93</td>
</tr>
<tr>
<td>6. Day 4 for comparison treatment...............</td>
<td>94</td>
</tr>
<tr>
<td>7. Day 4 for treatment condition.............</td>
<td>95</td>
</tr>
<tr>
<td>8. Day 5 for comparison treatment...............</td>
<td>97</td>
</tr>
<tr>
<td>9. Day 5 for treatment condition.............</td>
<td>98</td>
</tr>
<tr>
<td>10. Day 6: All students together- reading document set.........</td>
<td>99</td>
</tr>
<tr>
<td>11. Day 7: All students together- warrant selection and ranking</td>
<td>100</td>
</tr>
<tr>
<td>12. Day 8: All students together- warrant writing and essay....</td>
<td>100</td>
</tr>
<tr>
<td>e. Data Sources and Analysis..................</td>
<td>100</td>
</tr>
<tr>
<td>i. Research question group 1..................</td>
<td>101</td>
</tr>
<tr>
<td>1. Research question 1a....................</td>
<td>101</td>
</tr>
<tr>
<td>a. Warrant selection task...................</td>
<td>101</td>
</tr>
<tr>
<td>2. Research question 1b....................</td>
<td>103</td>
</tr>
<tr>
<td>a. Warrant ranking task...................</td>
<td>103</td>
</tr>
<tr>
<td>3. Research question 1c....................</td>
<td>104</td>
</tr>
<tr>
<td>a. Warrant writing task...................</td>
<td>104</td>
</tr>
<tr>
<td>ii. Research question group 2..................</td>
<td>106</td>
</tr>
<tr>
<td>1. Research question 2a....................</td>
<td>106</td>
</tr>
<tr>
<td>a. Cut scores for warrant selection and ranking....</td>
<td>106</td>
</tr>
<tr>
<td>b. Correlations..................................</td>
<td>107</td>
</tr>
<tr>
<td>2. Research question 2b....................</td>
<td>107</td>
</tr>
<tr>
<td>a. Item analysis of thinking types...........</td>
<td>107</td>
</tr>
<tr>
<td>3. Research question 2c....................</td>
<td>108</td>
</tr>
<tr>
<td>a. Warrant writing task...................</td>
<td>108</td>
</tr>
<tr>
<td>b. Essay writing task...................</td>
<td>109</td>
</tr>
<tr>
<td>iii. Research question group 3..................</td>
<td>110</td>
</tr>
<tr>
<td>1. Research question 3a....................</td>
<td>110</td>
</tr>
<tr>
<td>a. Sampling and overview think aloud and interview</td>
<td>110</td>
</tr>
<tr>
<td>b. Think aloud..................................</td>
<td>111</td>
</tr>
<tr>
<td>2. Research question 3b....................</td>
<td>112</td>
</tr>
<tr>
<td>a. Semi-structured interview..................</td>
<td>112</td>
</tr>
<tr>
<td>IV. RESULTS..................................</td>
<td>114</td>
</tr>
<tr>
<td>a. Pretest........................................</td>
<td>114</td>
</tr>
<tr>
<td>b. Research Sub Questions........................</td>
<td>117</td>
</tr>
<tr>
<td>i. Intercorrelation of study measures........</td>
<td>118</td>
</tr>
<tr>
<td>ii. MANOVA analysis of most reliable dependent variables......</td>
<td>121</td>
</tr>
<tr>
<td>iii. Analysis for research question 1...............</td>
<td>124</td>
</tr>
<tr>
<td>1. Research question 1a....................</td>
<td>124</td>
</tr>
<tr>
<td>a. Warrant selection task...................</td>
<td>124</td>
</tr>
<tr>
<td>2. Research question 1b....................</td>
<td>127</td>
</tr>
</tbody>
</table>
TABLE OF CONTENTS (continued)

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Selection of “correct” versus “not wrong” warrants</td>
<td>127</td>
</tr>
<tr>
<td>b. Warrant ranking task</td>
<td>128</td>
</tr>
<tr>
<td>3. Research question 1c</td>
<td>130</td>
</tr>
<tr>
<td>a. Warrant writing task</td>
<td>130</td>
</tr>
<tr>
<td>b. Essay writing</td>
<td>134</td>
</tr>
<tr>
<td>iv. Research question group 2</td>
<td>141</td>
</tr>
<tr>
<td>1. Research question 2a</td>
<td>141</td>
</tr>
<tr>
<td>a. Correlation between warrant selection and ranking</td>
<td>141</td>
</tr>
<tr>
<td>b. Correlation between warrant selection and writing</td>
<td>142</td>
</tr>
<tr>
<td>c. Cut scores</td>
<td>142</td>
</tr>
<tr>
<td>d. Cut score analysis for warrant selection task</td>
<td>144</td>
</tr>
<tr>
<td>e. Cut score analysis for warrant ranking task</td>
<td>145</td>
</tr>
<tr>
<td>f. Cut score analysis for multiple warrants subset</td>
<td>147</td>
</tr>
<tr>
<td>2. Research question 2b</td>
<td>150</td>
</tr>
<tr>
<td>a. Item analysis of thinking types</td>
<td>151</td>
</tr>
<tr>
<td>3. Research question 2c</td>
<td>152</td>
</tr>
<tr>
<td>a. Warrant writing task item 1</td>
<td>152</td>
</tr>
<tr>
<td>b. Warrant writing task item 2</td>
<td>155</td>
</tr>
<tr>
<td>c. Warrant writing task items 1 and 2 combined</td>
<td>159</td>
</tr>
<tr>
<td>d. Essay writing task</td>
<td>161</td>
</tr>
<tr>
<td>e. Predicate and side analysis for warrant selection</td>
<td>163</td>
</tr>
<tr>
<td>v. Research question group 3</td>
<td>165</td>
</tr>
<tr>
<td>1. Research question 3a</td>
<td>165</td>
</tr>
<tr>
<td>a. Think aloud</td>
<td>165</td>
</tr>
<tr>
<td>b. Think aloud for warrant selection task</td>
<td>169</td>
</tr>
<tr>
<td>c. Think aloud for warrant writing task</td>
<td>171</td>
</tr>
<tr>
<td>2. Research question 3b</td>
<td>174</td>
</tr>
<tr>
<td>a. Semi-structured interview</td>
<td>174</td>
</tr>
<tr>
<td>b. Historical thinking in semi-structured interview</td>
<td>174</td>
</tr>
<tr>
<td>c. What would make this work easier</td>
<td>177</td>
</tr>
<tr>
<td>d. What students learned versus typical instruction</td>
<td>178</td>
</tr>
<tr>
<td>e. Ivan’s response</td>
<td>178</td>
</tr>
<tr>
<td>f. Jose’s response</td>
<td>179</td>
</tr>
<tr>
<td>g. Ethan’s response</td>
<td>180</td>
</tr>
<tr>
<td>h. Sonia’s response</td>
<td>181</td>
</tr>
<tr>
<td>i. Follow up question</td>
<td>181</td>
</tr>
<tr>
<td>V. DISCUSSION AND IMPLICATIONS</td>
<td>184</td>
</tr>
<tr>
<td>a. Overview</td>
<td>184</td>
</tr>
<tr>
<td>i. MANOVA of most reliable study measures</td>
<td>185</td>
</tr>
<tr>
<td>ii. Warrant selection</td>
<td>185</td>
</tr>
<tr>
<td>iii. Warrant ranking</td>
<td>188</td>
</tr>
<tr>
<td>iv. Cut scores</td>
<td>189</td>
</tr>
<tr>
<td>v. Warrant writing</td>
<td>189</td>
</tr>
<tr>
<td>vi. Additional example of transfer</td>
<td>196</td>
</tr>
<tr>
<td>b. Limitations</td>
<td>197</td>
</tr>
<tr>
<td>c. Implications for Research and Instruction</td>
<td>202</td>
</tr>
<tr>
<td>i. Implications for research: Assessment design and analysis</td>
<td>202</td>
</tr>
<tr>
<td>ii. Implications for instruction: Characteristics intervention and theory</td>
<td>205</td>
</tr>
<tr>
<td>1. Backwards design</td>
<td>206</td>
</tr>
</tbody>
</table>
TABLE OF CONTENTS (continued)

2. Careful selection of texts ....................................................... 207
3. Close reading and heuristic use ............................................. 208
4. Scaffolding student writing .................................................. 211
5. Differentiating between more and less effective warrants ...... 211
6. Opportunities to discuss warrants and receive feedback ......... 212
7. Analyzing student writing to influence instruction ............... 213
d. Conclusion ............................................................................. 214

VI. REFERENCES ........................................................................... 220
VII. APPENDICES .......................................................................... 233
a. Appendix A ............................................................................ 233
b. Appendix B ............................................................................. 242
c. Appendix C ............................................................................. 245
d. Appendix D ............................................................................. 251
e. Appendix E ............................................................................. 253
f. Appendix F ............................................................................. 254
g. Appendix G ............................................................................. 255
h. Appendix H ............................................................................. 257
i. Appendix I ............................................................................. 258
j. Appendix J ............................................................................. 261
k. Appendix K ............................................................................. 265
l. Appendix L ............................................................................. 266

VIII. VITA .................................................................................. 267
IX. IRB APPROVAL ....................................................................... 274
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. TABLE 4.1: INTERCORRELATIONS FOR KEY STUDY MEASURES</td>
<td>120</td>
</tr>
<tr>
<td>II. TABLE 4.2: WARRANT WRITING TASK- COMBINED</td>
<td>122</td>
</tr>
<tr>
<td>III. TABLE 4.3: WARRANT SELECTION TASK- CORRECT</td>
<td>123</td>
</tr>
<tr>
<td>IV. TABLE 4.4: WARRANT SELECTION TASK</td>
<td>126</td>
</tr>
<tr>
<td>V. TABLE 4.5: SUM OF WARRANT SELECTION- NOT WRONG</td>
<td>128</td>
</tr>
<tr>
<td>VI. TABLE 4.6: SUM OF WARRANT RANKING SCORES</td>
<td>129</td>
</tr>
<tr>
<td>VII. TABLE 4.7: COMBINED PERFORMANCE ON WARRANT WRITING, 1 &amp; 2- QUALITY</td>
<td>132</td>
</tr>
<tr>
<td>VIII. TABLE 4.8: WARRANT WRITING TASK- SUM OF HISTORICAL THINKING</td>
<td>134</td>
</tr>
<tr>
<td>IX. TABLE 4.9: PERFORMANCE BY CUT SCORES ON WARRANT SELECTION TASK</td>
<td>145</td>
</tr>
<tr>
<td>X. TABLE 4.10: PERFORMANCE BY CUT SCORES ON WARRANT RANKING TASK</td>
<td>146</td>
</tr>
<tr>
<td>XI. TABLE 4.11: PERFORMANCE BY CUT SCORES ON WARRANT SELECTION- SUBSET OF STUDENTS</td>
<td>148</td>
</tr>
<tr>
<td>XII. TABLE 4.12: PERFORMANCE BY CUT SCORES ON WARRANT RANKING- SUBSET OF STUDENTS</td>
<td>149</td>
</tr>
<tr>
<td>XIII. TABLE 4.13: PERFORMANCE ON WARRANT SELECTION TASK- BY TYPE OF THINKING</td>
<td>151</td>
</tr>
<tr>
<td>XIV. TABLE 4.14: PERFORMANCE ON WARRANT WRITING TASK BY TYPE OF THINKING- ITEM 1</td>
<td>154</td>
</tr>
<tr>
<td>XV. TABLE 4.15: PERFORMANCE ON WARRANT WRITING TASK BY TYPE OF THINKING- ITEM 2</td>
<td>157</td>
</tr>
<tr>
<td>XVI. TABLE 4.16: PERFORMANCE ON WARRANT WRITING TASK BY TYPE OF THINKING- ITEMS 1 &amp; 2</td>
<td>160</td>
</tr>
<tr>
<td>XVII. TABLE 4.17: WARRANT SELECTION TASK PERFORMANCE BY SIDE AND PREDICATE</td>
<td>164</td>
</tr>
</tbody>
</table>
LIST OF TABLES (continued)

XVIII. TABLE 4.18: PERFORMANCE ON READING AND ENGLISH……….. 166
       ACT-PLAN VERSUS BENCHMARK

XIX.  TABLE 4.19: HISTORICAL THINKING IN WARRANT SELECTION.. 170
       TASK THINK ALOUD

XX.   TABLE 4.20: HISTORICAL THINKING IN WARRANT WRITING….. 173
       TASK THINK ALOUD

XXI.  TABLE 4.21: HISTORICAL THINKING IN SEMI-STRUCTURED……. 176
       INTERVIEW
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Figure 3.1: Example of a warrant selection task</td>
<td>103</td>
</tr>
<tr>
<td>2. Figure 3.2: Example of a warrant writing task</td>
<td>106</td>
</tr>
<tr>
<td>3. Figure 4.1: Warrant selection and ranking by category</td>
<td>147</td>
</tr>
<tr>
<td>4. Figure 4.2: Subset of warrant selection and ranking by category</td>
<td>150</td>
</tr>
<tr>
<td>5. Figure 4.3: Warrant writing task item 1</td>
<td>153</td>
</tr>
<tr>
<td>6. Figure 4.4: Warrant selection task for think aloud</td>
<td>168</td>
</tr>
</tbody>
</table>
### LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>American College Test</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>AP</td>
<td>Advanced Placement</td>
</tr>
<tr>
<td>CCSS</td>
<td>Common Core State Standards</td>
</tr>
<tr>
<td>CCSSO</td>
<td>Council of Chief State School Officers</td>
</tr>
<tr>
<td>DBQ</td>
<td>Document-Based Question</td>
</tr>
<tr>
<td>MANOVA</td>
<td>Multivariate Analysis of Variance</td>
</tr>
<tr>
<td>NAEP</td>
<td>National Assessment of Educational Progress</td>
</tr>
<tr>
<td>NCLB</td>
<td>No Child Left Behind</td>
</tr>
<tr>
<td>NGACBP</td>
<td>National Governors Association Center for Best Practices</td>
</tr>
<tr>
<td>NICHD</td>
<td>National Institute of Child Health and Human Development</td>
</tr>
<tr>
<td>PARCC</td>
<td>Partnership for Assessment of Readiness for College and Careers</td>
</tr>
<tr>
<td>PASS</td>
<td>Power Analysis and Sample Size</td>
</tr>
<tr>
<td>PISA</td>
<td>Program for International Student Assessment</td>
</tr>
<tr>
<td>SES</td>
<td>Socioeconomic Status</td>
</tr>
<tr>
<td>SPSS</td>
<td>Statistical Package for the Social Science</td>
</tr>
</tbody>
</table>
SUMMARY

This dissertation reports design-based research that determined the characteristics of an effective intervention to improve the historical argumentative writing of high school juniors. It compared a treatment that taught students to write warrants that used historical thinking to explain how evidence supports a claim, and a comparison treatment that taught students to find and evaluate evidence to support particular claims and sides. This research is important because historical argumentative writing is a core literacy practice of the discipline of history, and there is a gap in the research when it comes to explicitly teaching warrants. Historical argumentative writing reflects literacy challenges similar to those that students encounter when reading and evaluating documents online, and is essential for engaged citizenship in the 21st Century. When the most reliable study measures were combined and analyzed using MANOVA, there was a significant overall effect for the warrant treatment condition. I used the results of this study to establish the characteristics of an effective intervention that will be refined through further research. I also developed theory about how best to teach historical argumentative writing to adolescents that can inform work in similar contexts.

This study took place in the 11th grade history classes of one teacher in a diverse high school with approximately 50% Hispanic students and 12% students with disabilities. Building on previous studies of multiple text reading and argumentative writing, I first developed instructional prototypes with feedback from experts to ensure construct validity and one-on-one trials with students to judge practicality. The main study was eight days in length. Both groups used the same documents about the controversy surrounding the explosion of the battleship U.S.S. Maine in Havana Harbor at the start of the Spanish-American War. Students (N = 89) were randomly assigned to condition stratified by classroom with instructors (myself and the classroom teacher) counterbalanced across conditions. Pretesting of students’ experience of and
SUMMARY (continued)

feelings about argumentative writing and prior topic knowledge showed no significant differences between the two conditions.

The treatment condition taught students to use historical thinking tools such as sourcing and corroboration (reading across multiple historical accounts to compare content and identify differences, giving more weight to that information which is common across accounts), along with close reading as explicitly stated warrants through the use of warrant generating questions and templates. The comparison treatment focused on finding and evaluating evidence to support particular claims and sides. Both interventions included reading the same historical text sets, gradual release of responsibility to students, and opportunities for peer discussion and feedback. The performance of these two groups was compared in order to establish whether the instruction improved student ability to 1) select effective warrants reflecting different types of historical thinking, 2) generate their own effective warrants when a given claim and evidence, and 3) write more effective warrants in their own argumentative essay.

At the culmination of the intervention, students completed a multiple-select task where they chose effective warrants for given claims and evidence, with foils reflecting common errors in reasoning. Students then ranked the warrants they selected in terms of effectiveness. Next they were given claims and evidence and had to write a warrant. They also wrote an essay about whether or not the explosion of the U.S.S. Maine caused the U.S. to invade Cuba. A sample of four students from the treatment condition completed a think aloud protocol analysis and a brief semi-structured interview to understand student thinking during the task and their experience of the intervention.

Due to the relatively small sample size and limited number of assessment items, related measures were combined to increase reliability. Scores tied to selecting correct warrants were combined to create one overall warrant selection score, as were scores related to warrant
writing. The overall effect of these two combined scores was analyzed using MANOVA. This analysis revealed a significant overall effect for the treatment (heuristics as warrants) condition. Follow up ANOVAs indicated that the warrant selection score was statistically significant, though the combined warrant writing score did not reach the level of significance.

Additional analyses of other study measures were undertaken to determine what may have contributed to this overall effect. Mean scores were consistently higher for the treatment condition, though they did not reach the level of significance due to small sample sizes and corrections for multiple analyses from the same data set. Assessment items linked to corroboration often had the highest mean performance, followed by items linked to close reading and sourcing. Corroboration is important because it helps students realize the interpretive nature of history and make sense of conflicting accounts. Sourcing scores may be relatively lower because students had prior exposure to a routine of identifying sources. Post-hoc power analyses indicated that a moderately larger sample size may yield a significant treatment effect on several additional measures.

To aid in the finding of patterns, students were divided into quartiles based on performance on the warrant selection task, and these quartiles were used to analyze essay writing. There was qualitative evidence of a shift to more sophisticated types of historical thinking in the treatment condition in the third and fourth quartiles, with students showing greater levels of close reading and corroboration. Students in the lower quartiles relied almost entirely on sourcing when heuristic use was present. Qualitative analysis of the think aloud and semi-structured interview indicated that most students found the instruction valuable, and felt they would benefit from additional practice and feedback.

Based on these findings, I refined the characteristics of an effective intervention and began developing a theory of historical argumentative writing instruction. Refinements to the treatment intervention include targeted instruction in using heuristics in combination to
SUMMARY (continued)

strengthen an argument, and employing evidence organizers for reading across texts to support corroboration. They also include additional instruction in differentiating between more and less effective warrants for a particular claim. Prior to the next cycle of research, teachers will look at student writing to determine strengths and areas for growth in the use of warrants, targeting these areas for additional practice and feedback.

The emerging theory of teaching argumentative writing based on these findings includes the importance of explicit instruction in writing historical arguments warranted by historical thinking heuristics, including modeling, questioning strategies and writing scaffolds. The theory supports instruction organized in units centered on historical controversies that are backwards planned from the demands of the argumentative writing students will be expected to produce at unit’s end. It underscores the importance of teachers reading and writing historical texts themselves, identifying places in the text that call for heuristic use to target for close reading, and developing questions and scaffolds to address these areas. Finally, it requires teachers to design ample opportunities for students to discuss their warrants and receive feedback modeled after disciplinary norms, to apprentice students into the literacy practices of the Discourse of history.

In conclusion, this instructional intervention was shown to have a statistically significant overall effect on the most reliable study measures when examined through MANOVA. These findings were used to develop a theory of teaching argumentative writing that can inform work in similar contexts, and will be refined through further research. This approach to instruction may be appealing to history teachers because it provides a model of explicit instruction to improve historical argumentative writing that also keeps learning historical content as a central focus, unlike traditional content area literacy approaches that de-emphasize content. It may be more engaging for students than traditional textbook-centered instruction because students actively read, reason and write arguments about engaging historical controversies.
I. INTRODUCTION

Disciplinary literacy, a type of advanced literacy instruction that apprentices students into the literacy practices of disciplines such as history and science, has been sorely lacking for adolescents (Moje, 2008; Shanahan & Shanahan, 2008). This study focuses on a particular area of disciplinary literacy, namely historical argumentation. It aims to develop students’ warrants in their argumentative writing from multiple texts. In order to establish the goals and importance of this work, I will frame this study within the contemporary context of adolescent literacy education and provide an overview of its theoretical framework.

Adolescent literacy has long been neglected in both educational research and public policy, but that trend is finally reversing itself. The increased attention to adolescent literacy is due to two central reasons: attempts to prepare adolescents for the 21st century workplace (tied closely to concerns about U.S. global competitiveness), and failed efforts to “cure” literacy problems by only targeting early reading needs. Overall literacy levels for American youth have remained largely unchanged since the 1970s, when the U.S. educational system was at or near the top in performance on international literacy assessments. The influence of globalization, the impact of technology, and the reality of a 21st century workplace require a higher level of literacy skill than ever before (Carnevale & Desrochers, 2004), which means maintaining literacy levels is not sufficient. Six in 10 jobs are held by workers with some degree of postsecondary education or training; in 1959, the rate was 2 in 10. As many as two-thirds of new jobs will require a college education and high levels of literacy (Carnevale & Desrochers, 2004). Failing to adjust to the new literacy demands, the U.S. is now being outstripped by other countries that built on what the U.S. was doing right and refined it. A recent Program for International Student Assessment (PISA) report indicates that nine other countries scored higher in the combined reading scale than
American 15 year olds, with the U.S. mired in a group of 16 countries with scores not measurably different (Fleischman, 2010). Domestically, only 51% of total students tested on the ACT in 2005 met the benchmark for reading, a measure of college and career readiness (American College Testing, 2006).

Though the emphasis on adolescent literacy is a relatively new phenomenon, anxiety about the literacy performance of young people is not. The discourse about literacy in the latter half of the 20th century was dominated by concern about reading performance in general and the literacy levels of young children in particular. The modern standards-based reform era in education was initiated with the touchstone report, “A Nation at Risk” (Gardner, 1983), which sounded an alarmist note about America’s lack of commitment to educational excellence. This report described shortcomings in our education system as “an act of unilateral educational disarmament” (Gardner, 1983, p. 1). It stimulated efforts toward more rigorous standards, begun by the National Governors Association in 1989. It precipitated efforts to improve literacy achievement including the establishment of a National Education Goals Panel, the National Academy of Sciences report “Preventing Reading Difficulties in Young Children” (Snow, Burns, & Griffin, 1998), the National Reading Panel Report (NICHD, 2000), and the No Child Left Behind Act (No Child Left Behind [NCLB], 2003).

Many of these reform efforts employed a theory of action that literacy difficulties could be mitigated by a focus on early literacy achievement by catching and correcting reading problems early, thus “vaccinating” students from later difficulty (Shanahan & Barr, 1995, as cited in Shanahan & Shanahan, 2008). In fact, many state standards stopped addressing reading at the secondary level altogether. Early literacy reform efforts in the 1990s targeting basic literacy (including No Child Left Behind) (NCLB, 2003) fell into this category. The significant investment of money and time showed modest success in the form of National Assessment of Educational Progress (NAEP) score gains in early literacy, but did not lead to subsequent growth
for adolescent readers, whose scores have remained flat, or even declined (Grigg, Donahue, & Dion, 2007).

Thus, in recent years educational policymakers and literacy organizations have broadened their focus to include adolescent reading achievement, typically including grades 4-12 (Alvermann, 2001; Moore, Bean, Birdyshaw, & Rycik, 1999). The Common Core State Standards (hereafter CCSS), more rigorous than traditional standards and back-mapped from literacy levels needed for success in college and career, have also helped add to a sense of urgency around literacy achievement (National Governors Association Center for Best Practices [NGA Center] & Council of Chief State School Officers [CCSSO], 2010). The leveling off or decline in achievement in the middle grades, known as the “fourth grade slump”, has befuddled educational researchers and practitioners. It may be associated with students making the switch from learning to read, with primarily familiar, engaging, narrative text, to reading to learn, with unfamiliar informational texts in the disciplines, though the exact cause of the slump is up for debate. As teaching becomes departmentalized and disciplinary texts and language become more specialized in the secondary grades, students may associate the differences in ways of reading or writing in their content area classes more with those teachers’ expectations than with differences in the literacy practices of the disciplines (O’Brien, Stewart, & Moje, 1995).

In response to the well-documented difficulties faced by adolescent readers, researchers and schools responded by attempting to make every teacher a reading teacher through efforts at reading and writing across the curriculum (Moje, 2008). The vast majority of traditional “content area reading” reform efforts and research treat reading and writing as universal processes that operate largely the same regardless of subject area (Shanahan & Shanahan, 2008). Content area teachers have been resistant to these efforts (Alvermann & Moore, 1991) and see strategies and instruction for reading texts as cutting into their time and as out of touch with their understanding of what it means to learn in their discipline (Moje, 2008). The pressure many teachers feel to
“cover” content leads to the “pedagogy of telling” (O’Brien, Stewart, & Moje, 1995), which limits student opportunities to construct their own meaning from disciplinary text.

In order to meet the demands of 21st century careers and citizenship, students need instructional support in the highest and most specialized stage of literacy development, disciplinary literacy. The disciplines are the fields in which knowledge is actually produced; for instance, science as practiced by scientists, history as practiced by historians, and mathematics as practiced by mathematicians. Each discipline has its own cultural and social practices that have evolved over time with unique forms and uses of language. To not only comprehend the message of disciplinary text, but also to be able analyze, critique, and challenge it, and to create new knowledge, students need specific kinds of background knowledge, specific understandings about the specialized vocabulary and texts, and access to the particular ways of speaking, writing and listening in a discipline (Fang, 2012; Yore, Pimm, & Tuan, 2007). However, these elements are scarcely ever taught (Lee & Spratley, 2010; Shanahan & Shanahan, 2008).

**Theoretical Framework**

The theory that learning is inherently social (Vygotsky, 1978) has made a significant impact on literacy research beginning in the mid to late 20th century. Even literacy practices such as reading a book alone can be considered a social activity, because the reader is engaged with a text written by another person, in a language developed over long periods of social interaction. Each reader’s schema consists in part of other texts they have read, their interactions with others, and others ideas (Au, 1998). As Lemke (1990) argues, “We do not communicate by the transmission of signs or signals, but by creating and manipulating social situations” (p. x). In literacy research, the focus has shifted away somewhat from the study of individuals and what is going on in their minds as the unit of study and toward situating reading and writing within larger social practices (Gee, 1989; Lankshear, Gee, Knobel, & Searle, 1997). There has been a greater focus on discourse, or language as used in social interaction (Luke, 1995). Studies grounded in a
sociocultural framework often examine teaching and learning in actual classrooms as opposed to labs, focusing on teachers and children interacting and sharing ideas, and the language of these interactions. Interventions have dealt directly with designing new patterns of interaction or ways to influence the discourse that is happening in these classrooms.

This emphasis on the importance of language and its use has led to an examination of the challenges of literacy in the disciplines (Fang, 2012; Lee & Spratley, 2010), the literacy practices of experts (Shanahan, Shanahan, & Misischia, 2011; Wineburg, 1991), and ways to make these accessible for students. Shanahan and Shanahan (2008) developed a model of literacy progression that captures the increasing specificity of literacy practices, from basic, universal processes involved in any reading of text, to intermediate literacy that includes generic comprehension skills and reading strategies relevant to many tasks, to disciplinary literacy, the specialized literacy practices of the disciplines.

Disciplinary literacy is “based on the premise that students can develop conceptual knowledge in a discipline only by using the habits of reading, writing, talking and thinking which that discipline values and uses” (McConachie et al., 2006, p.8). These practices were identified in part by examining the thinking and strategy use of expert readers and writers in their respective disciplines (Shanahan et al., 2011) and examining the texts these experts read and produce (Fang, 2012; Schleppegrell, 2007). Unlike intermediate literacy, disciplinary literacy includes methods of generating new knowledge through research or synthesis and evaluating the knowledge and reasoning of others. It more closely approximates the kinds of complex reading and thinking that students need for learning in the 21st century than traditional views of literacy because experts in disciplines such as history frequently read multiple, often multimodal, texts of unclear origin and make sense of them; skills needed every day for reading online (Manderino, 2011; Wineburg & Reisman, 2015).
Through the lens of discourse, the disciplines are social spaces where researchers can examine established patterns of language use associated with them. ‘Capital D’ Discourses are “ways of behaving, interacting, valuing, thinking, believing and often reading and writing” (Gee, 1989a), differentiated from (lower case) discourse, or language-in-use. Each Discourse serves as an “identity kit” or a way of being in the world that others recognize (Gee, 1989a). People use and acquire these Discourses to show membership in social groups and gain social or economic capital. As spaces where new knowledge is constructed, the disciplines represent dominant Discourses. Expertise in a disciplinary Discourse brings with it potential access to status, employment, money, and power. Disciplinary literacy shifts the emphasis from a focus on training students in universal processes to apprenticing students into the ways experts use reading, writing, thinking and speaking to produce and critique new knowledge in these dominant Discourses (Schoenbach, Greenleaf, Cziko, & Hurwitz, 1999); see Moje, 2007 for a review.

One of the essential practices of disciplinary experts is reading across multiple texts. Research has established a theory of what happens cognitively when a child reads a single text (e.g. Kintsch, 1988; Kintsch & van Dijk, 1978). When expanding comprehension theory to multiple texts, researchers turned to expert readers in the disciplines (i.e. Wineburg, 1991), as comprehending multiple texts is foundational to disciplinary expertise. For instance, historical text cannot be understood except in relation to other texts, nor can a valid historical argument be made from reading a single document. Experts use heuristics, or specialized, short patterns of thinking, to analyze multiple texts (Rouet, Favart, Britt, & Perfetti, 1997; Shanahan et al., 2011; Wineburg, 1991).

Disciplinary experts spend a great deal of time reading and writing (Shanahan et al., 2011). There is research to indicate that reading and writing are reciprocal processes and reinforce each other (De La Paz & Felton, 2010; Fitzgerald & Shanahan, 2000; Shanahan, 1984). However, reading and writing remain largely and artificially separated in American education.
One of the challenges of writing historical arguments is the high cognitive demand of reading across texts. Therefore, the reciprocal nature of the cognitive and social processes of reading and writing is important to leverage when attempting to apprentice students into the complex literacies of the disciplines.

Another core process in the disciplines is argumentation. Argumentation, a process of systematic reasoning in support of an idea or theory, is an essential element of learning and making new meaning (Erduran, 2007; "Argumentation", n.d.). Though a single person can engage in argumentation, the process of making ones thinking and reasoning public can be considered dialogic by its very nature (Kuhn, 2002). The process of argumentation involves entering an ongoing conversation, an existing network of claims and counterclaims about a particular topic and acknowledging, building upon, or refuting the ideas of others. The argument is what is produced as a result of this process. There has been a move in academia from personal narrative essays and toward argumentative writing as the dominant writing mode for the past 30 years (Lunsford, 1999). It is now the core form of academic writing (Graff & Birkenstein, 2007). While research has historically focused on the product of argumentation, the argument, researchers have begun to focus on the process of argumentation as well (Kuhn, 2002).

However, this emphasis in academia has not trickled down to improved performance in our elementary and secondary schools. On a 2011 NAEP persuasive writing task requiring students to read texts and write a letter persuading a council to build or not build a discount store, only 19% of 12th grade students were judged to be at the “competent” level, while only 4% were at the “effective” level (National Assessment of Educational Progress, [NAEP], 2011). With the advent of the Common Core standards, which place a priority on argument writing, recent efforts are beginning to increase the amount of argument writing in schools, but the learning curve is steep (Newell, Beach, Smith, Vanderheide, & Andriessen, 2013).
Among the many approaches and frameworks for studying argumentation [(e.g. Walton’s argument schemes (2008); the pragma-dialectical theory (van Eemeren & Grootendorst, 1992)), no approach has been more influential than the work of Stephen Toulmin. Toulmin (1958) deviates from formal rules of logic for producing the “correct” conclusions based on premises, saying these are of little use in the real world. Instead he provides a “working logic”, a pattern that is grounded in how people actually argue or “reason from premises to conclusions” in particular social contexts. Toulmin determined all arguments include three central elements, a claim (the assertion one is making), based on data (the evidence that supports the claim), and the warrant, (the general, rule-like statement that explains how the evidence supports the claim). Since warrants may be challenged given the controversy or complexity surrounding a particular issue, Toulmin identified additional elements including the backing, which provides factual statements to support the warrant, the rebuttal, which gives the circumstances in which the argument may not apply, and the qualifier, which modifies the claim’s scope (Toulmin, 1958).

For the purposes of illustration, an example of the parts of an argument from the workplace follows: “I most likely (qualifier) deserve a raise (claim). Since I started working here, revenue has raised steadily every quarter (evidence). When revenues go up, the workers who contributed to the growth in revenue should benefit (warrant). I was recognized last staff meeting as the number one earner in the department (backing). Unless my numbers drop drastically (rebuttal), you should boost my salary.” While this simple example captures the basic elements, researchers often find they must articulate the process they used for determining the different argument elements as identifying discrete elements can be challenging and real-world arguments often leave elements unwritten for the reader to infer or place them in an unpredictable order.

Argumentation plays a key role in the creation of new knowledge. For example, scientists and historians use arguments to make knowledge claims, and use warrants and backings
to support them (Erduran, Simon, & Osborne, 2004; Toulmin, 1958). Warrants form the link between the data or evidence that a researcher may collect and the theories they generate (Osborne, Erduran, & Simon, 2004). While the structure of Toulmin’s argument pattern is field-independent, the types of warrants and backings experts use are field-dependent (Toulmin, 1958). From a disciplinary literacy perspective, each discipline has particular standards for what counts as a valid claim or powerful evidence and its own demands when it comes to warranting claims within argumentative writing (De La Paz, 2005; De La Paz & Felton, 2010).

The discipline of history has been examined in many foundational studies of argumentation because of its heavy literacy demands. Historical reasoning, or “working with and using evidence in the interpretive work of historians” (Monte-Sano, 2010), is challenging, as disciplinary arguments require the examination of several texts to gather information and to develop warrants, claims and backings (Larson et al., 2004; Le Bigot & Rouet, 2007; Perfetti, Rouet, & Britt, 1999). This work leads to deeper understanding of content (Brown, Bransford, Ferrera & Campione, as cited in Wiley & Voss, 1999). The critical, questioning stance and attention to subtext (VanSledright, 2004; Wineburg, 1991) has value far beyond the history classroom; it is critical for life in an information society (Luke, 1995; VanSledright, 2004; Wineburg & Reisman, 2015).

Having established the structure of an argument and the importance of studying argumentation, I will now focus on one often overlooked but nonetheless essential element of an argument, the warrant. Within the larger category of argumentation, warrants are worthy of study for several reasons. Arguments tend to hold or fall on the weight of their warrants. Warrants serve as inference licenses that explain the general rule one is using to arrive at their claim from their evidence. The general rule does not need to apply to all arguments, but it does need to be reliable for arguments of whatever type one is writing.
Warrants that align with the reasoning of the field or discourse in question are often more powerful as they call upon the authority of an established body of knowledge. While warrants may be implied but not explicitly stated in an argument, experts are more likely to include explicitly stated warrants (Crammond, 1998). In history, warrants are often specific to the discipline (Bruner, 1960). Since its very nature is interpretive, historians often include the reasoning they undertake to arrive at their claims (Mink, 1987, as cited in Monte-Sano, 2010). Rather than warrant claims through general rules, historians instead attend to who wrote documents, consider confirming and contradictory accounts, and analyze the context of that event to reason about the significance of available evidence (Monte-Sano, 2010).

**Statement of the Problem**

If our goal is disciplinary literacy, we must help students read and write warranted historical arguments. Even our most capable secondary students lack these skills. This is largely because they have not been taught or assessed. The ACT assesses only persuasive writing, and the affiliated College Readiness Standards do not include argument. Fortunately the Common Core Standards created a special place for argument in the standards (Graff & Birkenstein, 2007; Graff, 2004), in conjunction with specialized standards for reading in history (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). The Grades 6-12 Reading in History and Social Studies Standards include essential skills such as attending to source and reading across multiple texts.

Unfortunately, the writing standards for history are combined with standards for science and technical subjects. The CCSS writing standards speak to clarifying the relationship between claims and reasons, and reasons and evidence, but inconsistently refer to argument elements as claims, reasoning and evidence, or premises, claims and evidence without explanation for why terms are shifting (NGACBP-CCSSO, 2010). The focus in the CCSS on claim and evidence may send an unintended message to students and teachers that these are the only elements of argument
that really matter. The standards do not give enough guidance about what to teach to improve historical argumentative writing, and none of the standards give guidance on how to teach.

Even with the advent of the standards, students have relatively little exposure to reading historical arguments, because of the emphasis in instruction on primary sources and document analysis. They lack models of historical arguments as mentor texts. Though the expanded use of document-based questions outside of Advanced Placement (AP) classes seems promising for argumentation, the quality of writing students generate from such work is uneven, and the extent to which this approach has influenced instruction is debatable (Rothschild, 2000).

The lack of skill with warrants severely limits students’ ability to participate in high-level historical discourse. Though research has shown that with good instruction, students can identify and generate their own claims and find evidence to support them, they still struggle with warrants. They produce hardly any warrants at the beginning of middle school, produce only slightly more as the grades increase (Connor, 1990; Crammond, 1998; Knudson, 1992; McCann, 1989), and produce well fewer than experts (Crammond, 1998). Warrants are essential because the same data can be used to argue opposite claims if the warrant changes. Without an understanding of warrant students lack the knowledge to evaluate arguments or write their own. While disciplinary literacy in history generally aligns with the skills they need to be successful in the 21st century (Reisman, 2012), facility with warrants in particular is essential in the information age, because students are bombarded with evidence to support all sorts of dubious claims. They must constantly determine how the evidence does or does not prove the claim.

Existing research in the field of history also shows a gap when it comes to warrants. Research of students’ historical writing shows it often contains claims and evidence, but rarely includes warrants or evidence of heuristic use (Monte-Sano, 2008; Rothschild, 2000). Sam Wineburg and the group at Stanford have developed successful interventions for reading primary source documents (Reisman, 2012), and there are a handful of studies (e.g. De La Paz, 2005; De
La Paz & Felton, 2010; Monte-Sano, 2008; Monte-Sano, 2010b; Monte-Sano & De La Paz, 2012) that target historical argumentative writing and the use of heuristics as strategies to support reasoning. However, these do not emphasize warrants.

This current study answers the call of Fulkerson (1996) for research that adapts the Toulmin model “accurately yet creatively”, targeting warrants in historical argumentation. As Toulmin explains, warrants are not sweeping statements of absolute truth. They are simply inference rules, often made provisional by adding qualifications and rebuttals, but they do need to be reliable. Reliability is interesting because history itself is an interpretation based on incomplete evidence and always subject to revision. Historians read from a questioning, critical stance, approaching every claim with a healthy dose of skepticism. Unlike other disciplines such as mathematics that contain formal inference rules, or science that contain more linear methods for arriving at conclusions, warranting historical claims involves the reading, evaluation, and synthesis of multiple documents that never provide the complete picture.

While there are instances in which expert writers do not include warrants, they do this with an awareness of their audience based on an understanding of commonly accepted premises in the discipline. Students lack this expert knowledge. When they leave out warrants, it is typically for three reasons: Their argument is so simplistic that the warrant is unnecessary; they do not realize it is needed, or they are confused on what the link between their claim and evidence actually is (Crammond, 1998). In addition, the texts students traditionally read in history class do not support an understanding of warrants. History textbooks often only highlight claims, with little evidence and no warrants. They are presented as an unproblematic narrative of what happened. Since students lack common experiences or shared cultural backgrounds with the discourse community and exposure to models of effective warrants, they lack the sophistication to know when to leave out a warrant and should leverage disciplinary patterns of reasoning to establish common ground with their audience if they want their ideas to be heard. Therefore it
would be beneficial for them to get experience warranting their arguments with explicitly stated historical reasoning.

Expert studies (e.g. Wineburg, 1991) have identified heuristics or shorthand mental tools that support historians in reasoning and argument generation. Studies have begun to establish ways to help students use these tools in reading (e.g. Britt & Aglinskas, 2002; Reisman, 2012; Shanahan & Shanahan, 2008). Interventions that rely on social processes, procedural scaffolds, and a written final product show the most promise (Hynd et al., 2004).

How can we help students write warranted historical arguments? Evidence from research has shown that just telling students to use historians’ strategies is generally not enough to improve their use (Britt & Aglinskas; Reisman, 2012; Shanahan & Shanahan; 2008). Social processes are an important but overlooked element in argument instruction. Toulmin (1958) explained that warrants in disciplinary arguments often originate in a dialogic process where the author provides a warrant in response to a “warrant generating question”, e.g. “How did you get there?” (Freeman, 2011; Toulmin, 1958). Warrant generating questions can be modified to support thinking grounded in the discipline, so that the question prompts the use of heuristics. For example, “How does knowing who wrote this document and their perspective on this issue help support your claim?”. Warrant generating questions can be used to scaffold the dialogic nature of argumentation and promote audience-centered activity. This can ensure students ground their arguments in established patterns of disciplinary reasoning, or acknowledge these patterns in presenting alternative patterns of reasoning.

This study is a type of design-based research. Design-based research is an appropriate methodology to study the use of heuristics as warrants for a number of reasons. First of all, traditional intervention research in education has a mixed record of effectiveness, due in part to the complexities of doing studies in schools, along with a legacy of “ill-conceived and poorly constructed research that results in no significant differences, or, at best, in modest effect sizes”
According to the Design-Based Research Collective (2003), even when it does show a statistically significant effect, “educational research is often divorced from the problems and issues of everyday practice - a split that resulted in a credibility gap and creates a need for new research approaches that speak directly to problems of practice and that lead to the development of ‘usable knowledge’” (p.5). This is because experimental or quasi-experimental education research often treats aspects that are part of the larger system—such as teacher knowledge, professional development, curriculum, and assessment—individually, missing how a new innovation in one area will impact (or fail to impact) the others (Brown, 1992). According to Cobb (2006), “the first and most compelling argument for initiating design research stems from the desire to increase the relevance of research for educational policy and practice”. Indeed, Ann Brown, considered one of the originators of design-based research, began using this approach after finding the positive effects of her controlled experiments with students receiving expert instruction, failed to last under real-world classroom conditions (Brown, 1992).

Design-based research is more a collection of approaches than it is any one approach. It intends to “design and develop an intervention as a solution to a complex educational problem” while “advanc(ing) our knowledge about the characteristics of these interventions and the processes to design and develop them” (Barab & Squire, 2004). By connecting innovative designs with existing theory, or using the process of design to develop new theory, researchers can bridge the gap between theory and practice by developing theory-informed, practical solutions.

Before one can use a design experiment to develop theory, a design researcher must establish that the design “works” by producing change in the local context. Evidence of change in these contexts is “necessary evidence for the viability of a theory” (Barab & Squire, 2004). Such assessment makes the researcher “accountable for the results of (their) work to the children
themselves, to parents, to teachers, to local authorities, and, last but not least, to fellow scientists” (Brown, 1992, p. 2). For this reason, design-based research may employ methodologies more commonly associated with traditional intervention research.

However, there are some differences in approach to testing “what works” within a program of design-based research. According to Cobb (2006), rather than a traditional experimental design with a treatment and a control group, “designs advance best when the most promising design options are compared to one another”. He recommends that researchers “narrow the field” to two of the most promising approaches, then compare them directly in a study. This is because establishing one option as effective doesn’t mean another wouldn’t work even better.

If a design experiment is shown to be effective, it can lead to “design propositions” or principles such as “In situation S, to achieve consequence C, do A.” (Plomp & Nieveen, 2013, p. 32). Design interventions that are shown to work at the local level can also be used to develop new theories or validate existing ones (Plomp & Nieveen, 2013). These theories are intended to “transcend the environmental particulars of the contexts in which they were generated, selected, or refined” (Barab & Squire, 2004), meaning the theories can inform work in similar contexts.

Even more than other forms of research, design-based research follows a process of development. According to Plomp & Nieveen (2013) the first stage of research is focused more on content validity than anything else, with less emphasis on practicality or reliability. A researcher examines existing theory and research and develops guidelines or a framework that leads to an initial plan for the intervention.

The initial phase is followed by prototyping or development phase, where the researcher turns his attention to designing an intervention that is practical to implement while maintaining construct validity. The researcher then shifts the focus to the potential effectiveness of the
intervention by developing prototypes and getting expert feedback. This stage may also include trying out the assessment in one-on-one situations.

In the next phase, the researcher assesses whether or not the intervention prototype “works” by producing change at the local level, and whether or not the assessment is sufficiently practical for use by teachers in classroom contexts. Unless the first prototype produces the desired outcome in student learning, a sequence of prototypes are often tried and refined based on student formative assessment data. Data can include a range of sources beyond what is typically considered an assessment, including video and audio. A number of cycles of design and assessment may be undertaken until the educational outcome is realized or until a local theory grounded in patterns of student learning is developed (Plomp and Nieveen, 2003).

This study reports the findings of such a cycle of design-based research, from the development of an intervention, to the prototyping, expert feedback and piloting stage of both the intervention and the instrumentation to measure it, to its refinement and testing, to the reporting of evidence of change at the local level. The study reports the characteristics of an effective intervention to promote the use of historical thinking heuristics as warrants within the argumentative writing of adolescents, and the generation of theory to inform instruction in similar contexts. This initial study sits within a larger line of design-based research in which the intervention will be refined for use with students at the middle school level, leading to further refinements in theory.

While the development of an intervention and the development of instrumentation to measure its effectiveness are traditionally thought of as separate, this study allows the design of the intervention to inform the development of instrumentation and the interpretation of the findings in an iterative manner. In this study, the design of instruction and the measurement were done simultaneously so that they could inform one another and better support the development of theory. The intervention in this study was initially designed based on constructs
developed from how experts use certain heuristics to warrant arguments. As a field, we are still developing an understanding of how students learn to use these heuristics to warrant arguments. Developing both instruction and measurement simultaneously reduces the potential for something to be lost between the construct and the measurement of that construct, and better informs the development of theory around the best way to teach how to warrant arguments with historical thinking.

Though the approach has its advantages, the inherent difficulty in studying written warrants requires careful design and analysis of data. There is always a possibility that students are able to warrant their arguments and think of warrants but do not feel it is necessary to write them down. The explicit targeting of warrants in instruction and warrant writing tasks which cue students to generate warrants heightens student awareness and lessens the chance that they are able to warrant a claim but choose not to. There is also the additional confound of student writing ability. This study utilized a warrant selection and warrant ranking task to see if students could recognize effective warrants and differentiate between more effective and less effective warrants on an assessment that did not involve writing. These tests reduce the possibility that findings related to warrants were not merely a function of their writing ability. They also wrote warrants when given a claim and evidence, making it easier to interpret their warrant writing ability separate from their ability to make claims and provide evidence, and wrote an essay given a prompt, for which they could choose their side and evidence. A sample of students engaged in a think-aloud as they completed a warrant selection and ranking activity and wrote a warrant for a given claim and evidence. This think-aloud gave insight into the processes underlying student use of historical warrants. This range of data gave me a fuller picture of student ability to warrant arguments.

**Research Questions**
According to Barab and Squire (2004), design-based research reflects a “pragmatic” philosophy in which “the value in a theory lies in its ability to produce changes in the world” (p. 6). The overall design-based research question is “What are the characteristics of an effective intervention to promote the use of historical thinking heuristics as warrants within the historical argumentative writing of high school juniors?” As is consistent with design-based research, the goal is to use these findings to support the development of a local theory of how best to teach historical argument writing to students that can inform work in similar contexts.

In order to develop this theory, I need to determine whether or not the approach I am considering works in the local context. Cobb (2006) recommends comparing two promising approaches, rather than a traditional experimental design with an innovative treatment and a control reflecting how history is typically taught. The two treatments in this study include a treatment focusing on teaching students to use sourcing, corroboration and close reading as explicitly stated warrants through the use of warrant generating questions (Freeman, 2011; Toulmin, 1958) and templates (Graff & Birkenstein, 2007), and a comparison treatment focused on finding and evaluating evidence to support particular claims and sides. Both interventions focus on reading of historical text sets and gradual release of responsibility, and offer opportunities for peer discussion and feedback. I first combined the most reliable measures in this study in order to establish whether or not the intervention produces the changes in the local context needed to determine the characteristics of an effective intervention and to inform theory generation about argument writing.

The sub questions of this overall research question follow. The sub questions address specific elements of the difference between the treatment and comparison treatment, along with the type of thinking students do as they complete the treatment and their own experience of the intervention.
Research question group 1. The impact of the treatment on measures related to historical warrants.

1a. Does the treatment focused on heuristics as warrants improve student ability to select effective warrants, in comparison to a treatment focusing on finding evidence to support a claim? 1b. Do students in the treatment condition better discriminate between more effective and less effective warrants? 1c. Do students in the treatment write more effective warrants when given a claim and evidence than students in the comparison treatment?

Research question group 2. Student performance on items designed to reflect specific types of historical thinking and the relationship between skills related to historical argument writing.

2a. What is the relationship between students’ ability to select effective warrants, discriminate between more effective and less effective warrants, and write warrants demonstrating historical reasoning as measured by correlations and patterns in student responses? 2b. How do students perform on warrant selection items meant to reflect specific types of historical thinking? 2c. How do students perform on writing warrants given claims and evidence meant to stimulate specific types of historical thinking? What is the nature of students’ written warrants? Do students include warrants and historical thinking heuristics even without the explicit focus in terms of instruction?

Research question group 3. The thinking of students as they complete the assessment tasks and what they report learning from the treatment.

3a. What sort of thinking does a subset of students engage in as they complete the warrant selection, warrant ranking, and warrant writing tasks? 3b. How do these students feel the intervention is different from typical classroom instruction and what do they report they have learned?
Chapter 1 provided an introduction and defined the purpose of the study. It introduced the theoretical frameworks that guided the study design. It presented the research questions the study is designed to answer about the role of warrants in historical argumentation. The following chapter will review a representative portion of the research that shaped this study. A list of key terms from both the introduction and the literature review follow.

**Definition of Key Terms**

**Argument:** Produced as a product of argumentation, an argument is a statement that minimally contains a claim (an assertion), data (evidence to support the claim) and a warrant (a general statement explaining how the data supports the claim) (Kuhn, 1992; Toulmin, 1958).

**Argumentation:** Argumentation is a process of systematic reasoning in support of an idea or theory, and is an essential element of learning and making new meaning (Erduran, 2007; "Argumentation", n.d.).

**Argumentative writing:** Argumentative writing is the writing of an argument; a key form of creating and sharing new knowledge in the disciplines.

**Close Reading:** Close reading is a process of carefully analyzing an author’s language for three purposes: to learn what a text says, to determine how a text works, and to evaluate its larger significance in relation to other texts and ideas (Adler, 1965; Shanahan, 2013) (Adler, 1965; Shanahan, 2013).

**Content areas:** Content areas are subject areas associated with schooling (history, social studies, science, etc.).
Content area literacy: Content area literacy is a literacy effort attempting to infuse generic literacy skills and strategy instruction in subject area classes, guided by the belief that all reading and writing is largely the same (Shanahan and Shanahan, 2008).

Contextualization: Contextualization is a historical thinking heuristic that includes thinking of the significance of events in their historical context, and using this as a lens for interpretation.

Corroboration: Corroboration is a historical thinking heuristic that includes reading across multiple historical accounts to compare content and identify differences, giving more weight to that information which is common across accounts.

Discipline: A discipline is a field of study where new knowledge is produced, shared and critiqued. A discipline includes types of knowledge, topics of study, the approaches to inquiry, ways of interaction, patterns of language, and specialized ways of reading, writing, speaking, listening, and being.

Disciplinary literacy: Disciplinary literacy is advanced literacy instruction that apprentices students into the literacy practices of the disciplines (Shanahan & Shanahan, 2008).

discourse: language in use (Gee, 1989a)

Discourses: Discourses are established ways of “behaving, interacting, valuing, thinking, believing and often reading and writing” that others recognize (Gee, 1989a). Each discipline has its own Discourse.

Documents model- A documents model is a mental representation created after reading multiple texts. This includes the combination of the situation model of the individual texts and the intertext model of the connections among texts (Perfetti, Rouet, & Britt, 1999).
Heuristics: Heuristics are shorthand mental tools that support historians in reasoning and argument generation. These heuristics have been taught to students as strategies for reading across texts and generating historical arguments.

Historical reasoning: Historical reasoning is a process of analyzing and interpreting evidence to construct historical arguments, often leveraging historical heuristics (Monte-Sano, 2010).

Historical warrant: A warrant is a general rule-like statement used to explain how data supports a claim; also known as an inference-license (Toulmin, 1958). Historical warrants often leverage historical heuristics (e.g. sourcing, corroboration and contextualization) to explain how evidence supports the claim in a historical argument.

Intertext model: In multiple-text comprehension, an intertext model shows documents relate to each other (whether they agree, confirm, contradict, or support) and how a particular document relates to aspects of the overall situation (Perfetti, Rouet & Britt, 1999).

Situation model: A situation model is the creation of a mental image combining ideas in the text with prior knowledge (Perfetti, Rouet & Britt, 1999).

Sourcing: Sourcing includes identifying and evaluating the credibility of a source based on its author or its origin (Britt & Aglinskas, 2002).

Toulmin’s Argument Pattern: Toulmin’s Argument Pattern is an influential model of argumentation including a claim (the assertion one is making), data (the evidence that supports the claim), a warrant (the rule explaining how the evidence supports the claim), backing (factual support for the warrant), a qualifier (the degree of certainty assigned to the claim), and the rebuttal, (exceptions to the claim). The pattern is meant to capture how people actually argue or “reason from premises to conclusions” in social contexts, in contrast to formal rules of logic (Toulmin, 1958).
**Warrant-generating question:** A warrant generating question is a question posed in the dialogic process of argumentation where the person asserting a claim and providing data is asked to explicitly state the general rule or inference license they are using to link their evidence to their claim. A generic warrant generation question is “How did you get there?” (Freeman, 2011; Toulmin, 1958). A warrant generation question for a historical argument prompts students to use historical heuristics, or established patterns of reasoning from evidence to claim. An example for sourcing may include “How do the strengths or limitations of this source type influence how much I will rely on it to support my claim?”
II. REVIEW OF THE LITERATURE

This chapter provides an overview of research conducted within the theoretical framework introduced in Chapter 1 that informed my study. I will begin by situating this study within a sociocultural framework because argumentation is an inherently social process and it helps to capture the complexities of classroom interaction. I then will introduce a model of literacy progression to establish where this study of disciplinary literacy, specifically historical argumentative writing, falls in relation to how literacy has traditionally been examined.

After establishing these general frameworks I will explore the reading and writing connection, because historical argumentative writing depends on both reading and writing about multiple historical texts. I will examine the differences between how novices and experts write. Then I will review theories of comprehension of multiple texts in general and historical texts in particular. I will also examine the historical thinking heuristics historians use to read and reason with texts.

Finally I will review relevant work in the field of argumentation in general and argumentation in the disciplines. I will explain why Toulmin’s argument pattern is a helpful framework to examine historical argumentative writing. Finally I will examine warrants, the element of historical argumentative writing at the focus of this study, and explore how they function in history and can be defined as the place within a historical argument where students make their use of historical thinking heuristics explicit to the reader.

Sociocultural Framework

This study is situated in a sociocultural framework (Vygotsky, 1978), which treats language and learning as inherently social. Individuals use the signs of language as mediating
tools to think about the world (Wertsch, 1991), and they learn about language primarily through social interaction. Street (1984) articulates some of the implications of doing literacy research from a sociocultural perspective, distinguishing between two models of literacy, an autonomous and an ideological model. Literacy is traditionally viewed as autonomous, a set of discrete skills residing within a person, separate from the social practice of these skills (Street & London, 1996). Research from this perspective focuses on isolating the universal elements of literacy that exist separate from social practice and measuring and manipulating them. Illiterate children are viewed as deficient in some way; they lack essential skills. By improving the child’s literacy skills, one can then produce effects on the world.

In contrast, the ideological model asserts that literacy skills cannot be isolated from their actual practice with texts. Literacy is situated in social contexts, where issues of goals, values, beliefs, power structures, and economic and political conditions come into play. As Luke (2000) states, “As long as we locate literacy within human subjects, we will invariably find lack and deficit” (p. 460). A student’s success or failure at reading occurs within a particular social and cultural environment (which includes among other things the education they have received, their literate lives beyond the school, their socioeconomic status, and their own goals and motivations) and cannot be understood separate from that environment (Au, 1987).

While Street’s distinction between autonomous and ideological practices is helpful in understanding what it means to research from a sociocultural perspective, it represents a common pattern in the Discourse of literacy of reducing issues to ‘either/or’ binaries, as opposed to continua where one can simultaneously gain insights from different theoretical perspectives. Binaries such as whole language versus phonics or the current controversy of reading on one’s ‘instructional level’ versus reading complex texts do not advance our understanding of the complexities of literacy practices. These binaries often extend to the methodologies as well; traditionally research and researchers are either qualitative or quantitative, and operate from
either sociocultural or cognitive perspectives, rather than letting the questions they ask guide the methodologies and theoretical frames they choose. While literacy is embedded in social practices, assessing an individual student’s performance on a skill associated with that social practice is also valuable, if informed by the understanding that the skill level is indicative of not only the child’s performance, but the larger milieu. Modern sociocultural theory does not discount the importance of individual cognitive processes, but uses the sociocultural settings an important part of understanding these processes (Davidson, 2010, as cited in Tracey & Morrow, 2012).

A sociocultural perspective makes sense for studying the literacy practices of the disciplines because the disciplines themselves are socially constructed. The disciplines are also shaped within a particular cultural setting, and the processes they privilege vary based on the social needs and priorities of a disciplinary community (Barton, 2007; Gee, 2004; Luke, 2000; Scribner, 1984). The cultural and social histories of the various disciplines have led them to elevate certain practices that align with their values. This context determines the skills we come to associate with literacy in a discipline (Barton, 2007; Lee & Spratley, 2010; Moje, 2008).

A sociocultural framework is relevant for a study of argument in particular because argumentation is inescapably social in nature. Making an argument assumes an audience to receive and be influenced by the message. Argumentative writing is a vestige of oral argument, and argumentative writing is often preceded by oral argument. In frameworks such as Toulmin’s (1958), the different elements of an argument manifest in response to questions asked by a listener interrogating claims and evidence (i.e., “What is your claim? What do you have to go on? How do you get there?”) (Freeman, 2011; Toulmin, 1958). Children use higher-level mental processes such as argumentation first in social situations with peers, before eventually using them independently (Au, 1997; Hynd et al., 2004; Vygotsky, 1978). Argumentation in the disciplines
rests on a long tradition of socially developed and accepted patterns of logic and rules for warrant (Toulmin, 1958).

Another way in which the sociocultural framework fits a study of argumentative writing is the need for scaffolding, or temporary, flexible support that can guide students through more complex cognitive processes. As a complex task, writing a historical argument from multiple sources often requires supports, even for otherwise capable writers (Monte-Sano, 2010). Scaffolds can include things such as setting a purpose for reading, providing guiding questions, breaking a complex task into steps, or giving examples. Vygotsky uses the idea of the Zone of Proximal Development, or the place where students can do more with support than they can do on their own. Once students gain confidence, they can move forward without the same level of support. A perspective that recognizes individual cognitive demands and at the same time considers the social nature of learning is ideal for this study (Smagorinsky, 2001). However, providing just the right amount of support to spur growth, while allowing for productive struggle requires careful planning and responsiveness to student needs. Historians have specific ways they use evidence and theory to make arguments. While teachers tend to emphasize their use of evidence, they are often not specific on the particular ways historians use language to construct their argument, for instance, the way the sentences stating the warrant are constructed. This study provides those more explicit supports to scaffold not only the way historians think, but the way they use language as well.

**A Model of Literacy Progression**

Traditionally, literacy researchers have not focused their attention on the practices associated with particular disciplines. Instead, they have attempted to identify and examine universal elements of literacy that can be applied across them. However, since the 1990s, there has been a small but steady body of research that examines the specific literate practices of the
disciplines themselves, with a focus on what they have in common and what distinguishes them (Shanahan & Shanahan, 2008). While acknowledging that universal elements of literacy exist, this research has underscored that literacy is not one thing spread evenly across the content areas (New London Group, 1996). It hones in on the specific discourse practices that a disciplinary community uses as the ones to scaffold for students, as opposed to the generalizable processes common in all literate activity.

Part of the reason why adolescent readers continue to struggle is because of the ways we have traditionally conceptualized literacy development as a field. Traditional developmental models of literacy emphasize the earliest stages, implying once these crucial periods are passed, students will be successful readers for life (e.g. Ehri, 1987; Piaget, 1997). They also fit with the pervasive “simple” (Hoover & Gough, 1990) views of reading as Reading = Phonics x Listening Comprehension. This emphasis is apparent in our achievement data. In fact, according to the NAEP assessment, the majority of 17-year-old students have achieved Basic levels of reading. There are no significant differences between groups based on race/ethnicity or SES between students categorized as Basic. However at the Proficient level, there is a large achievement gap along the lines of race/ethnicity and SES. Interestingly, at the Advanced level, this achievement gap disappears because hardly anyone is able to comprehend complex text - Just fewer than 10% of 17 year olds demonstrate comprehension of complex texts (Lee & Spratley, 2010). This pattern indicates that while many students have achieved basic levels of literacy, there remains an achievement gap along the lines of race/ethnicity and SES at the Proficient level. However, students regardless of race/ethnicity and SES struggle with the most complex, disciplinary text (Lee and Spratley, 2010). This pattern also reflects the focus of instruction on basic literacy and generic comprehension strategies, but not the specialized ways of reading associated with complex, disciplinary text.
In contrast to overly simplified views of what it means to read, Shanahan and Shanahan (2008) present a Model of Literacy Progression that is more nuanced and articulates how many elements of literacy continue to develop over the course of a lifetime.

**Stage 1: Basic literacy.** At the first stage, basic literacy, students learn universal literacy skills such as concepts of print, decoding and sight word recognition, and build fluency with simple, short text. The National Reading Panel report showed strong evidence that certain universal skills such as phonics and phoneme awareness are important in teaching children to read (NICHD, 2000). The authors of the National Reading Panel Report (NICHD, 2000) went out of their way to emphasize that while this work is important, it does not constitute an effective reading program, though the degree to which that message was actually heard is debatable (Pearson, 2006). There has also been debate about the degree to which basic literacy skills, which are “constrained” in the sense that they do not continue to develop over the course of a lifetime, should be emphasized or whether “unconstrained” skills such as comprehension, vocabulary and writing that continue to develop over the course of a person’s lifetime, should be the focus (Paris, 2005). However, for certain windows of time, targeted instruction with basic literacy skills can be particularly effective and does correlate with future literacy levels. While important, these basic literacy skills represent only the first steps on a child’s literate journey.

**Stage 2: Intermediate literacy.** Stage two in the model of literacy learning consists of generic comprehension skills and reading strategies, strategies for learning more complex academic vocabulary, and improved reading fluency. The generic reading strategies that compose much of intermediate literacy learning are associated with research based on a psycholinguistic model of reading processes. This research began as studies of expert readers and contrasted that with the way novices read, in an attempt to identify the universal cognitive processes readers used to process information (Guzzetti, 1984). This cognitive, experimental research done from the early 1970s to the late 1980s was distilled into strategies for classroom
use. These strategies were thought to provide the answer for how students should read the texts they encounter, regardless of complexity or discipline. The influence of this work was seen in the push for content area literacy, with “every teacher a teacher of reading” (Moore, Readence, & Rickelman, 1983). This work focused on generic strategies and processes in content area classes, and based on the premise that the cognitive processes underlying all reading are pretty much the same (Shanahan & Shanahan, 2012). Examples of generic literacy strategies from this period that were applied to content area fields included Question-Answer Relationships (QAR) (Raphael, 1982) and K-W-L (Ogle, 1986). While these approaches help students learn, they are not specialized to the demands of disciplinary text. Similar approaches were taken in writing, with writing to learn strategies that they employ the same techniques (quick writes, journaling, etc.) regardless of the discipline.

While this research has been helpful in supporting struggling readers it is also limited. Many strategies informed by a cognitive processing framework show conceptions of students as ‘information processors’ who work individually to get information from text as opposed to reasoners, arguers, collaborators or users of multimodal text (Alvermann, 2001), processes which typify expert practice in the disciplines. Content area teachers often resist using “content area reading” strategies in their instruction, seeing them as an interference to covering content or out of step with their conception of learning in their discipline (O’Brien et al., 1995). Rather than actually allowing students to struggle through disciplinary text, many teachers end up telling students what the text says (Moje, 2008; Schoenbach, Greenleaf, Cziko, & Hurwitz, 1999). However, just getting the main idea or the gist of content is of decreasingly little value in today’s world, where information is available at a click of a button. Workers need to be able to constantly apply ways of thinking to new problems, learn and relearn skills, and seek needed knowledge for themselves. Content area reading strategies ignore the important differences in the
ways in which reading, writing, speaking, listening and knowing are used in the disciplines, and do not empower students to critique or produce their own disciplinary text.

**Stage 3: Disciplinary literacy.** In contrast to intermediate literacy that privileges the generic, disciplinary literacy involves using the specific literate practices employed by disciplinary experts to read, produce and critique disciplinary text. The research on disciplinary literacy primarily comes from three fields: expert/novice studies of the reading of disciplinary experts such as historians, scientists and mathematicians (e.g. Rouet et al., 1997; Shanahan et al., 2011; Wineburg, 1991), discourse perspectives (e.g. Gee, 1989; Luke, 1995; Moje, 2007; Moje, 2008), and systemic functional linguistics (e.g. Fang, 2012, Schleppegrell, 2007). These practices have been identified through research on expert readers in the disciplines (expert/novice), through research on how language is used in the disciplines (discourse) and through examining language itself in the discipline (systemic functional linguistics) and all indicate significant differences across the disciplines. The first two perspectives are most relevant the focus of this study, so systemic functional linguistics will not be reviewed. Expert-novice and discourse perspectives reflect how experts use argumentative writing and are situated in the discourse practices in which students should engage.

**Expert/Novice Studies**

In the 1980s a line of research began to use knowledge of subject matter as an important lens for understanding teaching and learning (Reisman, 2012). This research recognized the disciplines as having distinct ways of understanding and learning about the world. Experts in different domains possessed specific types of knowledge, namely declarative knowledge (factual knowledge or “knowing what”), procedural knowledge (knowledge of domain-specific processes and strategies or “knowing how”), and conditional knowledge (knowledge of “when and where” to access facts or employ processes/strategies) (Alexander & Judy, 1988). While expert-novice studies were used in cognitive science and were the source of many of the generic reading
strategies used in intermediate literacy, beginning in the 90s, a small group of researchers began
to conduct such studies to understand expert uses of literacy in the disciplines. They hoped that
by having experts think aloud as the read disciplinary text, they could distill strategies to support
novices’ access challenging disciplinary text. Such studies (e.g. Latour & Woolgar, 1979; Rouet
et al., 1997; Shanahan et al., 2011; Wineburg, 1991) determined that there were indeed
differences between the strategies experts and novices used when approaching disciplinary text.

For example, Wineburg (1991) contrasted historians’ think alouds with those of high
school students, identifying the heuristics of sourcing, contextualization, and corroboration.
Historians use these tools to move beyond what is explicitly stated in text, constructing
interpretations of events and testing models of what may have happened at a historical event.
Sourcing includes identifying and evaluating the source of a document, including the perspective
of an author and his claims to authority, which in turn influence how one interprets the content of
the document (Britt & Aglinskas, 2002). Contextualizing includes situating events historically,
drawing on declarative knowledge of what was happening in the world and participant
interpretations of events to understand their historical significance. Corroboration consists of
reading across multiple historical accounts to compare content and identify differences, giving
more weight to that information which is common across accounts. Interestingly, historians use
these tools profitably even when they do not possess knowledge about a particular event, drawing
upon their understanding of the nature of how knowledge is created, shared, and judged in the
discipline to shape their interpretation (Wineburg, 1991). In Wineburg’s study (1991), high
school students did not use heuristics. They struggled to evaluate the source of documents,
seeing the textbook as the most reliable source. While they could employ generic strategies (e.g.
finding the main idea, making predictions, determining importance, making inferences) they did
not acknowledge the differences in documents, and their interpretations were guilty of
presentism, or applying present values and perspectives to past eras. Again, this is not surprising
because such disciplinary skills are rarely taught and students have limited opportunities to struggle productively with complex disciplinary text.

Additional studies (Hynd, 1999; Rouet et al., 1997) have found that students do not use heuristics spontaneously (Wineburg, 2004). They have a tendency to read documents as individual pieces and not an integrated whole, and pick and choose which information they will consider, while ignoring contradictions. While recent interventions have had some success in improving students’ ability to source (Britt, 2002) and corroborate (Nokes, Dole & Hacker, 2007) contextualizing remains a challenge, as it requires a great deal of background knowledge (Wineburg, 1991; Nokes et al. 2007).

There have also been studies that have compared experts across the disciplines. Shanahan, Shanahan and Misischia (2011) conducted think-aloud protocols and focus groups to examine the strategies used by expert readers in three disciplines: History, Math and Chemistry. They analyzed transcripts and cross-coded using the heuristics common to each discipline. Along with helping understand how strategic behaviors differ, the results gave insight into underlying epistemologies of the disciplines as well. For instance, in math, rereading was essential because there was a high value on accuracy and precision. The value in knowledge came from being “true” or generating an error-free proof. Close reading was also essential, as even functional words such as “an” and “the” took on great significance and changed the meaning of the proof, for instance, in naming “an” angle (meaning any angle) in a geometric figure, versus “the” angle, meaning a particular angle. Compared to the disciplines of science and history, the heuristic of sourcing was less important; if the numbers worked out, the author was irrelevant and was not used to interpret the writing (Shanahan et al., 2011). Whereas in history; knowing an author was a Southern apologist would make a huge difference in judging their interpretation of a primary source document related to the Slave Codes. Given the way that these different
heuristics are applied in different disciplines, students need lots of instruction and practice with these patterned ways of thinking and reasoning in order to take ownership of them.

**Discourse Perspective**

The differences between the disciplines can also be understood through the lens of an examination of discourse. Discourse studies emanating from the New Literacy Studies (NLS) are a strong contributor to the field of disciplinary literacy. The New Literacy Studies (NLS) was an approach taken by researchers who examined literacy use in everyday life. Building on sociocultural theory, they used methodologies from anthropology, specifically ethnography of communication, to capture rich detail about how people actually use language. This focus on language-in-use caused scholars to recast literacy from a psychological, internal process of acquiring discrete skills taught in schools to read books, to a social, collaborative practice of making meaning with many different semiotic resources (Barton, 2007; James Gee, 1989b; Street, 2003) (Gee, 1991; Street, 2003, Barton, Hamilton & Ivanic, 2000). Since Gee (1989) defines literacy as the mastery of at least one secondary Discourse, a student whose home discourse does not align with the schooled Discourses faces serious obstacles. Delpit’s (1992) solution to this conundrum is that the “rules” to acquiring a Dominant discourse can and should be taught; accessing codes of power is an important outcome of schooling, especially for students from nonmainstream backgrounds. Acquiring Discourses of power does not imply assimilating to them- indeed, once Dominant discourses are learned they can be critiqued and transformed (Delpit, 1992).

Being able to navigate across home and disciplinary Discourses (the ways of thinking, talking, and being of each discipline) (Gee, 2004; Lemke, 1990; Moje, 2007) helps students become metadiscursive (New London Group, 1996). “To be metadiscursive means that people not only engage in many different discourse communities but know how and why they are engaging, and what these engagements mean for them and others in terms of social positioning.
and larger power relations” (New London Group, 1996). The work of teaching for disciplinary literacy is thus teaching for social justice because it has the outcome of helping students become empowered (Moje, 2007).

The areas of expert reader studies and discourse perspectives have contributed pedagogical approaches (see Moje, 2007 for a review), yet these approaches are not widely employed, particularly with students in poor, urban areas. While disciplinary literacy shows much promise, as a young field there are relatively few examples of studies that meet the criteria for “what works” questions. Approaches either have not or are just beginning to show evidence of a corresponding positive effect on standardized assessment scores (for an example of such evidence, see Reisman, 2012).

In an attempt to secure funding tied to test scores, many schools in poor, urban areas have resorted to “drill and kill” over the narrow band of skills covered on current high stakes tests, investing in remedial curricula designed to fill in skill gaps. Ironically, the kinds of instruction that would give students access to greater opportunities and financial rewards are often considered a luxury reserved for more affluent areas (Ladson-Billings, 1994). Teaching only what is easily measured through standardized tests provides sparse opportunities to develop the disciplinary literacy skills essential for life in the 21st century. What results is instruction that to students seems irrelevant and disconnected from their lives (Moje et al., 2004). Students never get the opportunity to engage in the real work of the disciplines (Schoenbach et al., 1999).

Next-generation assessments such as the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced show promise in that their performance assessment components as more closely approximating the kind of thinking students will need to do in the world of college and career. However, there has been a great deal of push back against these assessments for a variety of reasons. First of all, funding has been tied to states adopting these more rigorous assessments, making it seem to some as though the federal government is
trying to exert undue influence over education. Also, teacher evaluations have been tied in part to
student performance on these new assessments, most notably in New York State, causing uproar
from many teachers and their unions. Finally, more students have been identified as struggling
on these new assessments, leading to parents to feel they want to return to the days of easier
assessments and higher proficiency rates.

Ironically, an achievement gap has existed for years, and it is only now that middle class
parents are being told that their children are not college and career ready that there is a broad
outcry and a growing “opt out” movement, as the issue is hitting closer to home for these
influential groups. Meanwhile many leading civil rights organizations are strong supporters of
both standards and rigorous annual testing. However, some are concerned that the more
challenging assessments will only widen existing achievement gaps and further limit
opportunities for nonmainstream students.

One of the reasons why performance on next-generation assessments is weak is that
teachers are still figuring out how to teach students to read and write in the ways demanded by the
standards. Rapid adoption of the standards led to professional development and instructional
materials trailing behind what teachers needed to feel effective (Association for Supervision and
Curriculum Development [ASCD], 2012). Thus, studies such as this are timely and important
because they may yield recommendations for instruction that may support students in writing
arguments from complex, disciplinary texts, which is a key part of the standards themselves.

Disciplinary literacy instruction has the potential to bridge the opportunity gap for
students from disadvantaged backgrounds by shifting the emphasis for teachers away from
remediation or training students in universal processes. Rather, teachers of the discipline become
aware of the ways in which they use reading and writing to make and convey meaning as an
expert in their discipline (Achenbach et al., 1999), and find ways to scaffold these practices for
students. Opportunities to engage in real-world disciplinary literacy practices such as
synthesizing scientific studies about the benefits of vaccination and comparing this synthesis to messages in the media will help students apply disciplinary literacy skills to improve their own lives. Evaluating competing claims in a policy argument about immigration reform in the Presidential Primary debates, or “reading” reports from different outlets about free community college for bias and authors intent are examples of the kind of thinking that will help students become empowered as citizens. By emphasizing the advanced literacy skills needed to handle complex, disciplinary text and key literate practices, students can undertake work that is engaging and relevant to their lives as young people, while developing the skills they need to be successful in college and career.

**Reading and Writing Connection**

Literacy, encompassing both reading and writing processes, is a relatively new construct. It only came into regular use in the late sixties and early seventies, with the emphasis on the acquisition of language as a natural process (Halliday and Hasan, 1976). Despite the historical and empirical recognition of the connection between reading and writing, the processes still remain troublingly separate in the way they are theorized, researched and taught (Fitzgerald & Shanahan, 2000). The consequences play out in education where reading and writing are often separate subjects, taught by separate teachers, and where schools have a tendency to delay writing until students can read with confidence (Gesell, 1925, as cited in Cunningham & Shanahan, 2000). For instance, separate committees developed the Illinois Reading and Illinois Writing standards.

The Common Core State Standards do not share such a division and acknowledge the connection between reading and writing (with the emphasis on the centrality of evidence-based reading and writing), though the standards for both processes remain separate (NGACPB & CCSSO, 2010). The division of reading and writing instruction is particularly problematic in the area of disciplinary literacy, where the texts students read and write about become increasingly
more complex, and the texts students write become more specialized. Recent attempts to
determine the “value added” of teachers, or measure the impact of particular teachers on student
test scores in Chicago and other states have considered evaluating history teachers using the
reading score, while English (and writing) scores would be used to evaluate Language Arts
teachers.

There has been more than a half-century of research connected to the shared knowledge
and cognitive processes of reading and writing. From this theoretical perspective, reading and
writing should develop in tandem, their measures should be strongly correlated, and teaching
them in combination should lead to better learning outcomes. This approach gained steam with
the cognitive revolution in the 1970s that promoted an image of a reader actively constructing a
text in his or her mind as he read. This paralleled nicely with the process of text composition
(Tierney & Pearson, 1983, as cited in Fitzgerald & Shanahan, 2000). While readers form a
mental representation of thoughts written by someone else, writers formulate their own thoughts,
organize them, and create a written record of them using the conventions of spelling and
grammar. Though there is clear evidence that writing can be a vehicle for improving reading and
especially comprehension across content areas (e.g. Graham & Hebert, 2010), and that the time
spent writing their own texts can improve students’ ability to read texts written by others, it is
also clear that the same types of instruction do not improve reading and writing equally, and both
reading and writing require individual attention (Graham & Perin, 2007).

Fitzgerald and Shanahan (2000) describe a number of different types of knowledge upon
which reading and writing are thought to rely. Several of these areas are relevant to this study,
including “domain knowledge about substance and content” which includes both prior knowledge
and knowledge gained from reading and writing texts. It is similar to the knowledge that
Wineburg (1991) found historians to have about reading historical texts, which they could employ
regardless of their knowledge of the particular event in question. Another relevant type of
knowledge that applies to both reading and writing is procedural knowledge, which includes strategic behavior such as questioning or making predictions (Fitzgerald & Shanahan, 2000).

Fitzgerald and Shanahan (2000) articulate three relevant categories of research on reading and writing. The first emphasizes reading and writing as part of a process of communication (rhetorical relations). Students learn about this communication process by sending their own messages in writing and receiving messages from others through reading. The second approach treats reading and writing as functional procedures that can be combined to accomplish some academic goal. An example is examining how reading works in the process of revising papers (Beal, 1996, as cited in Fitzgerald & Shanahan, 2000). The third approach emphasizes the knowledge and cognitive processes that reading and writing share (Fitzgerald & Shanahan, 2000).

A meta-analysis of writing to read (Graham & Hebert, 2010) underscores the strength of the relationship by showing strong and consistent effects for many types of writing to learn activities. Another recent meta-analysis by Graham and Perin (2007) of learning to write includes some shared processes such as summarization, which underscores the relationship as well.

Reading and writing are often seen as mirror images of one another, though this is an overly simplistic view. While reading and writing depend on overlapping knowledge, skills and strategies, they are also significantly separate (Tierney & Shanahan, 1991). In studies of young children, Beringer and her colleagues used multiple measures of constructs and were able to explain upwards of .70 of performance. The stages of reading and writing development are often parallel, but the distance between the two parallel lines changes over time (Fitzgerald & Shanahan, 2000). This is also demonstrated by the fact that good readers are not always good writers, and of course the purposes of reading and writing differ. Attempting to find correlations between single reading and writing variables is limited in part by deficiencies in tests to capture reading and writing. Bivariate relationships testing only spelling or only vocabulary are difficult to establish, due to third variable problems and the lack of extensive multivariate analyses.
Nonetheless, since the processes are not identical, their combination is promising as it may result in more than the sum of its parts (Fitzgerald & Shanahan, 2000).

The reading and writing connection is particularly relevant to a study of historical argumentation, which is a process of reading across multiple texts, including primary and secondary sources, and constructing an argument by synthesizing these sources. Document-based writing in history requires strategies for reading texts with the outcome of writing a historical argument in mind. Readers must “reason about documents” by reading and making a mental representation of them; they must recognize and analyze conflicting claims, and they must evaluate documents produced by authors with varying motivations and levels of credibility, corroborating these sources with other texts. They must then integrate the evidence to construct their own argument (Kuhn, Weinstock, and Flaton, 1994; Rouet et al., 1996).

Writing experts critique the fact that writing is often taught as a universal process where the basic steps involved are nearly identical regardless of discipline and purpose (Hillocks, 2005). Rather, the process of writing is specific to the task in which students are engaged, meaning writing a historical argument differs significantly from writing a letter to the editor. Reading documents for their arguments, claims and warrants should help students subsequently craft their own arguments and give them more fluency with that structure to employ it in their own writing.

Knowledge Telling and Knowledge Transforming

Scardamaila and Bereiter (1987) found several differences between novice (elementary) and experienced (college-aged) writers and how they generate and interact with texts. They categorized these as the difference between knowledge telling and knowledge transforming. Knowledge telling characterizes novice writers who are heavily affected by their own knowledge and experiences. The writer begins with a task and topic ideas, and retrieves information that reflects “the straight-ahead form of oral language production, (requiring) no significantly greater amount of planning or goal setting than does ordinary conversation” (p. 9). Novice writers
generally write within a familiar genre. The focus is putting words on the page. They see no need for a plan or goal; they don’t have to consider a particular problem; they just write. All their energy is focused on recording their thoughts, even if the length or complexity of the task increases or if more time is made available. When given time to take notes, they tend to transcribe them directly into their writing. As soon as they get the words down, they feel finished. Knowledge telling shows a lack of writing strategies, little evidence of reasoning about content, and little rhetorical use of language or awareness of an audience.

Knowledge transforming characterizes more expert writers. According to this model, a writer’s knowledge and the text they are writing interact, creating a problem-solving process in which knowledge is transformed during the thinking and writing. The text produced is based on what a writer already knows and also on ideas and knowledge that are generated as text is created. Since experts have greater knowledge of the discipline in which they are writing, they can set the appropriate goals based on disciplinary norms. They use writing strategies independently and flexibly based on their need (Scardamalia & Bereiter, 1987; Graham & Perin, 2007). Skilled writers spend more time planning and revising and are actively engaging in researching, questioning and rethinking so that their initial draft is evaluated and revised to meet the needs of the audience and their own goals.

In knowledge transformation, writing is more than just self-expression. It is a part of critical thinking and problem solving. One problem authors need to solve is how to achieve their purpose given their audience and the norms of the discipline in which they are writing. Writers need to develop ideas, identify the information needed, and access that information (Hillocks, 1987). Depending on the task, different skills come to the fore. One of the explanations of the difference between novice and expert writers is that novices don’t have the content-based reasoning or research skills needed for knowledge transforming (Kuhn et al., 2004, Wiley & Voss, 1999).
Research informs some of the ways students may be supported in knowledge transformation. Support includes activities to make the audience clearer, framing to make the task clearer, and scaffolding to support expert thinking processes. A disciplinary literacy approach to argumentative writing could explicitly address these areas, helping novices develop as writers. More research needs to be done on how to support students’ knowledge transforming in the disciplines, however. The distinction between knowledge telling and knowledge transforming is relevant in this study because knowledge transformation is the target for students’ historical arguments and these tasks support expert thinking processes.

**Multiple Text Comprehension**

Multiple sources are important to historians because they realize history is constructed and historical accounts are “speech acts” (Stahl & Shanahan, 2004) written by human beings for particular purposes. For this reason, a single source has limited value. In order to understand the event, multiple voices must be considered. The very nature of historiography requires reading multiple texts (Stahl et al., 1994). Thus, before delving further into a review of the research on argumentation, it is valuable to examine the theories of comprehension in general and multiple source comprehension in particular that inform this research.

**Foundational research on comprehending single texts: Situation model.** Initial theories of text comprehension in cognitive science (e.g. van Dijk & Kintsch, 1983; Kintsch, 1988) focused on how a reader makes meaning from a single text. Ideally, a reader uses text to create a mental model of the concepts described within it. Early research had used students’ memory of a text as a stand-in for their comprehension of it, which many researchers found misleading. Thus, van Dijk and Kintsch (1993) established a more nuanced *situation model* theory of comprehension that indicated three levels at which a text could be understood. The first level includes the surface structure, where the reader encodes textual features such as words, sentences and the layout of the text. The second level is the textbase, which is the literal meaning
of the words. The third level is the situation model, a deeper level of understanding which
includes taking the textbase level and expanding on it through inference, integrating it with other
texts and understandings from the reader’s own experience and beliefs. This helps her
comprehend text by developing a mental representation of what she is reading, beyond just the
words on the page. Creating a situation model elaborates the difference between the content of
the text itself and understanding the situation it describes. For struggling readers, constructing a
situation model of even a single text can be challenging, and can be supported by the use of
explicit strategies, such as Questioning the Author (Beck et al., 1996).

There are two categories that influence the construction of a situation model – what the
text brings (e.g. the structure) and what the reader brings (e.g. the reader’s own prior knowledge,
goals for reading, and strategy use). Within a reader’s schema, a single text actually contains
connections to many other texts and relevant prior knowledge gained from previous reading of
the world and the word (Hartman, 1995), including ideas about how that type of text typically
functions, the role certain sorts of authors often play, and the predominant characteristics of that
genre, with links to impressions of other related texts (Manderino, 2011). The thoroughness of
one’s situation model helps determine whether one simply memorizes a text or actually learns
from it. While many readers can recall and even summarize a text sufficiently by relying on just
the textbase, this doesn’t mean they actually understand the text. Exclusive focus on the words,
sentences, and external text features limits the degree to which students can interpret a text.

If the goal were just to have students remember what is read and construct a textbase, we
could focus on supplying students with easy-to-process texts which, for instance, summarize the
main point of a paragraph in a topic sentence (Beck & McKeown, 1991). These “considerate”
texts could promote readability and recall and the formation of a textbase. However, the kind of
reading that students must do to be literate in the discipline of history is much different than that.
Comprehension must be thought of in terms of the more complex situation model due to the challenges of multiple perspectives, conflicting accounts, and the incomplete historical record.

Instructional moves such as assigning inference tasks or providing less coherent, more difficult texts can lead to a more robust situation model, though readers may not recall as much from the text itself, since their cognitive resources are not solely devoted toward recall (Wiley and Voss, 1999; Kintsch, 1994). Tasks that promote memory of information are not necessarily the ones that promote understanding, and the tests and quizzes that focus on recalling facts “may actually be biased against students who engage in deeper processing” (Wiley and Voss, 310). Paradoxically, reading can be too easy to be effective. If readers encounter no challenges that encourage them to create a situation model, the new information won’t be linked to their existing knowledge, and, while it may be remembered for the short term, true learning won’t take place (Kintsch, 1994, p. 302). A growing understanding of the importance of opportunities to engage with complex text has led many experts in the field to question the common practice of placing students in “instructional level” text which they can comprehend with relatively little teacher support, and a call for giving students opportunities to engage with text that would have previously been considered “frustration level.” There is a growing body of research to indicate more complex text (with additional scaffolding from teachers) can actually accelerate student progress (Shanahan, 2011).

**From single to multiple texts.** While the situation model theory of text representation accounts nicely for what happens cognitively when a reader makes sense of a single text, it does not account for multiple text reading, much as cognitive strategies developed from single-text comprehension models are not sufficient to help teachers support students to integrate and synthesize knowledge across texts. Since reading across texts is a central part of writing a historical argument, I will review research that explicitly addresses multiple texts, along with the instructional strategies that stem from such research.
Due to the quantity and sophistication of multiple-text analysis it requires, the discipline of history has been used as a proxy for studying the ways in which experts and students read across multiple texts. Wineburg’s (1991) work in the field of cognitive psychology built on the expert/novice studies of single text comprehension by examining how historians and high school students read across historical texts through think aloud protocols. There were not huge differences in topic knowledge between the two groups in terms of basic facts and dates. In fact, two of the high school students actually knew the answers to more identification questions of the topics. However, historians used particular heuristics and perspectives toward text that caused them to arrive at very different meanings than the high school students. Historians understood the texts as constructions, reading not just the text as written but the subtext (what the text didn’t say but what could be inferred about it) to create an event model of what they think happened. Historians used heuristics such as sourcing, which includes examining the perspective of the author and considering his or her influence on the document’s construction, including what is emphasized and what is left out, to interpret events. They used a corroboration heuristic to compare multiple accounts of the same events, interpreting the similarities and differences to make judgments about what may have happened. They used a contextualization heuristic to interpret when, where and for what purpose a document was written, using this context to better understand a document and its significance. While sourcing and contextualization are not explicitly multiple text strategies, sourcing requires an understanding of the types of sources typically examined in history (public records, diary entries, speeches, etc.) and a nuanced approach to contextualization requires understanding of multiple accounts that make up the expert’s knowledge of that historical period, often including maps, images, and other types of primary, secondary and even tertiary accounts.

The Documents model. Perfetti and his colleagues built upon Wineburg’s work to develop a comprehensive theoretical model of what goes on in one’s head when reading multiple
texts (Perfetti et al., 1999). The **Documents model** explains how readers make sense of information across multiple texts, including how they understand where information overlaps and contradicts itself. When a reader encounters multiple texts, she has to make some decisions about the content of her emerging mental model. She either ignores new information or makes deliberate choices about what to include (Britt, Rouet & Braasch, 2012). For instance, a reader may realize authors are both the source and the participant in events recounted in a document.

When examining a source, readers may also consider other information such as what was happening at the time period something was written, or who published the work, elements not considered in Kintsch’s model. The documents model has two interconnected components: The **intertext model** that shows how the documents relate to each other (whether they agree, confirm, contradict, or support) and how a particular document relates to aspects of the overall situation. The **situation model** addresses the situation referred to by each of the individual texts and may represent meaning across all of the text as a whole, or include several related situations from the texts, depending in part on the coherence of the text set and the purposes of the reader. In many instances, the reader reads across documents to arrive at the most thorough, accurate model of an overall situation. Since each text can present a different situation, there may be some situations that are incompatible. Authors of individual documents may be arguing that events should be interpreted in a certain way or that some information is relatively more important. Different documents might refer to the same events, but one author might say that some events caused others, or made a greater contribution to causing other events (Perfetti et al., 1999).

By working through the situation model, readers often decide on the best representation of what happened and why, and determine the documents that best support that version of events, or the reader’s own beliefs. They may also come up with different theories of what happened and the evidence for each. When the intertext model and the situation model are connected to each other, the documents model is considered complete. The relationship between the intertext model
and the situation model is iterative—those connections help shape the understanding of the situation as it is being developed in the mind of the reader (Perfetti et al., 1999).

Within the intertext model, readers construct a node or connection hub for each document and label the links between documents and their situation. Every node potentially contains information about the source, the rhetorical goals of that document, and the content. The three types of information in these nodes represent potential ways a reader can draw connections across multiple texts. Sourcing is the most elaborate of the three types of information, with separate slots for author, setting and form. Rhetorical goals include an author’s intent, or why the argument was written, and the intended audience, which often needs to be inferred by the reader. The content node includes either a statement of the main idea for an expository text, or a thesis statement for other texts such as an argument. In reality, students may only partially utilize the nodes, whereas experts would regularly use all nodes (Perfetti et al., 1999).

The intertext model also includes “intertext predicates,” which show the relationship between documents. This relationship is often stated in terms of whether it agrees with or contradicts, or gives evidence for or against. This can also include more nuanced relationships such as “incremental, familial-temporal, intellectual, or aesthetic” (Perfetti et al. p. 94). While these relationships aren’t the only way documents can be connected, the relationships among documents tend to be remembered even when the nuances of content, source, and rhetorical goals may be forgotten. An expert reader holds these relationships in their head and uses them to help construct their working theory of the situation model.

If expert writers were to synthesize several documents, they would include explicit markers of these relationships in well-written essays. They would do things such as cite other documents and acknowledge where documents confirm and contradict each other as part of their reasoning. Less capable writers are unlikely to build good documents models because they don’t have a coherent map of the document space and the relationships within and across documents.
This lack would be apparent in their writing (Perfetti et al., 1999). Since multiple text situations provide so many opportunities to make connections, readers also need to be more selective about what connections and explanations to make within and across texts (Lawless, Goldman, Gomez, Manning & Braasch, 2012).

Warrants are particularly connected to the Documents model because historical claims are often claims about a particular interpretation of what happened. Because of the interpretative nature of history, these can best be supported not by citing a single document or piece of evidence, but by explicitly stating the relationships among documents, or sharing an intertext model. Asking students to warrant historical arguments is in essence asking them to integrate their intertext model with their situation model, creating a more effective documents model that leads to better comprehension and also clearer writing and reasoning. Now that I have examined the theories of multiple text comprehension based on how experts engage with multiple texts, I will share research about students engaging with multiple historical texts.

**Students’ engagement with multiple historical texts.** Focusing on single text comprehension strategy work in history classrooms can mislead students about the nature of reading in history and discourage students from engaging in authentic disciplinary work (Shanahan & Shanahan, 2008). Stahl and his colleagues (1996) argue for the importance of multiple texts in building both content knowledge and disciplinary knowledge. While the traditional ‘one textbook’ approach sends the message that history can be learned from a single, authoritative source, multiple conflicting perspectives help students in creating a more nuanced mental model, forcing them to make links across texts and view the same event from different perspectives, while also improving their content knowledge. Treating textbooks, lectures, primary sources, maps, worksheets, etc. as historical accounts worthy of interrogation helps emphasize the constructed nature of history (Bain, 2005). Though developmental differences exist, reading multiple historical texts with the goal of creating a situation model of a particular
event or writing an explanation from a particular perspective is a challenging process for non-expert readers, regardless of specific grade level. It has been theorized that using multiple sources in a task may help students to construct a situation model. While single text comprehension limits a textbase to a single text, if more than one text is encountered, readers are forced to make connections and integrate information in a way similar to the way an expert reader does (Rouet et al., 1997; Wiley & Voss, 1999).

However, research suggests that students don’t construct document models automatically, even given multiple texts about the same event. High school students who are historical novices tend to read texts as discrete entities. If they notice similarities, they tend to emphasize those and ignore information that is unique, incomplete or contradictory (Stahl et al., 1996; Perfetti et al., 1999). The same pattern held true for graduate students who were historical novices in comparison to historical experts (Rouet et al., 1997). They may create a situation model for individual texts, but they don’t make a robust intertext model of the connections across texts, or leverage the expanding knowledge of texts as they encounter new texts to refine their situation model. Perfetti et al. (1999) argue that initial disciplinary knowledge influences whether or not students create a comprehensive Documents model. Students with less disciplinary knowledge may create a simple narrative of events, whereas students with disciplinary knowledge tend to emphasize the relationships between documents. When high school students attempt to create an overall situation model, they tend to stick with the impression they gained from the first one or two documents (Stahl et al., 1996). There is also evidence of a recency effect - students may prioritize information from the last document read.

While they can detect overt bias in historical documents, students tend to see the bias in documents as a source of weakness, preferring the “factual” tone of textbooks (Perfetti et al., 1993; Rouet et al., 1998; Stahl & Hynd, 1994). Whereas historians put limited value in any one source, students tend to look for the document that they feel is most valid or least biased, and
incorporate it into their situation model, failing to recognize that even the “best” document presents an incomplete view of the event. They fail to realize that a biased document may still be a great source of evidence depending on their purpose for reading or the argument they may be constructing. Think alouds confirm that the choices students make are often naive and based on ideas about someone being there or the authority of a textbook, the balanced tone, etc. Even when they revise their situation model by reading new texts, they don’t connect it to the intertext model (Manderino, 2011).

Some research has been done to document the developmental trajectory of learning in history, from students viewing history as one authoritative account, to thinking everything is just relative and “anything goes,” to understanding that while historical texts are constructed, there are better interpretations based on evidence, inference, and reasoning. The end goal may be an understanding of knowledge as constructed, moving students along the continuum of history learning (Lee & Ashby, 2000; VanSledright, 2002).

**Heuristic Use and Students**

Heuristics are tacit tools for thinking or solving problems. They are often short cuts for reasoning about a situation or issue based on experience. Wineburg (1991) identified key heuristics by listening to his study participants think aloud; the historians themselves did not have names for what they were doing. They had patterns of thinking that were developed over long periods of time working with texts in the discipline, based on their expertise. Students typically do not use heuristics for comprehending multiple texts without being taught to do so (Monte-Sano & De La Paz, 2012; Wineburg, 2001). When students do begin to source, they may initially use relatively naive strategies such as putting more stock in the perspective of someone who was actually there, as opposed to considering how they be motivated to present facts in a certain way to protect colleagues or present themselves in the best possible light. They may work backwards, figuring out the argument they want to make first, then finding the evidence supporting their
position and ignoring everything else (Monte-Sano, 2008). They have difficulty contextualizing because they lack the background knowledge, or fall victim to presentism (Lee & Ashby, 2000; VanSledright, 2002).

Many of the differences between students and historians in the use of corroboration have already been addressed in my discussion of multiple source comprehension. Whereas students often overlook differences, historians often seek these out. Historians likely will not discount a new perspective—In fact; they often use new information that has not previously been part of historical accounts to try to convince readers to revise their own narrative of what happened. In crafting an argument they may intentionally leverage differences, contrasting a publicly shared account such as a speech with private communication such as a letter, to present a more nuanced view of what happened or argue for a particular interpretation.

The various historical thinking heuristics overlap and connect with one another. Corroboration and sourcing are related because when accounts agree or disagree, historians examine the source and consider the context to try to determine author perspective. This helps them make judgments about which version of events to trust, and to what degree. Some students do pay attention to the source of a text, and students can correctly identify obvious bias. More experienced readers tend to make inferences about reliability based on document type, though unlike historians, they mistakenly rate textbooks higher than primary sources (Wineburg, 1991; Rouet et al., 1996). Some studies found that effective sourcing is predictive of comprehension, and that a sourcing scaffold increases comprehension (Bråten, Strømsø, & Britt, 2009; Wiley et al., 2009). Contextualization is perhaps the least studied of the three heuristics; it is also a mental process that is difficult because students lack historical background as a lens for interpretation. Britt and Aglinskas (2002) discuss the way in which the inability to properly source a document makes it difficult for students to contextualize a document - if they can’t figure out when it was
written and by whom, they won’t be able to activate that schema and interpret what was happening through the lens of that time period.

The relationship between these heuristics is also essential from the perspective of historical warrants. Historical arguments are often warranted by the very types of nuanced historical thinking that students find difficult. For instance, students may be able to identify an individual document as biased. They may use that as a warrant for how that evidence supports or does not support a claim. However, whether or not a single document is biased, while certainly important to know, is not nearly enough to be able to critique or construct a strong historical argument. Students may be missing the bigger picture by basing warrants on evaluation of single documents. If students struggle to source a document, they will not only have difficulty contextualizing (Britt & Aglinskas, 2002), they will have difficulty determining whether or not that document supports a particular historical claim, because the source, the context, and its relationship to other documents is as important as what the document itself says.

**Close Reading**

While not a historical thinking heuristic, close reading was also included in this study in both the instruction and the assessment. Close reading is an essential skill within the Common Core Standards (NGACPB & CCSSO, 2010). In this study, close reading is defined as a process of carefully analyzing an author’s language for three purposes: to learn what a text says, to determine how a text works, and to evaluate its larger significance in relation to other texts and ideas (Adler, 1965; Shanahan, 2013). The second purpose of close reading—determining how a text works—functions differently in English Language Arts than it does in History. In English Language Arts, determining how a text works typically focuses on the text alone, not the source or the content. However, when historians determine how a text works, they clearly think about the source and the historical context as they look at specific words and phrases within a text. This helps them infer what an author’s potential biases or motivations may be and what it reveals
about the time period as they construct an event model of what may have happened (Martin & Wineburg, 2008; Reisman, 2012). In this study, I targeted close reading of an author’s word choice, close reading of the argument (e.g. how an author uses evidence rhetorically to support her claim), and close reading of the structure of writing (i.e. how ideas are ordered in a text).

Along with historical thinking heuristics, close reading plays a key part of historical writing. Historical writing is essentially an argument for a particular event model, or interpretation of evidence and its significance. Unlike the warrants students encounter in their daily lives, historians usually do not warrant their claims through general rules. Rather they use heuristics such as analyzing the authorship of documents, the context, and where evidence overlaps and contradicts itself (Monte-Sano, 2010). Close reading helps historians target specific words and structures that make this reasoning possible, and helps them focus on how these elements reveal an author’s biases and motivations (Wineburg and Martin, 2008).

Much of historical writing is a “public display” of how one accounts for overlapping and contradictory evidence (Monte-Sano, 2010). Though historians do not always provide explicit warrants for their claim, warranting ones argument by explicitly stating the historical reasoning used to arrive at an event model is part of the public display that lends credibility to one’s historical writing. Close reading and heuristic use both help students identify and interpret the elements to “display” as a part of their historical writing.

**Why Study Argumentation?**

This study examined students’ warranting of historical arguments; thus the theory and research regarding argumentation is relevant. There are several examples of classroom-based studies of argumentation, including studies focusing on small group discourse, whole group discourse, teacher pedagogical approaches, and written arguments. Researchers have studied student arguments after interventions, including alterations in task, provision of models, inclusion of different kinds of texts (such as those with visible authors), instruction in strategies for
reasoning, or requirements for discussion or debate. Though recent meta-analyses of learning to write (Graham & Perin, 2007) and writing to read (Graham & Hebert, 2010) have reported the types of instruction that helps improve writing for quality and reading comprehension outcomes, they do not help us understand the best way to teach students to write a historical argument from multiple documents.

Writing is often viewed as the artifact of historical thinking, but there are rarely studies that combine interventions for both reading and writing. A relatively recent body of research has developed over the past 15 years focusing on the disciplinary demands of historical writing from multiple documents. Much of the work has been descriptive or qualitative in nature (Monte-Sano, 2008; Wiley et al., 2009). Studies have teased out the unique characteristics of argumentation in the disciplines of history (Pontecorvo & Girardet, 2012; Wiley & Voss, 1999) and science (Osborne et al., 2004). Instruction in argumentation has shown to increase content knowledge in both subjects. By giving students an authentic reason for engaging with disciplinary content, argumentation makes the consideration of multiple sources of evidence more manageable and meaningful (Wolfe & Goldman, 2005).

Researchers have increasingly documented and analyzed argumentation to understand the processes by which arguments are created, often examining discourse used during this process. For instance, Pontecorvo and Giradet (1993) drew on the theories of cognitive apprenticeship and distributed cognition to explain how collaboration in group discussion helped students internalize the reasoning processes of the discipline and helped facilitate their construction of fully-formed arguments consisting of claims, warrants, and evidence.

The relationship between shared argumentation and the arguments that students individually produce is complex. If the discourse during the social argumentation process is limited, consisting primarily of claims and counter-claims without much reasoning (see Pontecorvo & Giradet, 1983), a lack of complexity in an individual’s written argument is likely.
Even when the discourse during social argumentation appears complex and dynamic, it does not necessarily transfer to an individual’s written argument. There is a clear difference in children’s ability to produce oral arguments and their ability to produce written ones.

From a discourse schema perspective, written arguments are considered closed in that there is no peer providing models for the argument schema. No one is spurring the writer’s thinking or asking him or her to elaborate by providing additional evidence, warranting claims, or qualifying claims in response to rebuttals. Individual writers must rely on their own underdeveloped closed discourse schemata (Eftmeier, 1985, as cited in Knudsen, 1992), unless they can ask themselves warrant generating questions or engage in knowledge transformation (Toulmin, 1958). There is also a matter of the difference between declarative and procedural knowledge. Even though students may be able to state what an argument is and specify its elements, they may have difficulty producing arguments themselves (Knudson, 1992).

Theorists often consider writing in general as dialogic in nature. Individuals write to an audience and make connections between the texts they write and the texts they have read (Bakhtin, 1981; Rosenblatt, 1988). Argumentation is transparently dialogic since there is an expectation that one’s reasoning is explicit, evidence from various sources is used, and an argument is crafted to convince a particular audience. Argument writing requires students to consider the nature of argumentation, by considering how their claim, evidence and warrants will be received by a particular audience.

Argument writing takes time to learn: whereas a dialogic environment increases the quality of argument, the skills don’t develop quickly and are difficult even for college students. Tasks that ask students to read and evaluate arguments as models of the thinking and reasoning of more capable others, and lessons that require a meta-knowledge of argumentation, are essential. Tasks that ask students to co-construct arguments are important as well. To co-construct an argument, students must articulate and modify their reasoning based upon feedback. Argument
writing, due to its dependence on reasoning, makes students’ thinking public, allowing readers to assess the writer’s conceptual understanding and provide the writer with feedback to scaffold higher levels of thought (Erduran et al., 2002; Vygotsky, 1978).

**Why apply Toulmin’s Argument Pattern?** Toulmin’s conceptualization of argument is significant because it represents a departure from traditional perspectives on argument and because his thinking is clear enough to be embraced by scholars from a range of disciplines. Toulmin (1958) argued that formal rules for producing the “correct” inferences and conclusions given certain premises were of little use in the real world. Real-world arguments are non-absolute. They are probabilistic or statistical in nature, and hinge on the existence of certain conditions (Verheij, 2005). Toulmin theorized a “working logic”, to describe the way in which people actually argue or “reason from premises to conclusions” in particular social contexts (Kneupper, 1978; Driver, Newton, & Osborne, 2000). While his approach centered on the arguments individuals produce, he explained that arguments could be thought of as responses to questions by (absent) others. These “questions” emphasized argument elements such as claims (i.e. what do you have to go on?) and warrants (How do you get there?) (Toulmin, 1958).

*Toulmin’s Argument Pattern* can serve as an argument schema (Larson et al., 2004) that students can use to help them identify and evaluate the essential elements of an argument as they read, in the same way they may recognize other common patterns in expository text (Goldman & Wiley, 2004). Such a schema can be applied to structure students’ own writing as well. They can ensure they have the core elements (claim, evidence, warrant) and then include other elements (backing, rebuttal, and even counterargument), to refine their writing.

Toulmin’s approach to argument is particularly appropriate for a study framed in disciplinary literacy because he asserted “we must judge each field of substantial arguments by its own relevant standards” (1958, p. 34). Warrants and backing are areas of argumentation that Toulmin described as field-dependent. Toulmin’s framework has been applied flexibility in
many different fields including history, and his scheme has been modified depending on the emphasis of the study. Alterations have included combining two codes such as evidence and warrant, not coding for an element such as backing, and adding an additional element such as counter-argument, which Toulmin does not include in his framework, but many other experts consider essential (Kuhn, 1991; Walton 1996a).

**Why emphasize warrants?** Though they are the centerpiece of Toulmin’s model (Kock, 2006), warrants are perhaps the most maligned and understudied of the central elements of Toulmin’s argument pattern (Freeman, 2005; Grootendorst & Kruiger, 1984; Hitchcock, 2006; Johnson, 1996; van Eemeren, 1884). Toulmin (1958) makes the point that warrants are dependable inferences, but not universal, absolute statements, as the complexity of real-world situation rarely allows for universals. Pinto (2006) recommends thinking about the reasoning behind Toulmin’s warrants as not “If…then” statements, but rather more conditional statements such as “Data such as D entitle one to draw conclusions, or make claims, such as C” (Pinto, 2006). This approach nicely fits the probabilistic nature of historical reasoning.

There are critiques of the inclusion of argumentative elements such as warrant in the scoring of argumentative writing, because not all experts explicitly warrant all their arguments. However, research supports that while not all expert writing contains warrants for every argument, experts are much more likely to include explicitly stated warrants in their writing. In research conducted by Crammond (1998) there were no warrants in student writing at the 6th grade level and a slight, gradual increase at the 9th and 10th grade levels. McCann (1989) found a similar pattern. Compared to data and claim, novices have far fewer warrants (Connor, 1990; Crammond, 1998; Knudson, 1992; McCann, 1989). This indicates including explicitly stated warrants is a characteristic of quality argument writing (Crammond, 1998). Though they don’t always use warrants, experts realize that providing not just claim and evidence but justification
for how the evidence proves the claim is an important part of argument writing, particularly when the claim goes beyond universally held views.

Another criticism of analyzing writing or coding for different argumentative elements is that it is difficult to tell the function of a particular statement when isolated from the larger argument. This is due to the fact that the distinction is functional rather than surface level or structural. You cannot look at a statement in isolation and be confident it is a warrant without considering the role it plays in the larger argument. However, careful analysis that considers context and function can reveal valuable information about students’ use of warrants.

There are different methods for categorizing warrants (e.g. Freeman, 2005; Pontecorvo & Giradet, 2012). One of the most traditional comes from Brockriede and Ehninger (1960), who created a typology of warrants based on Aristotelian concepts of logos, pathos and ethos, substantive, authoritative and motivational warrants (as cited in Kock, 2006). Part of the appeal of categorizing warrants is that just because the author states a rule or inference license, does not mean that the inference license is valid, and categorizing warrants can help get at quality. Other factors that are helpful in determining quality can include whether or not a warrant is accurate, whether or not it contains a specific quote or addresses more than one text, and whether or not it reflects elements of historical thinking or close reading.

**Argumentation in the disciplines.** The use of argumentation in the disciplines is challenging for students. One challenge is that argumentation in disciplines such as history and science requires multiple text comprehension as a basis for warrants, claims and backings (Larson et al., 2004; Perfetti et al., 1999). Having students write in the disciplines produces unique benefits. To write arguments, students don’t just relay the ideas of a particular text, they combine ideas across texts in new ways, using them for their own purposes and transforming them (Scardamalia & Bereiter, 1987). To illustrate, Wiley and Voss (1999) did a study where they compared the writing students produced when asked to write an argument versus the writing they
produced when asked to write a narrative. They found that writing an argument created writing with more connections and more evidence of knowledge transformation (going beyond just telling what was in the text they read). An extension of the study showed that writing a summary produced results similar to writing a narrative, while writing an explanation produced results similar to writing an argument, but students who wrote arguments were still the most likely to engage in knowledge transformation. Cognitive theory explains that tasks requiring students to create complex mental representations of the relationships across texts promote better understanding of content (Brown, Bransford, Ferrera & Campione, as cited in Wiley & Voss, 1999). Writing an argument means defending and relating pieces of evidence, requiring sustained attention (Wiley & Voss, 1999). In addition, when students write arguments they may have more personal investment.

Argumentation is important for students not only because of the mental work involved, but also because it is a central part of disciplinary discourse. Argumentation is the process by which knowledge enters a field, as scientists and historians use arguments to make knowledge claims, and use warrants and backings to support those claims (Erduran et al., 2004; Toulmin, 1958). Arguments are also central to disciplinary inquiry because they form the link between the data researchers collect and the theories they generate (Osborne, 2004). Each discipline has its own unique demands for argument writing (De La Paz & Felton, 2010).

Bain (2005) argues that a central part of historical literacy is evaluating arguments for their plausibility. The epistemology of history emphasizes that history is an interpretation of the past, based on claims and evidence. Reading a history is in essence reading an argument that was constructed by an author who carefully selected and evaluated the evidence supporting his claims. Expert studies of historians indicate they read from a questioning stance, interrogating the author and looking for the evidence to support or refute his claim (Wineburg, 1991). Expert readers also
contextualize when thinking about the claims the author is making (Rouet et al., 1996; Wineburg, 1991), and use this as another lens to evaluate their reasoning.

However, to the uninitiated, the argumentative elements of historical accounts are buried, as historians tend to relate events in a narrative or expository fashion, hiding the decisions they made about what to include and what to leave out, and leaving the argument implicit. The language of nominalizations and passive voice also tend to hide the argumentative nature of the account (Achugar & Schleppegrell, 2005; Schleppegrell & Achugar, 2003).

History as taught in middle and secondary schools does little to help students understand its interpretive nature. American history is often taught through textbooks as a progressive march toward a more fully realized democracy, with change equated to progress (Lee & Spratley, 2010). The past is viewed negatively, as a time when bad things happened because people didn’t know what we know today. In history classrooms, potential sources for interpretation are everywhere (Bain, 2005), yet many classroom texts are written as if their authors didn’t exist (Paxton, 2002), and there are few opportunities to discuss elements that are implicit or hidden (O’Brien et al., 1995; VanSledright, 2004). Thus, engaging in the kind of historical thinking undertaken by experts in the field seems an “unnatural act” to students (Wineburg, 2004).

A disciplinary literacy approach to history instruction would make the argumentative, interpretative nature of historical reading, writing, speaking, listening and thinking more apparent. By empowering students with the tools that historians use, including sourcing, corroboration, and close reading, students can begin to understand how arguments are constructed and they can use these tools to construct their own arguments.

There are several research-based approaches to fostering historical argumentation. Studies have emphasized how task demands and introduction of additional primary sources help students engage in more complex reasoning as they co-develop support for their claims and construct their own historical narratives after building causal connections across documents.
(Perfetti, Britt, and Georgi, 1995). Rouet et al. (1996) found that presenting students with contradictory accounts helped students understand the constructed nature of history and that primary sources helped students understand the importance of evidence in supporting historical accounts (see also Wiley and Voss, 1999).

**Students writing warranted historical arguments.** This study focused on writing warranted historical arguments. The little existing research on helping students write explicit warrants in general, and historical warrants in particular, is reviewed here. The writing of warranted arguments, much like historical thinking, is an “unnatural act” (Weinberg, 2004). Yet since historians warrant their arguments with historical reasoning, we need to help students write better warrants to give them access to this powerful Discourse.

Whereas there are developmental differences in argumentative writing, warrants do not develop as quickly as other argument elements such as rebuttals, which distinguish the writing of 10th graders from 6th and 8th grade students (Crammond, 1998). Mature writers more often include warrants in their writing than immature writers, though they also include more claims and evidence (McCann, 1989). Unfortunately, few interventions exist to support students’ warrant writing. For instance, there are computer-based tools that support collaboration and map out arguments for students (Scheuer, Loll, Pinkwart & McLaren, 2010) without addressing warrants or helping students to write them.

Crammond (1998) indicates several possible reasons for the lack of warrants in student persuasive writing. Warrants may often be absent in an argument structure if implicitly understood by both arguer and audience (Crammond, 1998). Infrequent use of warrants may indicate that students tend toward arguments that rest on mutually understood reasoning, selecting only those arguments that rely on common understanding or commonly held ideas (Crammond, 1998). For instance, Toulmin uses the example of the claim, “Harry is a British Subject”, followed by the evidence, “Harry was born in Bermuda”. If the audience for the argument were
British and knew that Bermuda was a British territory, the argument could stop there. It would not be necessary to provide a warrant to explain how the evidence that Harry was born in Bermuda supported the claim. The warrant “People born in Bermuda are generally British subjects” does not need to be explicitly stated. However if the audience were people unfamiliar with the territories of Britain, it would be important to make that explicit (Verheij, 2005).

Since a warrant is a shared understanding that becomes the basis for reasoning, writers can only develop appropriate warrants if they have the understanding of their audience and idea of what their audience will accept. Less mature writers are less likely to realize when a warrant is needed, when the principles involved are unclear or require additional explanation (Lunsford, 2002). Students make many unsupported statements when asked to form an opinion, even though their notes show evidence they have read texts carefully and gathered appropriate information (Crammond, 1998, p. 341). They fail to view the writing of warrants as a rhetorical activity that adds to the strength of an argument by establishing “mutually agreed upon premises, beliefs, and feelings (by providing) a shared context with the audience”. The shared context allows the writer to “gain rapport and render the audience more receptive to the claims and arguments proposed” (Berthoff & Stephens, 1988, as cited in Crammond, 1998). Helping students to understand warrants as a rhetorical tool that helps to build a shared understanding with their audience may help demystify the idea of warrant and give students a concrete reason for their use.

Warrants as the ‘home’ of disciplinary thinking. Thanks to research, there is a clear picture of how experts use heuristics as they read documents. However, we have less information about how students can use these mental tools in their own writing. More research needs to be done to understand how educators can teach students to apply heuristics as strategic behaviors to make meaning from complex disciplinary text and construct their own arguments. Discussions about heuristic often focus on applying a heuristic as a strategy for reading or writing about texts, not as a tacit pattern of thinking as an expert would use a heuristic. Part of the problem is that we
teachers tend to talk about heuristic use to students in an amorphous way. We say, “This is what experts do, now do this in your own reading and reasoning”. Most interventions have focused around reading disciplinary text; fewer have emphasized heuristic use and historical reasoning in student writing.

In the structure of an argument, a warrant is a general rule about how the data supports the claim. It can be thought of as an inference license- a rule that supports the inferences the individual is making. When a historian makes sense of documents as he reads and reasons, he uses heuristics to make inferences about these documents. In that sense, the heuristics themselves are serving as inference licenses. As he constructs different arguments, he is using these established patterns of reasoning to build a reliable model of the relationship between documents and ideas. Because this pattern or shortcut of thinking has held up over time, he can be more confident in his conclusions. Much like Toulmin’s explanation of warrants, conclusions drawn from the use of heuristics are not accurate 100% of the time. They are probabilistic. When experts write down historical arguments, they may share inference licenses that they use to connect their claim and data. For instance, in some arguments, they may explain how they read two different documents written by different authors, and argue which perspective is most valid. In other instances they may not include the explicit reasoning behind how they got from evidence to claim, because they may know that no warrant is needed based on their understanding of the discourse community of history.

Helping students write arguments warranted by historical heuristics is essential; it helps them understand the norms for establishing common understandings in the discipline of history; yet, most students do not recognize that this is something they should be doing. The lack of awareness of the need for warranting stems from a lack of knowledge about the norms of the discipline (Crammond, 1998). If students do not understand what knowledge is common and what sort of logical reasoning is commonly used in the discipline, they will have a hard time
understanding when and why to use warrants. Since they are not experts, making the reasoning behind their arguments explicit is all the more important because their arguments are less likely to be given the benefit of the doubt (Crammond, 1998).

When students construct their own historical arguments, it is fruitful to think of the warrant as the place in their writing where that historical thinking ‘lives’. Student historical argumentative writing rarely contains warrants or elaboration. It often reads as a series of claims and evidence, particularly in the lower grades. Whether this is due to a preference for simplistic arguments, a lack of knowledge about the need for warrant, or confusion about exactly how the evidence and claim are linked (Crammond, 1998); the fact of the matter is that student writing would be improved by the use of warrants. This is also important from a disciplinary literacy perspective, where the goal is to have students gain membership in the discourse community by constructing and evaluating knowledge.

The way experts get from evidence to a claim about that evidence is in large part through heuristic use (Wineburg, 1991; Wineburg & Rosenzweig, 2008). Heuristics serve as a warrant, an inference license to bridge the evidence and the claim. Since students are using heuristics not as experts do but as a strategic behavior, it makes sense to explicitly state warrants in their writing. Once students develop the depth of disciplinary understanding to know when a warrant is or isn’t needed, one can remove the scaffold of providing an explicit warrant for each claim in their historical argument.

One of the reasons why students may have difficulty including historical reasoning in their writing is because educators don’t provide them with a structure for their writing that includes a place for that reasoning. When they are provided with one at all, structures typically resemble the five paragraph essay, or emphasize claim, evidence and elaboration. If we take a structure for argument writing such as Toulmin’s Argument Pattern, where the typical structure is claim, evidence, and warrant, and support students by telling them that the warrant is where you
explain how you used your heuristic/strategy to arrive at your claim, students should write better warrants, meaning warrants grounded in historical reasoning (namely sourcing and corroboration, and close reading in service of historical thinking).

Monte-Sano (2008) identified five characteristics of effective writing instructions in history classrooms. These included treating history as “evidence-based interpretation”, reading historical texts as interpretations, supporting both reading comprehension and historical thinking, having students develop their own historical interpretations supported with evidence, and using gradual release, from direct instruction, to guided practice, to independent practice, with feedback to teach evidence-based writing (Monte-Sano, 2008).

I used a similar approach with the addition of two supports for writing warrants that reflect historical thinking heuristics and close reading. Using the discussion of heuristics from the Stanford History Education group (www.sheg.org), I modified generic warrant generating questions to cue sourcing, corroboration and close reading. I also provided students with templates (Graff, 2006) that reflected the use of historical thinking heuristics. I modeled both the use of heuristics when reading primary and secondary documents, and the use of heuristics as warrants in historical arguments. This was intended to stimulate students to engage with the same mental tools experts use tacitly when they read and reason about disciplinary text (Wineburg, 1991) and scaffold students toward knowledge transformation (Bereiter and Scardamalia, 1987) by engaging in audience-centered activity. Following are examples of warrant generating questions and sentence frames used in this study.

*Examples of warrant generating questions.*

1. General warrant generating question: “How did you get there?”(i.e. from your evidence to your claim)?

2. Example of warrant generating questions that supports sourcing
“How do the strengths or limitations of this source type influence how much I will rely on it to support my claim?”

3. Example of warrant generating questions that supports close reading includes:

“What language does the author use to convince the reader? How do the words/symbols/images the author uses show their purpose or perspective?”

“What does the author choose to say first/last? What does this show about his purpose or perspective?”

4. Example of warrant generating question that support corroboration

“Is it surprising that these documents disagree? If the documents that disagree, which do I trust most? Which is most reliable?”

Examples of Sentence Frames

1. Example of sentence frame for close reading

The use of ________ (image/symbol) reveals the author is trying to ________ (connect to how this supports claim).

2. Example of sentences frames for corroboration

While these two documents disagree, ______ is the most reliable because of ____ (explain how that document is most reliable) (explain how that supports your claim).

While ______ (name of source) says _____, (name of other source) contradicts it by saying ________, which indicates ____ (explain how this supports your claim).

Summary

This chapter provided a review of research that informed this study. It was a study of advanced literacy skills specific to the discipline of history, namely historical argumentative writing. The study leveraged insights from many fields, including the relationship between reading and writing and understandings about the thinking heuristics historians use as they read and reason across multiple texts.
The design of the intervention and the instrumentation was also developed in light of the body of research on argument writing. It targeted the warrant, an aspect of writing that is often underdeveloped or totally absent in novice writing. It explicitly taught students how to use historical thinking heuristics and close reading to warrant their arguments. The intervention included discipline-specific warrant generating questions and sentence frames as additional supports, and gradually released responsibility to the students. In the next chapter I will examine the impact of this intervention on students’ ability to select effective warrants, rank warrants in order of effectiveness, and write effective warrants. I will also use qualitative data to examine students’ experience of the instruction, and their process for selecting, ranking and writing warrants.
III. METHODS

Participants and Setting

The participants in this study were students in four eleventh grade U.S. History classrooms taught by one teacher in a public high school, Northwest High (pseudonym). According to the district website, The U.S. History class is “a chronological study of the major political, economic, social, and cultural events from the creation of the Constitution to the present. It includes preparation for state and U.S. Constitution exams, which are a requirement for graduation”. There were a total of 89 students in the study, though not all students provided data for all measures.

At the junior level at Northwest High, all students must take a history course of some sort, but students have different options based on their academic performance. Though information came from the district website, I will avoid citations for school-specific reports which could jeopardize participant confidentiality. All students are eligible to take U.S. History and U.S. History-American Studies (which combines U.S. History and Literature). This is known as the “College Prep” track, which is essentially the general education track. Higher performing students are eligible to take U.S. History-Honors and American Study Honors. These students are also eligible to take Sociology honors and Economics honors, courses traditionally offered to seniors. This is known as the “Honors” track. The highest performing students may take U.S. History AP or Psychology AP. Psychology is also offered as an elective to all students.

Northwest High is located in a working-class suburb immediately adjacent to a major Midwestern city. It is a large, diverse school with a total student population of 1,733. 47.4% of the students at Northwest High are Hispanic, and 45.2% of the students are White. Asian
students make up 3.4% of the student population. 1.6% of the population is Black, and 1.3% is identified as two or more races. American Indian students make up 1% of the population, while less than 0.1% of the population is Native Hawaiian/Pacific Islander. 11.5% of students have an Individualized Education Program. 7.4% of students are identified as low income, 6.6% of students are categorized as Limited-English-Proficient, and 0.6% of students are considered homeless.

The attendance rate at Northwest High is 92%, and the high school dropout rate is 1.3%. The average class size is 21.2 students. Eighty percent of teachers in the school district, which is comprised of two high schools, have a Master’s degree or above. The faculty numbers are not reported separately for this school.

In terms of performance, the average ACT Composite score for the graduating class of 2014 is 19.5, which is in line with the state average. The average ACT Composite score in Reading is also 19.5; it is 18.8 in English. Overall 40.2% of students at Northwest High are considered “Ready for College Coursework” based on their ACT Composite scores. 31.7% of students met ACT Benchmarks in Reading, while 57.1% met ACT Benchmarks in English. Postsecondary enrollment 12 months after high school graduation was 64%; it grew to 70% 16 months after graduation.

While overall performance at Northwest High is on par with the state average, the average ACT-PLAN score of students in this study is 15.54, reflecting the fact that this is a general education track. The higher performing students are enrolled in Honors and AP classes.

Eleventh grade students were chosen as the focus of this study for several reasons. First of all, 11th grade is a high stakes testing year, with students in this state typically taking both a required statewide exam and the ACT. This year, students will be taking the online PARCC assessment for the first time (PARCC Writing Evidence Tables, 2015). These students could benefit from additional support in strategies for reading across multiple texts and writing
arguments, skills emphasized on the PARCC exam. They also could benefit from the multiple-select items and ranking items on the warrant selection and warrant ranking of the primary assessments in this study, as the PARCC exam includes similar innovative item types. Eleventh grade coursework periodically incorporates document-based questions as students use these to prepare for the types of writing tasks in AP or college classrooms, so the focus on writing about multiple historical texts fits well with the 11th grade curriculum.

Though this particular school was a sample of convenience thanks to a colleague’s existing relationship, I intentionally chose a school serving a diverse population, and a general education classroom for two reasons. First, I have a social justice perspective and want to design strategies that actually work in the contexts where they are most needed. Second, from a discourse perspective, students from nonmainstream backgrounds may not have access to powerful Discourses such as academic argumentative reading and writing. The ability to teach them to read, critique and produce their own historical arguments could be powerful skills to give nonmainstream students more agency in their own lives (Birr-Moje, 2008; Lee & Spratley, 2010).

There were a significant number of diverse students and English Language Learners in the study population. The classroom teacher Ms. Jones (pseudonym) had a total of 15 students who were categorized as special education mainstreamed on her roster across the four class periods involved in the study. She had a wide range of students for whom English was not their first language, including some students who had arrived in the United States during their high school years. U.S. History is the first class that is not sheltered for English Language learners. This includes Hispanic and Eastern European students whose home languages include Polish and Bulgarian. According to Ms. Jones, students entering Northwest High from Eastern European backgrounds typically have a stronger content background about U.S. History than students entering from Hispanic countries, due to differences in schooling in their home countries, but both groups typically lack knowledge of the explosion of the U.S.S. Maine.
There were two adults involved in the study; the U.S. History teacher of these four sections, Ms. Jones (pseudonym) and myself. Ms. Jones was classified as a regular classroom teacher. She is a veteran teacher who has received professional development on Document Based Questions (DBQs) in the past, which emphasize writing a historical argument from multiple sources. However, she said her school “uses the term loosely”. Her instruction typically focuses around single texts or paired texts. She does not use Toulmin’s argument pattern or emphasize the role of warrant in her instruction, but she has taught about the parts of an argument. The classroom contained a poster with the parts of an argument as “claim, evidence and reasoning”, with an example of an everyday argument about the best fast food. Reasoning was defined as explaining how the evidence supports the claim, so it was used similarly to how I used the term “warrant”, but without the disciplinary emphasis that was at the focus of this study.

I was also qualified to teach history. In college, I minored in teaching history, and had previously taught US and World history at the high school and Jr. High levels for a total of three years. The remaining eight years of my teaching experience was as a Reading and English teacher in urban and suburban settings. For the past five years I have been out of the classroom but I have an active certification for teaching U.S. History at the high school level. I was a literacy coach in a junior high school in this large Midwestern city and an instructional coach in “turnaround” high schools which included some of the lowest performing high schools in the city. These roles included coaching history teachers as well. I also served as a literacy coordinator in the same “turnaround” network of elementary and high schools. My work focused on helping teachers and leaders in our network make the transition to the Common Core State Standards (NGACBP-CCSSO, 2010). The remainder of my teaching career I was a Reading and English teacher at the middle school and high school levels.

Ms. Jones became interested in disciplinary literacy because a colleague was part of a research study, which focused on reading, evidence and argumentation in disciplinary instruction.
Ms. Jones was not part of the study, but felt that her students needed advanced literacy skills as well. She had tried a few materials from the Stanford History Education Group “Reading Like a Historian” website (www.sheg.stanford.edu/rlh). However she felt these activities were challenging for her students and that they needed a great deal of scaffolding.

The History department at Northwest High agreed as a whole that students needed to have a claim and provide evidence and reasoning in their argumentative writing. However, according to Ms. Jones, “No one could agree what that meant, nor how to grade it, or how to or present it to kids”. She did try to introduce the term warrant last year, but her students associated it with getting arrested which was distracting, so she did not continue with the term. This year, she works periodically with a literacy instructor on the district literacy team. They work together twice a month, with the focus on lesson planning.

Ms. Jones taught history chronologically. Prior to the study, students began the year by studying slavery and then transitioned to the 1850s. The study took place at the beginning of a unit on imperialism. Ms. Jones said that recently the “curriculum changed drastically.” She said that culminating assessments, which used to consist of multiple choice questions, now were exclusively comprised of argument writing tasks. However, according to Ms. Jones, there was no professional development for teachers regarding how to teach students differently to prepare them for these assessments. She wanted help in preparing kids to do the kind of thinking the new argument writing assessments required. She felt that the abrupt switch could amount to “pulling the rug out from these kids.” She hoped to use what she learned as a participant in this study with her teaching team.

Both Ms. Jones and I were participants in this study. In order to determine whether or not the idea works at the local level before using it to generate theory to inform instruction in similar contexts, I developed a comparison study with random assignment of subjects stratified by classroom with teachers counterbalanced across the treatment groups to limit the possibility for
differences in quality of treatment or Hawthorne effects. In addition, the instruction was scripted to minimize teacher effects, including the key questions and prompts we used. We met each day prior to instruction to run through the lesson and to debrief the effectiveness of the intervention and what we were learning about student use of warrants, including potential changes to the protocol, as is common in design-based research.

**Materials**

The primary instructional materials for this study consisted of a background video introducing the controversy surrounding the explosion of the U.S.S. Maine and four primary sources related to the Maine explosion (Rosenzweig & Wineburg, 2008), the handouts and overheads for the treatment and comparison treatment, and the scripted lessons. The tools that were provided to students included the handouts explaining different elements of historical thinking containing warrant generating questions and sentence frames for the treatment condition and the two-column note taking forms and scaffolds for gathering evidence for the comparison treatment group. There were also overhead transparencies of these documents used for modeling purposes.

The video was downloaded from the website “Historical Thinking Matters” (Rosenzweig & Wineburg, 2008). It was approximately three minutes long and provided historical background on the explosion of the U.S.S. Maine. This video was shown to orient students to the topic and to build knowledge for the argument writing task, to assist students in comprehension of the documents, and to help mitigate the impact of varying levels of knowledge on students’ written arguments, along with the series of documents about the U.S.S. Maine.

**Texts for instrumentation.** All eight primary sources for this study were adapted from the collection on “Historical Thinking Matters” (Rosenzweig & Wineburg, 2008), an award-winning website that contains both meaningful collections of documents around a particular
historical question, and video of historians doing think-alouds with some of the documents in the set. The work is related to Wineburg’s foundational work with expert historians reading historical documents (Wineburg, 1991). These documents were selected to support students engaging in the use of historical heuristics, namely sourcing, contextualization, corroboration and close reading. The texts themselves are included in Appendix A.

Wineburg and Martin (2008) have studied students and historians engaging with these text sets. They have conducted think alouds (Ericsson & Simon, 1993) with historians and novice high school age readers, and have had students write essays in response to a similar prompt, so they have been vetted for opportunities for students to engage in the heuristics described in this study. Though the site makes it possible for students to view the historians’ reasoning as they read, and includes guiding questions to support student comprehension of the documents, these elements were not part of this study, and students did not visit the website.

The texts included a range of types. They were all print-based; no video, audio, or political cartoons were included. Though texts of different modes are essential for studying history, I wanted to focus on warranting documents using textual evidence, so I stuck to written texts. The text set included eight primary sources and one tertiary source (a textbook). The primary sources included a range of text types: two newspaper articles, a song, two speeches, and two telegrams, one containing an unsigned enclosure. The range of text types was important because historians source not only based on who wrote a text, but also the type of text and its strengths and limitations to answer different historical questions (for instance, a private letter may reveal more of an individual’s own thoughts than a public speech). The textbook was included because previous studies have shown that despite being tertiary sources, students often identify textbooks as more trustworthy and useful than other sources (Rouet et al., 1996) and as sources of accurate, reliable information (Wineburg, 1991), though historians approach them from a
skeptical, questioning stance. Including a textbook also may help students reflect on the range of sources they typically encounter in class.

Though the sources were already vetted as part of the Historical Thinking Matters website, I took additional steps to ensure the documents were appropriate for this study. This included determining the Lexile level of each text, determining opportunities to corroborate texts, and informal pilot testing of the documents. The source and Lexile level of each of the eight sources in the text set used for the warrant selection, ranking and writing tasks was determined on Lexile.com (https://lexile.com/analyzer/). Six of these sources fall within the ranges recommended for grades 9-10 (1050L-1335L) or 11-12 (1185L-1385L). One document, a speech, fell just below the 9th grade Lexile range (940L). This may have been due to the desire to appeal to a wide audience as it was given by a senator seeking reelection. One source was above the 11-12 grade band (1660L) (NGACPB & CCSSO, 2010). This document was a telegram and the higher Lexile may have been due to long, multi-clause sentences describing a series of images.

Documents were modified slightly from their original state in order to be accessible to all readers (Wineburg & Martin, 2009). Since these were historical, some of the complexity (or lack thereof) related to the speaker and nature of discourse and could not be modified without losing the character of the documents themselves. Such variation is part of the uniqueness of examining the historical record, and helping students read and interpret these differences is an important part of the work of this study. The sources from “Historical Thinking Matters” (Rosenzweig & Wineburg, 2008) included a brief explanation of the document, providing context and a definition of some words. This format was mirrored for documents students examined during instruction.

Sources were also categorized by type (e.g. primary, secondary, tertiary), and whether or not the source was unique (containing information not contradicted or corroborated elsewhere in the document set), contradicted or corroborated. This information was intended to ensure the text
set provided a range of opportunities for historical thinking (Manderino, 2011). The text characteristics can be found in Appendix B.

**Texts for instruction.** The texts used for instruction were different than those used in the instrumentation, but of similar type and content. The first source was a contemporary U.S. History textbook (Prentice Hall, 2013). This source was read aloud to all students first in order to build knowledge, and was also analyzed by students on the first day of the study. The second source was an article from the New York World that was retrieved online from the website “History Matters” (www.historymatters.org). The third source was an article from the New York Times, also retrieved online from the “History Matters” website. The final document was an excerpt from the Monroe Doctrine, which was also adapted from the “Historical Thinking Matters” website. The sources provided opportunities to model the use of historical thinking heuristics and writing warrants (for the treatment condition) and the selection and evaluation of historical evidence (for the comparison treatment) that was relevant to the work students would be doing for the instrumentation. The texts for instruction are found in Appendix C.

**Models of writing.** Models of writing have been shown to be effective in writing instruction (Graham and Perin, 2007). In particular, the treatment condition included models of how the heuristics of sourcing and corroboration along with close reading can be used to warrant historical arguments. I wrote these models, which were used with a think aloud to demonstrate the process of warranting a claim. They were written at a mean Lexile level of 980 to lessen the chances comprehension interfered with students’ understanding of the central elements of the essay, yet still present a model that was complex enough to provide an exemplar for students’ own writing.

**Instrumentation for main study: Types of claims.** One of the characteristics of warrants that make them difficult to study is their dependence upon their claim and evidence partners. The type and quality of claim influences the warrants being employed, along with the
relevance of evidence that students’ gather. I theorized that traditional approaches to helping students write warrants by coming up with general rules that link claim and evidence (Hillocks, 2011) are less effective in history classes because rather than rely upon general rules, historians often employ specific heuristics to handle the unique demands of reasoning in the discipline. While the simplicity of Hillocks’ take on warrants is appealing, a straightforward “as a rule” statement is often not sufficient to capture the relationship between multiple pieces of evidence and a historical claim. Rather, heuristics (e.g. sourcing and corroboration) and close reading, help historians address unreliable or uncertain narrators, notions of bias that shift based on the question asked, the incomplete record of events, and multiple, often contradictory sources. Students need to understand that heuristics themselves often ‘live’ in the warrant of an argument—they are serving the specific purpose of linking claim and evidence.

Due to the specialized and nuanced ways historians use these heuristics (e.g. Wineburg, 2004), students need examples of particular types of claims, and how warrants function to support them. This research pinpointed specific of claims in order to be precise about what students can and cannot do and what impact may or may not be seen on different sorts of warrants. The warrant selection instrument itself had a 4x4 design with four different types of claims, crossed with four types of warrants. I will provide the rationale for this design and briefly describe the types of claims and warrants this instrument addressed.

According to Britt, Kurby, Dandotkar, and Wolfe (2008), there are three elements of a claim. These include the theme, the side, and the specific predicate. The theme is basically the subject in question. The side is literally the side taken with the claim, whether it is for or against, etc. The predicate is what the claim is saying should or should not be done. Consider the statement “The explosion of the U.S.S. Maine did cause the U.S. to invade Cuba.” The theme is the explosion of the U.S.S. Maine. The side is “did cause”. The predicate is “the U.S. to invade Cuba.” According to Britt et al. (2008), while it is generally easy for people to recall the theme
and the side of a claim, people are significantly more likely to incorrectly identify an argument predicate (e.g. should be banned vs. should be limited) despite the clear difference in the sort of evidence and reasoning used (Britt et al., 2008). Even college students have difficulty in this regard (Wolfe et al., 2009). This indicates more must be done to help students attend to such differences and employ precise reasoning for their particular claim.

College students also have difficulty arguing for a side they do not agree with, so I intentionally designed an instrument that had students selecting and ranking warrants for both sides of an argument. As part of the data generating tasks, students were asked what side they agreed with to determine potential impact of “Myside bias” on student performance (Wolfe, Britt, & Butler, 2009).

**Pretest.** The students completed a pretest to determine what students already knew about argument writing and how effective they feel they are at it (See Appendix D). The pretest questionnaire had two parts. Part one asked them about their experience with argument writing, both in history and other classes. It also asked them to name the parts of an effective argument. Finally, students self-evaluated their ability in their argument writing ability on a Likert scale. Part two asked students to answer an open-ended question directing them to write down everything they knew about the U.S.S. Maine and the Spanish-American War.

The teacher also completed a questionnaire to help me determine both her own knowledge of the role of warrants in historical argumentation and to confirm whether or not the students had prior instruction in warranting arguments (Appendix E).

Finally, the teacher provided a small representative sample of argumentative writing at the high, medium and low levels that students produced for their most recent DBQ essay, along with a brief description of the classroom instruction that accompanied that writing sample. These materials were not analyzed in depth as part of this study, but provided additional context for the data gathered in terms of what was happening already with argument writing prior to the study.
Warrant selection task. The types of warrants emphasized within the instruments and within instruction were intentionally chosen as well. They included warrants that leverage sourcing, corroboration and close reading, along with non-historical warrants. Non-historical warrants are not necessarily invalid in historical arguments, and can be employed effectively. In secondary sources, including those written by historians, authors of historical arguments occasionally make everyday appeals, to logic, intellect, or even common sense, to reel in or engage the audience. However, they interweave these with claims warranted by historical heuristics. Using everyday appeals is only a problem when not combined with a preponderance of historical evidence and when not supported by conclusions made through the use of historical heuristics.

Though contextualization was a key heuristic identified by Wineburg, it was not included in this study. It relies heavily on knowledge of the historical time period to draw more meaning from the document. Students using contextualization use what they already know about what it was like to be alive at that time as a lens for interpretation, drawing on their knowledge of the issues people faced and how they probably thought about the world, and comparing how that time was similar or different to the current time to put the language in the proper sociohistorical context (Reisman, 2012).

Students in this study are just beginning to learn about this era in history, and it would be difficult to build deep background knowledge in a short study, so I will not be addressing contextualization. However, there is some consideration of when a document was written within the sourcing heuristic. This study teaches students to begin their reading of a document by sourcing, or determining when the document was written and by whom, and then use that to identify author’s position and purpose, and reason about their potential believability and trustworthiness. This work is done in relation to other documents in the text set. It asks when this document was written, but does not focus on what else was happening at that time, or use in-
depth knowledge of the historical context as a lens or interpretation. This is consistent with how the Stanford History Education Group (www.sheg.stanford.edu/rlh) treats sourcing.

Each warrant selection task item included five answer choices. Each set of answer choices included responses that represented warrants including sourcing, corroboration, close reading, and non-historical. Across the 12 items, there were three correct and three incorrect responses each for sourcing, corroboration and close reading, with non-historical filling in the additional items. Each item included more than one “correct” answer. The amount of “correct” answers varied from 2-3 items per every five-item prompt. Students were not cued about how many answers could be correct.

Warrant subtypes. Within each type of warrant reflected on the warrant selection task there were subtypes that referred to particular varieties of these warrants. These subtypes of warrants were introduced during instruction. For instance, when each of these types of thinking was introduced, examples were given of each of the subtypes, and the students received a handout that had examples listed for each of the subtypes. However, students did not get extended practice with each of the subtypes based on the short duration of the study. Therefore, while these warrants and subtypes were also included in analysis and reporting, they are more important in ensuring that the breadth of the type of thinking reflected by that heuristic is covered. Caution should be taken in making assumptions based on performance on a particular subtype of warrant due to the short duration of a study and the small number of total distractors linked to any one warrant subtype. See Appendix F for an overview of warrant subtypes and foils.

There were four subtypes for sourcing. The first, Sourcing A (SoA) referred to examining the source in terms of when the event happened, the placement of the event in relation to other events, and how that impacted the interpretation. An example follows: “This song, written after the sinking of the U.S.S. Maine but prior to the war, reflects the passionate public opinion to avenge the U.S.S. Maine following the explosion. This outrage influenced the
decision to invade Cuba.” This was not considered contextualization because it was focused on the information that could be found within the source itself- when an event occurred- and not information about the larger social context or comparing what was happening at that time to how things are now as a lens for interpretation.

The second sourcing subtype, Sourcing B (SoB), refers to strengths and limitations of that particular type of source (for instance, a politician’s diary versus a politician’s campaign speech). An example follows: “The fact that this is a war message to Congress means that McKinley is making a case that he hopes will persuade his audience. He’s only going to give evidence of why we should declare war that will be widely accepted. Therefore just because he states the U.S.S. Maine is an "obvious" reason to invade, it doesn't mean this was necessarily the root cause.” Though it is focused on the type of source, students may also experience these prompts as prompting them to reflect on the advantages of that particular source (and the fact that it is a speech, for instance).

The third type, Sourcing C (SoC), deals with author’s perspective or role as a tool for interpretation of the document. An example follows: “The Secretary of the Navy has expertise about ships, so the fact that he says this makes it likely that others believed him and blamed Spain for the sinking of the U.S.S. Maine, which helped lead to the invasion.”

The fourth (SoD) is sourcing used in combination with corroboration. An example follows: “This sort of Yellow Journalism described in the textbook (Doc. J), and exemplified in the New York Journal (Doc A), uses emotion to and exaggeration to engage the readers. While it may not be the most factually accurate, it shows how passionate people felt, and had a big influence on the public opinion to go to war”.

There are three subtypes for corroboration. Corroboration A (CoA) is corroboration in the sense of “vote counting”, or referring to other documents or knowledge from the document set that also said the same thing. An example follows: “The song (Doc C) and the New York Journal
(Doc A), conveyed the attitude that Spain was to clearly to blame for the explosion, and they had to respond. Though McKinley's speech does not directly state Spain blew up the U.S.S. Maine, he blames them for not protecting it in the harbor. All these add up to show how the U.S.S. Maine helped cause the invasion”.

The second, Corroboration B (CoB), demonstrates using corroboration to explicitly address contradictory information. Corroboration is typically thought of as reading across multiple historical accounts to compare content and identify differences, giving more weight to that information which is common across accounts. However, another historical thinking skill related to corroboration is also examining where documents disagree, and using that to reason about the evidence.

An example follows: “President McKinley says Spain can't secure the navy in the harbor, and echoes the upset tone of doc A. Of course they can't secure against an explosion of coal, which document B and the textbook (Doc J) said was a likely cause. Therefore, they should not have invaded Cuba following the explosion of the U.S.S. Maine”.

The third subtype of corroboration, Corroboration C (CoC) addresses corroboration in combination with sourcing, and how the source of the corroborating information contributes to interpretation. An example follows: “This is a campaign speech so he is pulling out the stops to show the advantages of his plan of expansion into Cuba. This connects to Doc D, in that they saw the region as an extension of their influence and power. Doc H refers to the U.S. deciding what to do with Cuba when they actually got the lands. These sources show there were economic and political reasons to invade beyond the explosion of the Maine”.

There are also three subtypes for close reading-historical. The first type is Close Reading A (ClA), which is close reading of word choice, zooming on a particular word or phrase and its significance. An example follows: McKinley's use of words like “intolerable” and "obvious sign" shows he is trying to remove doubt from what they are doing, and make it seem like it would be
silly not to intervene in Cuba. This shows how the explosion of the U.S.S. Maine was used to help the case for invasion.”

The second subtype of close reading is Close Reading B (ClB), or close reading of argument, which deals with the way the argument was constructed in terms of the claim the author made, evidence provided, et cetera. An example follows: “The President makes the claim that things in Cuba are intolerable and gives the destruction of the U.S.S. Maine as evidence. He warrants this by saying the Spanish government should be able to keep the American Navy safe. But if it was a coal explosion (e.g. Doc B), there is nothing that the Spanish government could have done to protect the U.S. Maine. The way he warrants the argument is faulty, so this shouldn’t be a reason for invasion.”

The third subtype of close reading- Close Reading C (ClC) is the close reading of word order/structure, which emphasizes the way words or ideas are ordered in a document. An example follows: “This document is structured with the factual information first, then includes a stirring first-hand account with example after example of how horrible things are in the Reconcentration camps. By the end of that passage, it is clear the only remaining step is to intervene.”

In addition to the types and subtypes of historical thinking heuristics reflected in the study, there are two subtypes for non-historical warrants. Since there are so many potential ways to warrant an argument, the categories were intentionally kept broad: everyday reasoning that is relevant to the historical event (NHA) and everyday reasoning that is irrelevant (NHB). An example of NHA follows: “People don't want war ships for nothing. If he asking for warships, that means he wants to go to war, or at least scare someone. And once you get ships down there, you feel obligated to use them. So the U.S.S. Maine wasn't the only cause; they were already anticipating other conflicts.” An example of NHB follows: “If someone blows your stuff up, you have to do something about it. You can't allow that to go unchecked, or people will get the idea
you are soft. Therefore, they should have invaded Cuba.” While students were not instructed in how to do non-historical reasoning, they were informed that non-historical reasoning could also be valid, and not to automatically discount it.

**Foils.** Finally, there are codes for foils which cut across the warrant types in question. There are five types of foils. The first type of foil includes statements that are inaccurate (i), which means a connection that is factually wrong. The second subtype is warrants that are accurate but irrelevant (ii). These connect claim and data in a way that is factually correct, but not relevant to the question. The third subtype is relevant but out of proportion (iii), which is using a relevant warrant but ascribing too much or too little meaning to it by over or undergeneralizing. The fourth subtype is misused heuristic (iv), which includes common errors in heuristic use, such as assuming primary sources are necessarily more reliable because the person was there. The final foil is an error that conflicts with another heuristic (v). This means that the warrant makes a claim based on one heuristic that is contradicted by another. For example, attempting to corroborate by finding the opinions of three people in the U.S. government that agree about what happened, without realizing they all have the same perspective, and seeking no alternate perspectives. In this instance the subtype of Corroboration A (CoA or corroboration by vote counting), conflicts with the heuristic of Sourcing C (SoC), sourcing by thinking about the perspective of that particular individual and their motivation.

**Warrant ranking task.** After completing the warrant selection task, students ranked the warrants they selected as effective from most effective to least effective. They typed the letters of the warrants they selected and the ranking they assigned them. The idea was to get a better picture of student ability to discriminate between more effective and less effective warrants. There were a total of 12 warrant ranking task items, one following each warrant selection task. A sample item and link to the entire warrant selection and warrant ranking task is available in Appendix G.
**Warrant writing task.** On the day after completing the warrant selection and warrant ranking tasks, students completed a warrant writing task in which they were given a claim and evidence and were directed to write a warrant that explained how the evidence supported the claim. They were not prompted to use historical thinking heuristics, though the items were designed to give students the opportunity to use historical thinking heuristics through the nature of the claims and evidence provided. The items referenced texts for the same data set. They had 15 minutes to complete this task. They also were given 30 minutes to respond to an extended writing prompt about the explosion of the U.S.S. Maine. The text of the warrant writing task can be found in Appendix H.

**Overview of Validity and Fidelity**

To ensure the validity of the conclusions, both the treatment and the comparison treatment groups examined the same documents in the same order; however, the focus for the treatment and comparison treatment differed. The treatment group emphasized using historical thinking heuristics as warrants, while the comparison treatment group focused on gathering and evaluating evidence to support their claim. Instruction was scripted for each group, and both instructors taught both conditions. An important part of design-based research is the development and testing of prototypical instructional approaches that are adjusted as a result of their application in the classroom. We huddled at the beginning and middle of each day to discuss how the lesson was going and discuss any issues that arose during instruction and potential changes to the protocol for future research. However, since we were both teaching each condition, we were sure to remain consistent in the way we were teaching each group.

**Assignment to condition.** Within each class, students were randomly assigned to a condition. The researcher and teacher were initially randomly assigned to condition and then the assignments were balanced so that they were teaching each condition for each class period an equal number of times on alternating days. One condition met and received instruction in the
library; the other group remained in the classroom. This alternated each day to control for any differences in performance related to the location of the treatment.

**Pilot testing.** As part of the prototyping phase of the design of my research, I arranged an informal pilot test of the documents with a small group of similarly diverse high school students in the same region to ensure that these documents were still accessible and appropriate for high school juniors. The teacher who led the pilot said that the students were able to comprehend the documents and found them appropriately challenging. She also led a group of five students through a pilot test and a think aloud of a subset of items on the assessment. Based on her feedback, the assessment directions were made clearer and the number of items was reduced to address issues of practicality (Plomp & Nieveen, 2003).

**Internal validity.** In addition to the pilot testing, expert judgment has an important role in design-based research. Three history experts (a professor of Literacy Education who is a former AP U.S. History teacher, a PhD student in U.S. History Education with a focus on disciplinary literacy, and the cooperating teacher of this study) examined the text set. They all felt the documents were appropriately complex for diverse 11th grade students and felt the documents provided a range of perspectives necessary for a nuanced understanding of the topic. In addition, they previewed the test items and found them challenging but reasonable for high school juniors. The PhD student in U.S. History also took the warrant selection assessment herself and provided feedback. In addition, the professor of literacy education and I went through the warrant selection and warrant writing items item by item to make sure that there was construct validity; meaning that the items sufficiently reflected the types of historical thinking that they were intended. This included taking the assessment ourselves and coming to agreement on needed changes. These expert judgments helped ensure the assessments were practical and could be used by actual educators, which reflects the commitment to construct validity and practicality in this stage of design-based research (Plomp & Nieveen, 2013).
Overview of Procedures

The following section will explain the procedures undertaken in this study. I will begin by giving an overview of the typical lesson sequence for the comparison treatment and treatment condition, and then explain the procedures in detail for each day of the study.

Overview of comparison treatment. The purpose of the comparison treatment was to have students engage with the documents and do meaningful work gathering evidence for both sides of the argument, without an explicit focus on warrants. It is possible that careful reading of texts for textual evidence may be sufficient for students to generate warranted arguments. The students in this group spent a greater portion of time reading and engaging with the texts directly than did the students in the treatment condition. They read the same documents each day that the treatment condition did, and the instruction followed a gradual release approach (Pearson & Gallagher, 1983), beginning with the teacher modeling the process of reading documents in order to find evidence to support a claim. The primary note taking format for the comparison treatment was a simple two-column template with a space to find and record evidence from the document that supported each side of the claim. Then at the bottom they wrote which side they felt had the most evidence to support it. They learned that evidence could be a direct quote or a paraphrase and how different predicates (e.g. “did cause the U.S. to invade” and “U.S. should have invaded” affect the evidence used to support an argument. On the last day, they looked across the whole document set and determined which side had the most evidence to support it and which evidence was the most convincing overall. Like the treatment group, they were given opportunities to write about the evidence they found and periodically share what they wrote with peers.

To be clear, the comparison treatment did not reflect how history is traditionally taught. Typically, students learn a single narrative of what happened as presented in a textbook (Reisman, 2012). It was designed as an effective comparison treatment to an emphasis on warrants, because evidence is also an essential part of crafting an effective argument, and the
Common Core Standards put a great deal of emphasis on finding evidence from text. It is another promising approach to argument writing. By comparing it alongside an emphasis on warrants, I can accelerate the development of my design by testing an intervention against another promising method. Needing to identify evidence that supports a particular argument is a key component of the PARCC exams, though this instruction was not directly modeled after the format of these exams (PARCC Writing Evidence Tables, 2015). See Appendix I for a scripted lesson excerpt and instructional materials from the Comparison Treatment.

**Overview of treatment condition.** The goal of the treatment condition was to teach writing a warrant as the part of the argument where students use heuristics and close reading to explain how their evidence supports the claim. This included teaching historical heuristics through the process of warranting an argument. Many popular approaches to instruction with heuristics focus on them as a tool for reading, and do not emphasize their usefulness for writing (e.g. Reisman, 2012). The instructor explained that once the claim was established, it become a lens to examine additional documents. The whole idea of reliability or bias of a document or source only takes on meaning when in service of a particular claim. Operationalizing warrants in this way was intended to help students understand that heuristics are thinking tools that historians created to help them reason about text and make arguments. The instructor also consistently modeled how the same evidence could be used to support either side of a particular claim, depending on the warrant being used, to help students understand the importance of warrants and the interpretative nature of history.

The principles of effective strategy instruction incorporating gradual release of responsibility (Pearson & Gallagher, 1983) were applied here. Gradual release of responsibility has been shown to be effective in improving writing achievement (Fisher & Frey, 2003). The strategy students learned is the warranting of claims through the process of answering warrant generating questions. Students were provided the additional scaffolding of sentence frames that
reflect different types of historical thinking (Graff, 2006; Stanford History Education Group website). Students had some time to write their own warrant before using that warrant in a paragraph. Like the comparison treatment, students also had practice writing warrants for different sides and predicates, and had time to share their writing with peers and get feedback. A scripted lesson excerpt and instructional materials for the Treatment Condition are available in Appendix J.

**Day 1: Pretest and intro to argument (all students together).** On this day, students took an initial pretest questionnaire to establish any differences between treatment and comparison treatment prior to the study in terms of their knowledge and feelings about argument writing, and to determine their prior knowledge of the Spanish American War and the explosion of the U.S.S. Maine. The teacher also completed a pretest to determine her knowledge of warrants and her previous work with the students in terms of argument writing.

After completing the pretest, all students were given a brief overview of the parts of an argument. They were asked to turn and talk then share out their response about the parts of an argument. Then the instructor presented a minilesson on three parts of argument: Claim, evidence and warrant. The instructor distributed a handout that included helpful hints for writing a warrant (Hillocks, 2011). They were shown an example of an argument about a contemporary basketball player and there was a brief opportunity to practice writing an everyday warrant given a claim (i.e. “Northwest High is the best high school in our conference”). They were asked to provide evidence (facts, examples quotes, etc.) and a warrant (how the evidence supports the claim. Finally the instructor framed that next week, they would be using what they know about basic argument structure and applying it to historical arguments, which have the same basic structure, use but use claims, evidence and warrants differently.

**Day 2 for comparison treatment.** Day two for the comparison treatment began with a review of the basic argument structure (claim, evidence and warrant) that was introduced on day
one. Students were given a timeline listing the events they had studied and the explosion of the U.S.S. Maine simply to orient them to the events in relation to what they had studied previously. They watched a brief introductory video from the “Historical Thinking Matters” website (Rosenzweig & Wineburg, 2008) which shared the events leading up the Spanish American War. They learned the central question for this mini-unit, “Why did the U.S. invade Cuba following the explosion of the U.S.S. Maine?” They were oriented to the possible claims (“Yes, the U.S.S. Maine explosion did cause the U.S. to invade Cuba” or “No the explosion of the U.S.S. Maine did no cause the U.S. to invade Cuba”).

The instructor then read Document A, The Spanish American War (textbook excerpt) aloud to build background knowledge. Students were prompted to underline anything they thought was important, and circle anything they thought was confusing as they read, though they were permitted to annotate in other ways as well. The instructor did a think aloud of the first three paragraphs, modeling what she thought was important and circling what was confusing, and annotating by adding questions or thoughts (we came to agreement on the words to be underlined prior to the study as we would both be teaching each condition). At the end of these three paragraphs, the instructor read the remainder aloud, and prompted students to continue to underline and circle what they found important and confusing. The instructor had students pair up and share what they underlined and circled, and had pairs share with the group, addressing any fundamental misconceptions.

The instructor prompted the students to read the document again, this time looking for evidence that would help them answer the research question. They created a T-Chart divided into two columns: “Yes, the explosion of the U.S.S. Maine caused the U.S. to invade Cuba” and “No, the explosion of the Maine did not cause the U.S. to invade Cuba”. The instructor modeled finding evidence from the text for each side of the argument. Students were instructed to read and gather evidence for the remainder of their period. At the end of the period, they turned to
their neighbor and shared one piece of evidence that they had that their neighbor did not, and added it to their charts. At the end of the period they had an exit ticket asking them what claim they felt had the most evidence to support it, and to explain their answers.

**Day 2 for treatment condition.** The treatment condition did everything the same as the comparison treatment until after the timeline was shared.

The focus of the first day was on sourcing. The instructor explained that sourcing included asking oneself warrant generating questions: “When (and where) was this written? What might the author’s perspective of purpose be in writing this, and how might that influence the version of events they shared? How does knowing that help me interpret the document? How does knowing the source relate to the focus question I am answering?”

The instructor read the first three paragraphs aloud, modeling thoughts about when the source was written, the source type, and the author. The students were then prompted to continue to read the document themselves, underlining anything they thought was important and circling anything they thought was confusing, thinking about the source of the document as they did so. At that point the instructors asked the group to share any questions they had. Then they made the connection to sourcing as a powerful way to warrant an argument. They modeled writing a warrant to explain how the claim “The sinking of the U.S.S. Maine caused the U.S. to invade Cuba”, and the evidence “After the Sinking of the Maine, ‘War fever’ gripped the nation. Spain agreed to abolish the reconcentration camps and make other concessions, but it was ‘too little too late’” were connected. Then the instructors showed how the claim, evidence, and warrant would look in the form of a paragraph. Students were then prompted to use sourcing to make the opposite claim, that the U.S.S. Maine did not cause the U.S. to invade Cuba. They were given a handout with the claim and evidence filled in, and they wrote a warrant connecting them. Finally they were prompted to write their response in paragraph form.
**Day 3 for comparison treatment.** Students in the comparison treatment traveled with the teacher or researcher, whoever was assigned to this group, to the computer lab. They were asked to summarize what they learned yesterday then were given a new document titled “Not an Accident, Capt. Sigsbee Says” to see what evidence it gave them to address the question, “Did the explosion of the U.S.S. Maine cause the U.S. to go to war with Cuba?” They again were prompted to read the document through once, underlining anything they thought was important and circling anything they thought was confusing. Then the instructor read it aloud again, prompting students to find anything they thought might be evidence to help address the research questions. The instructor then modeled how to use a T-chart on the overhead, divided in to two columns, “Yes the explosion of the Maine caused the invasion of Cuba,” and “No, the explosion of the Maine did not cause the invasion of Cuba.” The instructors modeled one example with a direct quote and one example with a paraphrase. The instructors called on students to find an additional piece of evidence and explain what side of the argument the evidence supported. Students were given a handout with a T-Chart and read and gathered evidence in the T-Chart for the remainder of the period. With ten minutes remaining, the instructors read aloud the fourth paragraph and had a volunteer put it in their own words to address any comprehension issues. Then they asked students to place the evidence in that paragraph into their T-Charts. At the end students turned to a neighbor and in a strategy called, “Give One, Get One” added one piece of evidence from their neighbor that they did not have, and shared one piece of evidence that their neighbor lacked. At the end of the period they completed an exit ticket. They were prompted to look back at that day and the previous day’s documents and answer the question “Now based on both documents as a whole, which claim do you think has the most evidence to support it? Explain.”

**Day 3 for treatment condition.** The treatment group was reminded that the previous day they had learned how to use historical thinking heuristics such as sourcing as a warrant to explain
how their evidence supported their claim. They were told, “When you write a historical argument, your warrant is where your historical thinking ‘lives’. Your historical thinking goes in your warrant.” They were reminded about the two claims they could make about their research question, “Yes the explosion of the U.S.S. Maine caused the U.S. to invade Spain”, or “No, the explosion of the Maine did not cause the U.S. to invade Spain,” and reminded how they used sourcing to write a warrant showing how the evidence supported the claim. The instructors then introduced the same document the comparison treatment had read, “Not an Accident, Capt. Sigsbee Says”. They modeled using sourcing and underlining what they thought was confusing and important. Then they put up an overhead with the claim and evidence filled in, and had students write a sourcing warrant together with a neighbor. After they were finished the instructors shared their warrant. Next they looked at the same evidence and thought about how it could be used to support the other side of the argument, using sentence frames.

At that point the instructors made a transition to talk about the next way to warrant an argument, close reading. They handed out a handout about close reading, and modeled how close reading for structure could be used to warrant the argument from the textbook by examining how the author chose to structure the chapter, dividing it into sections and showing all the events that built up to the U.S.S. Maine explosion. Ending with the explosion made it seem as if the explosion was the inevitable conclusion of the events.

Students were prompted to look at the World article and examine what came first and what came later, and to think about how they could use how an author chooses to structure the sequence of evidence to make a warrant about their argument. They were given a claim, “Yes, the explosion of the U.S.S. Maine caused the U.S. to invade Cuba”. As evidence, they examined the title, “Not an Accident, Capt. Sigsbee says” and the subtitle, “Disaster was due to an enemy”. They discussed how the author intentionally chose these as headlines, and how the headlines the author chose gave the impression that the explosion was caused by an enemy. However, later in
the document it stated that Captain Sigsbee actually was not confident in his assessment and wanted to keep his thoughts confidential until he could conduct an investigation. By choosing to structure the article with this information first, as the headline, and burying the Captain’s reservations later in the article, they saw how the media coverage of the Maine explosion could have helped cause the U.S. to invade Cuba. Students wrote their warrant to connect the evidence and claim. At the end of the period, warrants were shared and collected.

**Day 4 for comparison treatment.** Students in the comparison treatment were reminded about why evidence is important in making an effective argument and that it can be specific facts or details from a document, or information that is paraphrased. The instructors shared an example of a student response on the overhead. They talked about what made the response effective (that it was a complete idea, that multiple pieces of evidence were used, that evidence came from both documents, and that the source of the evidence in was included, e.g. (Doc A). Students had three minutes to improve their response, underlining anything they added that they didn’t include before.

The comparison treatment did a similar process of collecting and organizing evidence. However, on this day they were given a different predicate: “Should the U.S. have invaded Cuba following the explosion of the U.S.S. Maine?” The instructor shared that as they read, students could refine their claim to adjust the certainty or degree, but they were going to simply think in terms of “Yes the U.S. should have or invaded” or “No the U.S. should not have invaded.” Instructors discussed how students could start with a question and find evidence to determine what their claim should be, or generate the claim based on evidence as they read. The instructors introduced the fact that, even though the question asked “should”, it was not an opinion question. Students needed be impartial like a judge, weighing the evidence and deciding what side was strongest. They were asked to test their evidence to make sure it supported their claims and that they didn’t over or under generalize. They were then introduced to the document from the New
York Times titled, “The Maine Disaster.” They were asked to see what evidence it provided to address the question, “Should the U.S. have invaded Cuba following the explosion of the U.S.S. Maine?” Students were prompted to read the document through, underlining anything they thought important and circling anything that was confusing. Then the instructor read it aloud a second time, and asked students to find any evidence to support the research question.

The class made a T-Chart again, this time with the sides “Yes the explosion of the U.S.S Maine should have caused the U.S. to invade Cuba” and “No the explosion of the Maine should not have caused the U.S. to invade Cuba”. The instructors explained what those statements meant in terms of evidence, modeled a direct quote, and then had students do one together as a class. Ten minutes before the end of the lesson they read the last paragraph. The instructors had students turn to their neighbor and share their evidence with them. Students gave each other feedback on their evidence—whether it said why they should or should not have invaded, whether it gave a specific quote or paraphrase, and whether it was clear what the evidence was saying without returning the passage. At the end of the class period, they completed an exit ticket: “Based on the document you read today, what claim do you think has the most evidence to support it: Yes the U.S. should have invaded Cuba following the explosion of the Maine, or No, the U.S should not have invaded? Give evidence to support your position.”

**Day 4 for treatment condition.** The students in the treatment condition were reminded that the warrants they wrote the previous day were meant to reflect structure and how argument structure might have influenced people who read the document. The instructors put up an example and students were prompted to tell what this student did well. Then they were prompted to turn to a neighbor and discuss why warrants were important. The instructors called on a group to share, then reinforced the importance of warrants - that they explained how the evidence supported the claim and that evidence can be used to support either side of an argument.
Students were then informed that they were going to be writing claims, evidence and warrants that addressed a different question. Instead of “Did the explosion of the U.S.S. Maine cause the U.S. to invade Cuba” they were going to be addressing the question, “Should the U.S.S. Main have caused the U.S. to invade Cuba? They were prompted to think about how even though this question seemed like more of a value judgment, they would still need to use evidence to and warrants supporting their side. This time they were oriented to the New York Times article, “The Maine Disaster.” They were cued to the sourcing information and showed how the source was different from that of the previous day (the New York World).

Students were then informed that along with close reading of structure, they could also read closely for argument or word choice. They referred to their Close Reading handout and the warrant generating questions and sentence frames that accompanied that writing type. The instructor modeled closely reading the New York Times document for word choice. They modeled underlining what they thought was important and circling what was confusing, focusing on word choice. Examples of warrant generating questions were, “What language does the author use to convince the reader (words, images, symbol, etc.)? How do these words show the author’s perspective?” The instructors modeled underlining the word “theory,” and how the fact that this is a theory only supports the idea of whether or not the U.S. should invade Cuba. The instructors read the document aloud; prompting students to continue to underline and circle what they thought was important and confusing. They modeled writing a claim “The U.S. should not have invaded Cuba following the explosion of the U.S.S. Maine,” and the evidence “only theory as to the cause of the disaster.” Then they modeled how to write a warrant using close reading to explain how the evidence was tied to the claim, using sentence frames followed by brief guided practice.

Finally students were asked to choose a side for the argument and then choose evidence, namely a specific word or phrase that revealed something important that supported their
argument, or a word that showed what the author was trying to do or how they felt about the event. Then they were asked to write a warrant.

Students gave each other feedback on their warrants, picking one focus area for feedback (whether or not they were using historical thinking correctly, or whether it was accurate, relevant, appropriate in degree, or fitting with other evidence). In their meta-analysis of research on effective writing instruction, Graham and Perin (2007) found that collaborative writing, or “developing instructional arrangements whereby adolescents work together to plan, draft, revise and edit their compositions” had a strong impact on improving student writing. This was by no means collaborative writing writ large, but it was an opportunity for students to give each other feedback and revise and refine their warrants.

Day 5 for comparison treatment. Students used the same procedure as the previous day, still addressing the question “Should the U.S. have invaded Cuba following the explosion of the U.S.S Maine?” but this time with a different document, The Monroe Doctrine. They were told that this document was written over 70 years before the incident, from another president, James Monroe. He gave a speech saying the U.S. would not always stay isolated. He laid out the circumstances in which they would get involved in the affairs of other countries. The instructors read the document aloud and asked students to annotate by underlining what they thought was important, circling anything confusing, and writing questions in the margins. The instructors modeled finding evidence and putting it in the T-Chart for each side of the argument. Students were then asked to get out all the documents and evidence organizers from previous days. They were shown the two questions they examined, “Did the explosion of the U.S.S. Maine cause the U.S to invade Cuba?” and “Should the explosion of the Maine have caused the U.S. to invade Cuba?” With a partner, they looked at the T-Charts they completed for each text. They were given a document titled “Bringing it all together.” Of the four claims, they were asked to star the evidence they felt was most effective. Across from the evidence, they were asked to explain why
they thought that particular evidence was so effective. They did this collaboratively with their partner and the instructor circulated and gave feedback.

**Day 5 for treatment condition.** The treatment group began by reviewing the two forms of historical thinking they had studied so far, sourcing and close reading. This day the focus was on corroboration. They were told that the strongest historical arguments use evidence from more than one document, since no single document tells the full story. Instructors passed out a corroboration handout that explained three types of corroboration: corroborates with source (vote counting), corroboration of contradictory information, and corroboration in combination with sourcing. They were given an example of why employers ask for more than one reference on a job application to illustrate the importance of corroboration.

Instructors modeled the corroboration of contradictory information within documents for the claim, “The U.S. should not have invaded Cuba following the explosion of the U.S.S. Maine.” They thought aloud about how The World stated that the Maine was the “work of an enemy”, whereas The New York Times stated that there was “only theory” as to what happened, and the evidence tended to indicate an accident. Students were given warrant generation questions and sentence frames to help write the sentence. An example of a sentence generated from this work follows: “While the New York World and the New York Times disagree about whether or not the explosion was an accident, the New York Times is more reliable; the New York World tends to exaggerate and the Times is a more respected source. This supports my claim that the U.S. should not invade Cuba.”

Though the label “corroboration of contradictory information” may seem like a misnomer, one important historical thinking skill is not only reasoning about where documents agree, but also reasoning about when they disagree, and using those disagreements to warrant an argument. For the purpose of this study, this was included as an element of corroboration,
determining where documents agree and disagree, in order to expand the usefulness of the thinking involved for students.

Students then examined the Monroe Doctrine document. The instructors read the document aloud while students underlined parts that were confusing and important. The instructors modeled annotating the document. The instructors asked why this document was relevant, having been written 70 years beforehand. Then they thought aloud as they corroborating the evidence in this document with the rest of the documents in the set. They worked with partners to write a warrant together, and then paired up, giving each other feedback on their warrants, while the teacher circulated to give feedback to individual groups (Graham and Perin, 2007).

**Day 6: All students together- reading document set.** On Day 6, the treatment and comparison groups’ instruction was complete. Treatment and comparison groups came together in the classroom. Students were given a day to read a new set of documents about the same content and similar to the previous documents that would be used for the instrumentation. They were permitted to annotate documents as they saw fit, but were not prompted to do. Students had one class period to read the documents. They were informed they would be taking an assessment the next day.

On the end of the 6th day, five students in the treatment group completed a think aloud for one warrant selection and one warrant writing task. The purpose of the think aloud was to provide a snapshot of the thinking that students do as they complete the warrant selection and warrant writing tasks. An explanation of the think aloud process is available in Appendix K. These same students then completed a semi-structured interview with the researcher, in order to understand their perspective on what they learned this week and how it was similar or different from typical instruction. The semi-structured interview questions are included in Appendix L.
These students were removed from the larger data set because the experience of the think aloud may change their performance on the other assessments.

**Day 7: All students together- warrant selection and warrant ranking task.** On Day 7, students completed the warrant selection and ranking task. The directions were read aloud to everyone. The task was generated using Google Forms and the documents for students were arranged on a Google site. More and more, students are expected to engage with documents online. In addition, the students will be taking the online PARCC assessment, which is entirely screen-based. However, the teacher felt strongly that the students should interact with print copies of the documents, since she felt it was easier for them and it would be consistent with what they had done during the year, so we provided each student with a hard copy of all documents. However, they completed the warrant selection and ranking, the warrant writing, and the essay task on google forms embedded into a simple google site.

For the warrant selection task students were prompted to select effective warrants. Each item contained 2-3 correct responses, though students were not told how many items to select. For the warrant ranking tasks, students ranked the warrants that they had already selected in order of effectiveness.

**Day 8: All students together- warrant writing and essay writing.** On Day 8, students completed two warrant writing tasks and an essay writing task. They had 15 minutes to complete the two warrant writing tasks, and the remaining 30 minutes of the period to complete the essay. These students had been taught by the classroom teacher to think of a historical argument as a single paragraph, so the majority of students responded in the form of a single paragraph, while a handful of others wrote more than that for the essay task.

**Data Sources and Analysis**

In the following section I will discuss the data sources and the analysis I performed for each data source, arranged by research question, including information about interrater reliability.
for each analysis where relevant. I will also discuss my examination of the range of scores to determine whether the measures were sensitive to various degrees of student performance, and to contribute to interpretation of the results.

Due to the similar nature of many of the measures, the most reliable measures were identified and combined, and a MANOVA was run to determine whether there was an underlying effect. The Warrant Selection: Correct task was identified as one of the most reliable because it more closely reflected what students were actually taught during the intervention. In addition, several warrant writing measures related to historical thinking and quality were combined to create a single, more reliable measure. These two scores were then analyzed through MANOVA. Following the MANOVA, I ran a series of t-tests as a follow up to look for patterns that may exist in the data and may have contributed to any overall difference between the groups. I acknowledge the reliability problems in this sort of analysis, including higher likelihood of finding differences that due to multiple statistical analyses on the same data set. I am simply using this approach to identify patterns that may exist to help inform the design of the characteristics of an effective intervention and aid in the development of a theory of teaching historical argumentative writing to adolescents. I applied statistical corrections for the repeated analyses on the same data set.

**Research question group 1:** The impact of the treatment on measures related to historical warrants.

**Research question 1a.** Does a treatment teaching students to use sourcing, corroboration, and close reading as explicitly stated warrants through the use of warrant generating questions (Freeman, 2011; Toulmin, 1958) and templates (Graff & Birkenstein, 2007) improve student ability to select effective warrants, in comparison to a treatment focusing on finding evidence to support a claim?

**Warrant selection task.** Overall student ability to select effective warrants was measured by the warrant selection task. There were 12 total items. Each item had five possible answer
choices. Each choice of the 12-item assessment was scored as a true/false question, meaning every item had five possible points, for a total of 60 points.

SPSS was used to create a total sum score for each student. Students initially received a point for each possible response of the multiple select items. Since there were five points possible for each item, with a total of 60 points for the assessment, this meant they got the same number of points for selecting an effective warrant as they did for not selecting an ineffective warrant.

In order to get more information about students’ ability to select effective warrants, the five total points possible for each warrant selection task item for each were broken into “correct answers” and “not wrong” answers. There were 2-3 correct answers possible per item for a total of 30 possible “correct answer” points. SPSS was used to create a Warrant Selection: Correct sum for each student, which was included in the MANOVA analysis of the most reliable study measures. Following this MANOVA, I screened for normality, outliers, and homogeneity of variances and ran an independent-samples t-test with a Bonferroni correction to see if there was a significant difference in the performance of the two groups on the overall Warrant Selection Task.
Research question 1b. Do students in the treatment condition better discriminate between more effective and less effective warrants?

The Warrant Selection: Correct score provided some insight into student ability to discriminate between more effective and less effective warrants because I examined their ability to identify correct warrants separately from their ability to avoid selecting incorrect warrants. For instance, a student may select the two correct warrants, and an additional incorrect warrant for a given item, indicating he may have an idea of what makes a warrant effective, but has difficulty avoiding less effective warrants.

Warrant ranking task. Another way that the ability to discriminate between more effective and less effective warrants was measured was through the warrant ranking task. Immediately following the warrant selection task, students ranked the answers they selected as
effective by order of effectiveness. Students received one possible point per item. The points were totaled to make a Sum of Warrant Ranking score, with a maximum score possible of 12 points.

Scoring this task was complicated by the fact that students could have selected any number of correct answers. Because of the nature of the assessment design, some students chose only one response, which meant there was nothing for them to rank. For students that only initially selected one warrant, a point for ranking was awarded if the one warrant they selected was effective and no points were awarded if it was ineffective. For students that selected more than one answer, I looked at the initial pair of items that the student ranked and assigned one point if the more effective warrant was in the primary spot. An independent samples t-test with a Bonferroni correction was run to compare the performance in the treatment and comparison treatment groups.

By selecting only one correct answer, it could be argued that these students were already discriminating between warrants of varying levels of effectiveness, and they should be retained in the data set. However, it is also problematic to treat them as the same since they did not complete the ranking task. Therefore, I analyzed the subset of 21 students who selected at least two or more answers for every item of the 12 item set to compare those findings to the results of the larger data set with students who selected only one warrant included.

**Research Question 1c.** Do students in the treatment condition write more effective warrants when given a claim and evidence than students in the comparison treatment?

**Warrant writing task.** The ability to write effective warrants when given a claim and evidence was measured through two warrant writing task items. Students were given a claim and evidence, and were prompted to write a warrant explaining how the evidence supported the claim. They were not cued to use historical reasoning.
The writing that students produced for the two warrant writing tasks was read several times and coded for evidence of historical thinking using the same codes that were used to develop the warrant selection instrument, related to sourcing, corroboration, and close reading. Students were given a point for each type of historical thinking their response demonstrated. Therefore, there were a total of three possible points for each warrant writing task item. The points for each type of thinking were added across both items to create total scores for each type of thinking and an overall Warrant Writing: Historical Thinking score. I calculated the mean and standard deviation for these categories for treatment and comparison treatment.

Responses were also coded for other indicators of quality including addressing the prompt, accuracy, including a direct quote, and including explicit references to two or more texts. They were given a point for each of these elements. The points for each indicator of quality were added across both items to create total scores for each indicator of quality and an overall Warrant Writing: Quality score.

There were two items, seven codes per item, and 84 participants scored, for a total of 588 codes. I addressed interrater reliability by coding a subset of 10% of items with a literacy professor who was also a former AP history teacher. We were in agreement on 88% of the codes on this subset of the items.

I used SPSS to calculate the mean and standard deviation for each of these measures for the treatment and comparison treatment conditions for both the Warrant Writing: Historical Thinking and the Warrant Writing: Quality measures. In order to increase the reliability of these measures, I created a sum of points earned on both of these measures to create an overall Warrant Writing: Combined score. This combined warrant writing score was examined along with the Warrant Writing: Correct score as part of the MANOVA. Following this, I examined the differences between individual elements of historical thinking and quality using a series of t-tests.
with Bonferroni corrections, in order to begin to understand what may have contributed to the overall effect. An example of a warrant writing task is shown in Figure 3.2.

**Directions**

There are two paragraphs from two different historical argument essays about the explosion of the U.S.S. Maine below. Each paragraph has a claim and evidence, but it is missing a warrant. Finish each paragraph by writing a warrant in the space below each paragraph that explains how the evidence supports the claim. Be sure to read the claim and evidence carefully. It does not matter whether or not you personally agree with the argument. Your job is to write the best warrant you can. You have the remainder of the class period.

3. The explosion of the U.S.S. Maine did not cause the U.S. to invade Cuba. In his State of the Union Address (Doc D), President McKinley stated there was “serious injury” to the trade and business of Americans who were either working in Cuba, or who had businesses there. He said “reckless destruction” of property on the island had been occurring as the “fight waged for years” between the Spanish and the Cuban rebels. McKinley argued this had left many of these Americans with their “property destroyed and themselves ruined” (Doc D). 

Now write a warrant that explains how the evidence supports the claim.

*Figure 3.2. Example of Warrant Writing Task*

**Research question group 2**: Student performance on items designed to reflect specific types of historical thinking and the relationship between skills related to historical argument writing.

**Research question 2a**: What is the relationship between students’ ability to select effective warrants, discriminate between more effective and less effective warrants, and write warrants demonstrating historical reasoning as measured by correlations and patterns in student responses?

**Cut scores for warrant selection and warrant ranking**: In order to better characterize the data set and how the performance on warrant selection and warrant ranking compared between the two conditions, I set cut scores for performance on the warrant selection and warrant ranking
tasks (Dwyer, 1996). The cut scores created categories of student performance. There is no one approach to setting cut scores, and they always are a judgment call. Cut scores inherently misrepresent data because there is not often a clear difference between passing and failing performance, when it in essence exists on continuum. Whether or not established cut scores are valid depends in the knowledge of the person setting the cut score (Dwyer, 1996). However, creating cut scores can be useful for characterizing a data set when there are not high stakes attached and when the purpose is to draw conclusions about students’ relative strength in the treatment and comparison treatments.

The cut scores served to divide student performance into three categories: Advanced, Proficient and Below Proficient, for both the warrant selection task and the warrant ranking tasks. I created a bar graph to visually represent the categories in which students fell for each task, in order to better illustrate trends in student performance and the relationship between student performances on each task.

**Correlations.** Finally, correlations were run in SPSS in order to better understand the relationship between student performance on the warrant selection and warrant ranking tasks. A separate correlation was also run to determine the relationship between student performance on the warrant selection and the warrant writing tasks.

**Research question 2b.** How do students perform on warrant selection items meant to reflect specific types of historical thinking?

**Item analysis of thinking types.** In the design of the assessment, warrant selection items were crafted to reflect different types of thinking, namely sourcing, corroboration, close reading, and non-historical and specific subtypes of thinking for each larger thinking type. Across the 60 individual answer choices there are 16 items each meant to reflect sourcing, corroboration and close reading, with the remaining 12 items reflecting non-historical thinking. Performance on items linked to each type of thinking was determined using a series of independent-samples t-tests.
with Bonferroni corrections. Finally, the performance on items excluding the non-historical thinking items was determined, since non-historical thinking was not a primary focus of the study and it was possible that the inclusion of those items was adding noise to the data.

**Research question 2c.** How do students perform on writing warrants given claims and evidence meant to stimulate specific types of historical thinking?

**Warrant writing task.** The warrant writing items were designed to give students opportunities to engage in historical thinking, without limiting the potential heuristics students could use. In this task, students were given a claim and evidence and asked to write a warrant. The task could stimulate different sorts of historical thinking by the type of evidence it provided. For instance, by presenting evidence from sources of varying levels of credibility or relevance to the claim it could stimulate sourcing. By providing contradictory evidence, it could stimulate corroboration of conflicting evidence. By providing dates that were significant, it could stimulate sourcing that attended to when a document was written. By including quotes of significant words or phrases with figurative meaning or significant tone, it could stimulate close reading.

In addition to an overall historical thinking score and analysis of other indicators of quality, I analyzed the evidence of each type of historical thinking (sourcing, corroboration, close reading) and determined the mean scores for each type of thinking. I situated this analysis within what I would expect based on the types of thinking the items were likely to stimulate.

The items were also designed to assess student performance based on the predicate and side of the argument. Each of the two predicates (Did Cause and Should Have Caused) had six total warrant selection task items, for a total of 30 possible points each. Each predicate was broken into two sides (Yes, Did Cause/No, Did Not Cause and Yes, Should Have Caused/No, Should Not Have Caused) for a total of 15 possible points for each side of each predicate. In order to determine if “Myside Bias” (Wolfe, Britt & Butler2009) was a factor in student performance, students were asked what they personally believed.
Essay writing task. Finally, students were asked to write an essay response to one of the central questions that they were considering this week, namely “Did the explosion of the U.S.S. Maine cause the U.S. to invade Cuba?” Students were directed to use the text set and what they learned in the past week to answer the question. They were given 30 minutes to write a “well-developed response.” Mrs. Jones told me that students had been taught to respond to essay prompts in a single well-developed paragraph in their previous instruction. She said that they would likely assume the prompt meant to limit their writing to a paragraph. In order to not limit students, I included in the prompt the language “Your response can be one paragraph or more”.

In order to examine this essay writing data, I created quartiles in SPSS based on the primary data set for this study, which was the warrant selection task. Then I determined the mean of all scores within each quartile. The division between the first and second quartile (25th percentile) was a score of 34 (out of 60). The division between the second and third quartile (50th percentile) was a score of 38. The division between the third and fourth quartile (75th percentile) was a score of 42.

Then I determined the mean of the scores within each quartile. The first quartile contained scores ranging from 25-34, and the mean was 29.57. The second quartile contained scores ranging from 35-38, with a mean of 36.5. The third quartile scores ranged from 39-42, and the mean was 40.5. The fourth quartile spanned from 43-51, with a mean of 47.24.

Finally, I selected the two papers each for treatment and comparison treatment that were closest to the mean of that quartile. I analyzed these further by identifying the elements of Toulmin’s argument pattern (Toulmin, 1958) and then analyzing the types of historical thinking present within the warrant of each, in the same manner I did for the warrant writing task, coding for evidence of sourcing, corroboration, and close reading. Within each heuristic code, I additionally coded for the subtype of that heuristic, using the same framework I used for the design of the warrant selection instrument.
**Research Question Group 3**: The thinking of students as they complete the assessment tasks and what they report learning from the treatment.

**Research Question 3a**: What sort of thinking does a subset of students engage in as they complete the warrant selection, warrant ranking, and warrant writing tasks?

*Sampling and overview of think aloud and semi-structured interview.* A small group of five students completed a think aloud and a semi-structured interview. However, one student was removed from this portion of the analysis because she did not complete the warrant writing think aloud due to an issue with the school schedule.

The stratified sample was selected based on students’ sophomore PLAN score for a think aloud protocol analysis (Ericsson & Simon, 1993, Smagorinsky, 1989). There is a long history of research using think-alouds in reading in general (Kucan & Beck, 1997) and of reading complex primary sources (e.g. Afflerbach & VanSledright, 2001). However, Wineburg’s work and subsequent research focused on comprehension of documents, whereas mine included a protocol analysis of their item responses and the generation of warrants, which is a subset of the composition process (Smagorinsky, 1989). I conducted think alouds with these students to gain insight into the thinking students were doing as they selected warrants, ranked warrants, and wrote their own warrants when given a claim and evidence. I also conducted semi-structured interviews with these students to get their perspective on what they learned during the week and how similar or different the intervention was from normal classroom instruction. This qualitative data helped me interpret the quantitative findings and think about the implications of these findings for instruction and future research. Think alouds were one class period (40 minutes after transition time) in length, with time divided equally between the think aloud and the semi-structured interviews.

*Think aloud.* The session began with a think aloud of one item from the warrant selection task and continued with the corresponding warrant ranking task, followed by one warrant writing
Student data was analyzed qualitatively, also using the categories of types of historical thinking (sourcing, close reading, and corroboration), subtypes, and categories of incorrect responses. Prompting was minimal. I modeled for them what a think aloud entails, with an example related to writing, but not to the topic in question. If students stopped explaining their decisions or if they did not elaborate on a particular decision, I prompted students using undirected probes to explain their thinking, e.g. “Why did you write this?” or “Keep going” (van Someren, Barnard & Sandberg, 1994) with as little interference as possible. Direct prompts such as “Did you use this warrant to show how the evidence proved your claim?” which may have unduly guided the participants thinking, were avoided.

The think alouds were transcribed and coded for evidence of sourcing, corroboration, or close reading. Since some students expressed difficulty in doing the think aloud and working at the same time, the researcher also used retrospective reports immediately after the task while much information was still in short term memory (Ericsson & Simon, 1993).

There has been some criticism of think aloud protocols as empirical data (e.g. Nisbett & Wilson, 1977) due to the claims that the prompting may modify the cognitive process the researcher is meaning to understand or that people are often unable to provide clear reports of higher order or inferential processes. However, prompts meant to not modify the cognitive process were used (Ericsson & Simon, 1993). Likewise, verbal reports have been used in the study of writing in particular (Smagorinsky, 1989), and can be depended upon when conditions are relatively natural, e.g. a classroom setting as opposed to a laboratory environment, being sensitive to variables that may affect the transcript, situating it with what we know about the writer, and emphasizing “level I” verbalizations, in which respondents explain their thinking and no new information is introduced (Smagorinsky, 1989).

**Research question 3b.** How do these students feel the intervention is different from typical classroom instruction and what do they report they have learned?
**Semi-structured interview.** The students were given the semi-structured interview after the completion of the instruction and immediately following a think-aloud task that they completed with the researcher. The interview was divided into clusters of questions meant to examine different aspects of their thinking about the week’s instruction.

The first cluster contained two questions that addressed student’s thinking about the events they studied this week (1. Do you think the explosion of the U.S.S. Maine caused the U.S. to invade Cuba? Why or why not? 2. Do you think the U.S. should have invaded Cuba following the explosion of the U.S.S. Maine? Why or why not?). These questions were meant to capture their processing of the content of the event.

The second cluster of questions dealt with argumentation and the role of warrants in general (3. What are warrants? 4. What makes a good warrant? 5. What role do warrants play in an argument?). These were coded for general structure of argument. The purpose of these questions was not to elicit historical thinking, but to gather information about how they thought about warrants.

The third cluster of questions contained four questions that prompted students to be metacognitive and reflect on what they were doing or thinking about and how that was similar or different from what they usually do in English class (6. What have you learned this week? 7. Tell me more about the warrant selection task. How did it go? 8. Tell me about the warrant you wrote. How do you think you did? 9. How is this week similar or different from what you normally do in U.S. History class?).

There was a final question that asked students to reflect on what might have been helpful to prepare them to complete the warrant ranking and warrant writing tasks (10. What would make this work easier, For instance, more time to practice, more feedback, etc.?). The codes for this task were straightforward, based on whatever students said would make it easier.
Items 1 and 2 and 6-9 were coded for historical thinking using the same codes used to develop the Warrant Selection and Ranking instrument because there was potential for these to include discussion of historical thinking heuristics. However, it is important to note that students were not prompted to discuss historical thinking heuristics, and it was possible to answer these items without mentioning historical thinking heuristics.

Responses were parsed into comments, which were bursts of speech that followed a student reading or interacting with a text, and events, or idea units that represented a unique idea on a particular topic, often reflecting a reading or thinking process (Wolfe & Goldman, 2005; Manderino, 2011). Relevant idea units were coded with the type of thinking (i.e. sourcing, close reading, or corroboration).
IV. RESULTS

The previous chapter detailed the methods I employed in my research and the rationale behind those methodological decisions. In this chapter I provide the results of the data I collected. First, I present the MANOVA analysis of the most reliable measures in the study to determine the overall impact of the different approaches to instruction, in order to determine if the treatment approach “works” at the local level. Then I organize the remaining analysis by the corresponding sub questions they address to help determine the characteristics of an effective intervention to promote the use of historical thinking heuristics as warrants within the historical argumentative writing of adolescents. This will enable me to generate theory about teaching argumentative writing which will inform work in similar contexts.

The analyses focus on student performance on a pretest, a warrant selection task, a warrant ranking task, a warrant writing task, and an essay writing task. They also include analysis of data from a short think aloud protocol and semi-structured interview with a subset of four students in the treatment condition, to provide texture to the results of the primary analyses. This chapter will report results corresponding to each of my research questions.

Pretest

Students took a brief survey prior to treatment in order to determine level of prior knowledge about the explosion of the U.S.S. Maine and the Spanish-American War and to determine whether or not I needed to control for any prior differences between treatment and comparison treatment groups. The majority of students said they knew nothing about the Spanish American War. Only two students had any specific knowledge about the U.S.S. Maine. The
random assignment to treatment helped to insure any differences would be randomly distributed across treatment and comparison treatment conditions so I did not examine this data further.

In addition, the students were asked two questions about their prior knowledge of argument writing: “What makes a written argument effective?” and “Name the parts of an effective written argument.” These questions were asked to determine what students already knew about argument structure and to determine whether there were any differences between treatment and comparison treatment. The answers to the question, “What makes a written argument effective?,” varied widely and were not amenable to quantitative analysis, so I included that only in adding texture to the think aloud data in later analysis.

The answer to the question “Name the parts of an effective written argument”, however, was analyzed further. First of all, a point was assigned for each of the three core elements of Toulmin’s argument pattern (claim, evidence, and warrant). Since this was a pretest and students were not introduced to the term “warrant” in their class, I accepted words that were similar, and gave students the benefit of the doubt. For instance, I accepted the word “opinion” and the phrase “the point you are trying to make” for claim and “facts” for evidence. I accepted the word “reasoning”, “elaboration” and “explaining the evidence” for warrant. Though they did not exactly reflect a warrant, I wanted to give students credit for their prior knowledge of argument structure. 37 of 64 students, or 57.8% named all three parts of an argument. Nine additional students or 14.1% named two elements of an argument, and nine students or 6.3% named only one part of an argument. Fourteen students or 21.9% named no parts of an argument. Examples of responses that did not name parts of an argument included “I don’t know” or “the body paragraph.” With that being said, of students who correctly identified the three parts of an argument, the majority of these referred to the parts as claim, evidence, and reasoning, reflecting the fact that the teacher had taught these parts of an argument to students previously.
Students were given one point for each of the three parts of an argument, with a total of three points possible. The mean score was 2.08 overall, 2.0 for the treatment condition and 2.14 for the comparison treatment. Though this difference is not statistically significant, it is notable that the comparison treatment had a higher mean score of the parts of an argument prior to the intervention. This data was used as part of intercorrelation with other measures, and was also used in analysis of the four students that were part of the think aloud and semi-structured interview task.

If I were to have been more strict in my criteria for student definition of a warrant (e.g. requiring students to actually use the word warrant in their responses) I would have missed that the majority of students did have prior knowledge of an element of argumentative writing that functions similarly to a warrant. I would also have missed the potential for confusion between the terms “reasoning” and “warrant”, an issue that one of the students raised during the think aloud. While students had been exposed to generic argument structure, classroom discussion indicated they thought about heuristics as tools to use exclusively during reading, and did not realize they could be used within writing to warrant claims.

The survey also asked students to “self-evaluate your ability at argument writing”. They were given a scale of 1-5, one being the lowest rating, not at all effective, and five being the highest rating, extremely effective. The overall mean score for 64 students was $M = 3.13$. Again, the mean score of the treatment condition ($M = 3.04$) was slightly lower than the comparison treatment ($M = 3.19$). I ran an independent samples t-test and determined that this difference was not statistically significant ($M = -0.15$), 95% CI [-0.54 to 0.24], $t(62) = -0.78$, $p = 0.33$.

Next, the survey prompted the students to answer the question “How do you feel about argument writing, compared to other things you study in school?” They were again given a scale of 1-5, one being the lowest rating, one of your least favorite things to study in school, and five being the highest rating, one of your favorite things to study in school. The overall mean score for
64 students was 2.48 and the range was 3. The mean of the treatment condition \( M = 2.52 \) was slightly higher than the mean of the comparison treatment \( M = 2.46 \). I ran an independent samples t-test and determined that this difference was not statistically significant (\( M = 0.06 \), 95\% CI [-0.42 to 0.54, \( t(62) = 0.25, p = 0.40 \]). Since these were not statistically significant, they were not analyzed further, but were used to inform analysis of the data of specific students in the think aloud/semi-structured interview conditions.

**Research Sub Questions**

This portion of the research addressed the three sub question groups that follow:

- **Research question group 1**: The impact of the treatment on measures related to historical warrants.

  - **Research question 1a.** Does a treatment teaching students to use sourcing, corroboration, and close reading as explicitly stated warrants through the use of warrant generating questions (Freeman, 2011; Toulmin, 1958) and templates (Graff & Birkenstein, 2007) improve student ability to select effective warrants, in comparison to a treatment focusing on finding evidence to support a claim? 1b. Do students in the treatment condition better discriminate between more effective and less effective warrants? 1c. Do students in the treatment write more effective warrants when given a claim and evidence than students in the comparison treatment?

- **Research question group 2**: Student performance on items designed to reflect specific types of historical thinking and the relationship between skills related to historical argument writing.

  - **Research question 2a.** What is the relationship between students’ ability to select effective warrants, discriminate between more effective and less effective warrants, and write warrants demonstrating historical reasoning as measured by correlations and patterns in student responses? 2b. How do students perform on warrant selection items
meant to reflect specific types of historical thinking? 2c. How do students perform on writing warrants given claims and evidence meant to stimulate specific types of historical thinking?

**Research question group 3:** The thinking of students as they complete the assessment tasks and what they report learning from the treatment.

**Research question 3a.** What sort of thinking does a subset of students engage in as they complete the warrant selection, warrant ranking, and warrant writing tasks? 3b. How do these students feel the intervention is different from typical classroom instruction and what do they report they have learned?

**Intercorrelation of study measures**

In order to further investigate the validity of my measurements, I conducted a bivariate correlation (Pearson’s r) among the different variables. I also included pretest data to examine the relationship between pretest and outcome measures. In this instance, I included Likert scale data with the data that I generated as an outcome of my research. I also included an additional measure of pretest data regarding the structure of students’ written arguments, because this data could be considered interval in nature and these numerical scores actually reflected a good deal of variation in student responses to the prompt.

I ran a Pearson’s product-moment correlation in SPSS to obtain the value of the Pearson correlation coefficient. I used Cohen’s (1988) guidelines about interpreting the strength of a correlation by using the coefficient value. One of the reasons to examine intercorrelations of study measures is to establish convergent validity, that measures that should theoretically be related (for instance, that test similar constructs) are in reality related. The related idea of divergent validity states that there should be a relatively lower correlation between measures that address theoretically different constructs. In the analysis below, I describe the relationship between the different measures in the study. In my study I did not design items that intentionally
tested constructs not at the focus of my study so I will focus on convergent validity in my analysis. While there were several assessments where there was no significant relationship, there was a relationship between some of the measures in this study. I will discuss significant and moderate relationships; the remaining results appear in Table 4.1.

There was a moderate or strong relationship between many of the pretest measures. There is a moderate correlation between the “Feeling about argument writing” pretest measure and the pretest measure of identifying “Parts of an argument”, \( r = .39 \). There is also a moderate correlation between the pretest “Self-evaluation of argument writing” and the “Parts of an effective argument” measure, \( r = .34 \). There is a strong correlation between the pretest measures “Self-evaluation of argument writing” and “Feelings about argument writing”, \( r = .51 \).

In terms of the relationship between measures in the study, there are several measures which have small but not significant correlations. However, there is a strong relationship between the measure of historical thinking within the warrant writing task, and the measure of writing quality within that same task. This is not surprising as the quality measures (i.e. addressing the prompt, accuracy, including a direct quote, and including explicit references to two or more texts) may enable historical thinking. There was also a moderate relationship between the historical thinking students showed in their arguments and the pretest “Parts of an argument” measure, \( r = .29 \).

Though the strength of relationship varied, there was evidence of correlation between pretest measures of argument writing and students’ ability to write arguments that reflected historical thinking, indicating that the measure may be discriminating between writers of different aptitudes. The lack of a significant relationship between the ability to select effective warrants and write effective warrants may have been impacted by the complexity of the warrant selection task.
Table 4.1

Intercorrelations for Key Study Measures

<table>
<thead>
<tr>
<th></th>
<th>M (SD)</th>
<th>Select</th>
<th>Select Corr.</th>
<th>Rank</th>
<th>Write HT.</th>
<th>Write Qual</th>
<th>Pretest Parts.</th>
<th>Pretest Feel</th>
<th>Pretest Self</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select</td>
<td>37.85(6.06)</td>
<td>1</td>
<td>.67**</td>
<td>.12</td>
<td>.15</td>
<td>.17</td>
<td>.11</td>
<td>.04</td>
<td>.16</td>
</tr>
<tr>
<td>Sel.Corr</td>
<td>17.42(4.73)</td>
<td>1</td>
<td>-.01</td>
<td>.06</td>
<td>-.01</td>
<td>.17</td>
<td>.18</td>
<td>.20</td>
<td></td>
</tr>
<tr>
<td>Rank</td>
<td>6.84 (1.49)</td>
<td>1</td>
<td>-.01</td>
<td>.07</td>
<td>.09</td>
<td>-.05</td>
<td>.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writ.HT</td>
<td>1.75 (1.22)</td>
<td>1</td>
<td>.60**</td>
<td>.29*</td>
<td>.21</td>
<td>.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writ.Q</td>
<td>3.68 (1.28)</td>
<td>1</td>
<td>.25</td>
<td>.24</td>
<td>.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Part</td>
<td>2.08 (1.24)</td>
<td>1</td>
<td>.39*</td>
<td>.34**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Feel</td>
<td>2.48 (0.94)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.51**</td>
<td></td>
</tr>
<tr>
<td>Pre-Self</td>
<td>3.13 (0.77)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

** Correlation is significant at the 0.01 level (2-tailed)

*Correlation is significant at the 0.05 level (2-tailed)

Notes. N’s range from 53 to 65 due to occasional missing data. Select = Warrant Selection; Select Corr. = Warrant Selection: Correct Responses; Rank = Warrant Ranking; Write HT = Warrant Writing: Historical Thinking; Write Qual = Warrant Writing: Quality; Pre-Part = Pretest: Parts of an Argument; Pre-Feel = Pretest: Feelings about Argument Writing; Pre-Self = Pretest: Self-evaluation of argument writing.
MANOVA analysis of most reliable dependent variables

My study had limited power due to small sample sizes. In order to increase the power and determine whether or not there was an overall effect from the two instructional approaches in this study, I selected the most reliable measures for this analysis. It is reasonable to assume that the treatment in the study had a greater effect on students’ ability to select effective warrants than it did on avoiding incorrect warrants since it did not emphasize discriminating fine variations between more and less effective warrants.

Since students earned points for both selecting a correct response and not selecting an incorrect response, I calculated a Warrant Selection: Correct variable in SPSS that tallied only the points for selecting correct warrants. This included the scores from items tagged to all types of thinking (i.e. sourcing, close reading, corroboration, and non-historical). I screened for normality by running descriptive statistics in SPSS and dividing the skewness and kurtosis statistics by their standard errors, and determining that the number was not larger than the absolute value of 2. I screened by outliers by converting the Sum of Warrant Selection: Correct to a z score and determining students two or more standard deviations from the mean. There were two outliers, but after conferring with the classroom teacher they did not represent a different population and were retained in the analysis. The descriptive data is reported in Table 4.3.

There was homogeneity of variances for Warrant Selection: Correct, as assessed by Levine’s test for equality of variances \((p = 0.77)\). The mean scores of the treatment condition \((M = 19.16, \ SD = 4.26)\) and the comparison treatment \((M = 15.97, \ SD = 4.67)\) were determined.

I then added the Warrant Writing: Historical Thinking and the Warrant Writing: Quality score into a more reliable Warrant Writing: Combined score. These scores were also analyzed for normality and homogeneity of variance separately and are reported in the subquestion analysis. I used MANOVA to see the overall effect of these measures, and then did follow up
ANOVAs to determine whether or not either individual measure had a statistically significant effect.

The descriptive statistics for the treatment and comparison treatment conditions are reported in Table 4.2 and Table 4.3.

Table 4.2

Warrant Writing Task - Combined (Historical Thinking and Quality)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>53</td>
<td>5.43</td>
<td>2.24</td>
<td>1-11</td>
</tr>
<tr>
<td>Treatment</td>
<td>25</td>
<td>6.00</td>
<td>2.43</td>
<td>3-11</td>
</tr>
<tr>
<td>Comparison</td>
<td>28</td>
<td>4.93</td>
<td>1.96</td>
<td>1-9</td>
</tr>
</tbody>
</table>

Note: total possible points = 14. 3 possible points for historical thinking (sourcing, corroboration, close reading) and 4 possible points for quality (addresses prompt, accurate, references 2 texts, contains quote) times two items.
Table 4.3

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>55</td>
<td>17.42</td>
<td>4.73</td>
<td>5-27</td>
</tr>
<tr>
<td>Treatment</td>
<td>25</td>
<td>19.16</td>
<td>4.26</td>
<td>12-27</td>
</tr>
<tr>
<td>Comparison Treatment</td>
<td>30</td>
<td>15.97</td>
<td>4.67</td>
<td>5-26</td>
</tr>
</tbody>
</table>

Note: Total points possible = 30 (2 or 3 correct answers spread across 12 items)

I began by ensuring my study met the assumptions of MANOVA, namely that there were independent observations and adequate sample size in each group. The data had been previously screened for univariate normality and univariate outliers. I screened for multicollinearity by examining the Pearson correlation among the new writing variable combining writing quality and historical thinking, and the “Warrant Selection: Correct” variable. There was no multicollinearity as assessed by the Pearson correlation ($r = .03, p = .87$). I tested for homogeneity of variance-covariance using Box’s M test of equality of covariance. Box’s M test was not significant, $p = 0.33$, meaning the homogeneity assumption was not violated.

There was a statistically significant difference between treatment and comparison treatment groups on the combined dependent variables for Warrant Selection: Correct and Warrant Writing: Combined, $F(2, 38) = 4.68, p = .015$; Wilks' $\Lambda = .80$; partial $\eta^2 = .20$. Follow
up univariate ANOVAs showed that there was a statistically significant difference in Warrant Selection: Correct scores between students from different groups, $F(1, 39) = 6.95, p = .012$; partial $\eta^2 = .151$, with a Bonferroni-corrected alpha level of .025 for two dependent variables. However, there was not a statistically significant difference on the Warrant Writing: Combined variable between students from different groups, $F(1, 39) = 2.09, p = .156$; partial $\eta^2 = .05$, with a Bonferroni-corrected alpha level of .025. This means that while the overall effect was significant, the Warrant Selection: Correct score had the greatest contribution to the overall significant effect. When examined separately, only the Warrant Selection: Correct effect was statistically significant.

After determining that there was indeed an overall effect, I did a series of analyses including t-tests to understand what may have contributed to this difference. I understand the limitations of this approach and used a Bonferroni correction to account for the increased possibility of finding differences. My purpose in this portion of the study was to determine patterns in the data to inform the characteristics of an effective intervention and the development of theory of teaching historical argumentative writing to adolescents. I also did additional qualitative analyses of pretest and interview data to add additional insight to these patterns.

**Analysis for research question 1:**

**Research question 1a.** Does a treatment teaching students to use sourcing, corroboration, and close reading as explicitly stated warrants through the use of warrant generating questions (Freeman, 2011; Toulmin, 1958) and templates (Graff & Birkenstein, 2007) improve student ability to select effective warrants, in comparison to a treatment focusing on finding evidence to support a claim?

**Warrant selection task:** There were 25 participants in the treatment group and 31 participants in the comparison treatment for this portion of the study. One outlier was removed from the comparison treatment, bringing the comparison treatment total to 30. The initial task
that students completed was the 12-item warrant selection task, in which students were given a claim and evidence, and selected as many warrants as they thought were effective from a list of five warrants. Each item had either two or three effective warrants; students were not cued how many answers to select. Each answer correctly selected (or correctly avoided) was awarded a point, for a total of 5 points per item. The total points earned per item were totaled into a Sum of Warrant Selection score for each student, out of 60 possible points for the Warrant Selection Task.

I ran an independent samples t-test to determine if there was a significant difference in the ability to select effective warrants between treatment and comparison treatment conditions. I screened for normality by running descriptive statistics in the statistical software SPSS and dividing the skewness and kurtosis statistics by their standard errors, and determining that the number was not larger than the absolute value of 2. The descriptive data is reported in Table 4.4.

I screened for outliers by converting the Sum of Warrant Ranking score to a z score and identifying students two or more standard deviations from the mean. Three students were above two standard deviations from the mean, all in the comparison treatment. I inquired with their classroom teacher about these students and she indicated that the student with the largest standard deviation from the mean ($SD = 2.35$) was misplaced in her class and should have been in an Honors section. In addition, her ACT-PLAN composite score of 24 was significantly higher ($SD = 3.45$) than the rest of the class ($M = 16.32$). I determined she represented a separate population from the remainder of the class and removed her from this and all further analysis.

There was homogeneity of variances for sum of warrant selection scores, as assessed by Levine’s test for equality of variances ($p = 0.78$). The mean scores of the treatment condition ($M = 39.32$, $SD = 5.76$) and the comparison treatment ($M = 36.63$, $SD = 6.13$) were determined. The difference in mean scores $M = 2.69$, 95% CI [-0.55 to 5.93], $t(53) = 1.66$, $p = 0.10$ was not
statistically significant, particularly after a Bonferroni-adjusted significance level of .005 was calculated to account for the increased possibility of type-I error. There is not a significant difference between treatment and comparison treatment with relation to warrant selection ability.

Since this test was not significant, a post-hoc power analysis was conducted using the statistical software PASS (NCSS Statistical Software). Group sample sizes of 80 each for comparison and treatment and a total \( n \) of 160 are needed to achieve 80.46\% power to reject the null hypothesis of equal means when the population mean difference is \( \mu_1 - \mu_2 = 39.3 - 36.6 = 2.7 \) with a standard deviation for both groups of 6.0 and with an alpha level of 0.05 using a two-sided two-sample equal-variance t-test.

Table 4.4

<table>
<thead>
<tr>
<th>Warrant Selection Task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td>Treatment</td>
</tr>
<tr>
<td>Comparison Treatment</td>
</tr>
</tbody>
</table>
Research question 1b. Do students in the treatment condition better discriminate between more effective and less effective warrants?

Selection of “correct” vs “not wrong” warrants.

In addition to the Warrant Selection: Correct score, I also calculated the Sum of Warrant Selection: Not Wrong score, in order to determine whether there was any difference in ability to avoid selecting incorrect warrants. I used the same procedure to screen for homogeneity of variances and normality. There was one outlier that did not represent a different population. The descriptive results are reported in Table 4.5. The means were nearly identical for the treatment and comparison treatment, though the comparison treatment was slightly higher. The mean difference $M = -0.31$, 95% CI [-2.74 to 2.12], $t(53) = -0.25$, $p = 0.80$ was not statistically significant at the Bonferroni-adjusted significance level of .005, indicating no difference in avoiding incorrect warrants, though the range was narrowed for the comparison treatment.
Table 4.5

*Sum of Warrant Selection - Not Wrong*

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>55</td>
<td>20.33</td>
<td>4.44</td>
<td>6-29</td>
</tr>
<tr>
<td>Treatment</td>
<td>25</td>
<td>20.16</td>
<td>5.00</td>
<td>6-27</td>
</tr>
<tr>
<td>Comparison Treatment</td>
<td>30</td>
<td>20.47</td>
<td>3.99</td>
<td>12-29</td>
</tr>
</tbody>
</table>

*Note: 2-3 possible responses per item for 12 items for a total of 30 possible points.*

_Warrant ranking task_. After participants completed the warrant selection task, they immediately completed a warrant ranking task that was part of the same google form. They were prompted to rank the warrants they selected in the warrant ranking task in order of effectiveness from most to least effective. Due to the instrument design, some students selected only one effective warrant, which meant that they did not rank their warrants in the same matter that the other students did. These students who selected one answer were given a point if the one warrant they selected was effective and no points if the one warrant they selected was ineffective.

I screened for normality by running descriptive statistics in SPSS and dividing the skewness and kurtosis statistics by the standard error of the skewness and kurtosis statistics, respectively, and determining that the number was not larger than the absolute value of 2. I screened for outliers for by converting the Sum of Warrant Ranking Scores to a z score and
determining students two or more standard deviations from the mean. Though there were two outliers, they did not represent a different population and they were retained in the analysis. The descriptive data are reported in Table 4.6.

There was homogeneity of variances for Sum of Warrant Ranking Scores, as assessed by Levine’s test for equality of variances \((p = 0.54)\). The mean scores for the treatment condition \((M = 7.12, SD = 1.30)\) and the comparison treatment \((M = 6.60, SD = 1.61)\) were determined. The difference in mean scores, \(M = 0.52, 95\% \text{ CI } [-0.28 \text{ to } 1.32], t(53) = 1.30, p = 0.20\), was not statistically significant. Because two total outcome measures were tested, a Bonferroni-adjusted significance level of .025 was calculated to account for the increased possibility of type I error. There is no significant difference between treatment condition and comparison treatment with relation to warrant ranking ability.

Table 4.6

Sum of Warrant Ranking Scores

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>55</td>
<td>6.84</td>
<td>1.49</td>
<td>3-11</td>
</tr>
<tr>
<td>Treatment</td>
<td>25</td>
<td>7.12</td>
<td>1.30</td>
<td>5-10</td>
</tr>
<tr>
<td>Comparison Treatment</td>
<td>30</td>
<td>6.60</td>
<td>1.61</td>
<td>3-11</td>
</tr>
</tbody>
</table>

*Note: one point possible per item across 12 items for a total of 12 possible points.*
The conclusions that can be drawn from the warrant ranking data are limited. Since some students had only selected one answer, they were not engaging in the same sort of task for the warrant ranking activity. In retrospect, they should have been recalibrated by being prompted to rank all the responses in order of effectiveness, rather than continue with the same items they had selected in the warrant selection task. However, if I limited this analysis to only students who never selected only one answer, I would be left with only 21 students, 11 in the comparison treatment and 10 in the treatment condition, which is not amenable to statistical analysis. However, the mean ranking scores for this smaller sample were very similar for the treatment \(M = 6.55\) and the comparison treatment \(M = 6.10\), in line with the larger data set. It is challenging to draw any conclusive interpretations based on this data set alone.

**Research question 1c.** Do students in the treatment write more effective warrants when given a claim and evidence than students in the comparison treatment?

**Warrant writing task.** Students were given a claim and evidence and directed to write a warrant for two different items as part of the Warrant Writing Task. First, the writing was coded for indicators of quality separate from historical thinking heuristics.

A participant response received a point for “Addresses Prompt” if it contained a statement that operated as a warrant for that specific claim. For instance, on the first item, the warrant needed to explain how the evidence supported the claim “The explosion of the U.S.S. Maine did not cause the U.S. to invade Cuba.” They were not given credit if their warrant did not provide a reason why the explosion of the U.S.S. Maine caused the U.S. to invade Cuba. Some responses did not receive a point because they were a summary. An example of a statement that was not given a point for “Addresses Prompt” follows: “This doc states that the explosion of the U.S.S. was an accident what [sic] was likely caused by a coal bunker fire that ignited the ships ammunition causing the explosion however the New York Journal states that the U.S.S. was purposely destroyed by Spanish mine [sic].” This statement does not explain how the evidence
supports the claim that the explosion of the U.S.S. Maine did not cause the U.S. to invade Cuba. Statements that included a warrant that supported the opposite side of the claim did not receive a point for “Addresses Prompt” either, because they were prompted to write a warrant linked to the claim that the Maine explosion did not cause the U.S. to invade.

A participant response received a point for “Accurate” if it included factually accurate information. If a student made a long statement that was accurate overall but contained one minor detail that was not precise or a misinterpretation, he or she still received a point for accurate. In the later coding for historical thinking, a student could still receive a point for historical thinking even if they had an inaccurate statement, if they used the historical heuristic correctly. An example of a student response that did not receive a point for accuracy follows: “The U.S. had no conflicts with neither [sic] country, they only had conflicts with themselves. Spain kept Cubans in concentration camps, leading up to Americans finding out. This caused Spain to destruct [sic] us.” This statement is inaccurate in saying the U.S. had no conflicts with either country (i.e. Cuba and Spain) and in the statement “this caused Spain to destruct us.”

A participant received a point for “Quote” if they contained a direct quote from a text, whether the quoted information was also contained in the evidence that was given in the prompt or information taken directly from the text. They also received credit for direct language from a text that was not put in quotation marks.

A participant received a point for “Reference Two Texts” by explicitly referencing two texts in their warrant. They could explicitly cite one text and reference another text explicitly without citing it and still receive a point for this category.

The mean scores for these coding categories were combined for the two items and are reported in Table 4.7. The means scores are nearly identical for treatment and comparison treatment in these categories. “The Addresses Prompt”, “Reference Two Text”, and “Quote” scores are slightly higher for the treatment condition, and the “Accurate” score is slightly higher
for the comparison treatment. The performance on having a quote and referencing two texts was particularly low for both groups.

Table 4.7

Combined Performance on Warrant Writing Items 1 and 2- Quality

<table>
<thead>
<tr>
<th></th>
<th>Total Possible</th>
<th>Group</th>
<th>N</th>
<th>M (SD)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address Prompt</td>
<td>2</td>
<td>Overall</td>
<td>53</td>
<td>1.51 (0.67)</td>
<td>0-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treatment</td>
<td>25</td>
<td>1.56 (0.65)</td>
<td>0-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
<td>28</td>
<td>1.46 (0.69)</td>
<td>0-2</td>
</tr>
<tr>
<td>Accurate</td>
<td>2</td>
<td>Overall</td>
<td>53</td>
<td>1.55 (0.67)</td>
<td>0-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treatment</td>
<td>25</td>
<td>1.48 (0.65)</td>
<td>0-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
<td>28</td>
<td>1.61 (0.69)</td>
<td>0-2</td>
</tr>
<tr>
<td>Reference 2 Texts</td>
<td>2</td>
<td>Overall</td>
<td>53</td>
<td>0.26 (0.45)</td>
<td>0-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treatment</td>
<td>25</td>
<td>0.32 (0.48)</td>
<td>0-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
<td>28</td>
<td>0.21 (0.42)</td>
<td>0-1</td>
</tr>
<tr>
<td>Quote</td>
<td>2</td>
<td>Overall</td>
<td>53</td>
<td>0.36 (0.65)</td>
<td>0-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treatment</td>
<td>25</td>
<td>0.44 (0.71)</td>
<td>0-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
<td>28</td>
<td>0.29 (0.60)</td>
<td>0-2</td>
</tr>
</tbody>
</table>

*Note. Treatment, n=25. Comparison, n=28. Includes combined performance scores of items 1 and 2 on quality factors other than historical thinking.*

The two Warrant Writing Task items were also coded for specific types of historical thinking (i.e. sourcing, corroborating and close reading). Students received a point for each instance of historical thinking, with a maximum possible total of two points per type of historical thinking.
thinking. The descriptive data is reported in Table 4.8. Historical thinking codes were totaled across both items and the mean scores for the treatment condition ($M = 2.20, SD = 1.29$) and comparison treatment ($M = 1.36, SD = 1.03$) were calculated. Though the mean difference, $M = 0.84$, 95% CI [0.20 to 1.48], $t(51) = 2.64, p = 0.01$ would have been statistically significant at 95% CI, a Bonferroni-adjusted significance level of .003 was set to account for increased possibility of type I error associated with 17 outcome measures. Therefore these results showed no statistically significant difference.

Since this test was not significant, a power analysis was conducted using the statistical software PASS. Group sample sizes of 75 and 75 achieve 80.02% power to reject the null hypothesis of equal means when the population mean difference is $\mu_1 - \mu_2 = 2.2 - 1.4 = 0.8$ with a standard deviation for both groups of 1.3 and with a significance level (alpha) of 0.003 using a two-sided two-sample equal-variance t-test.
Table 4.8

*Warrant Writing Task- Sum of Historical Thinking*

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>53</td>
<td>1.75</td>
<td>1.22</td>
<td>0-4</td>
</tr>
<tr>
<td>Treatment</td>
<td>25</td>
<td>2.20</td>
<td>1.29</td>
<td>0-4</td>
</tr>
<tr>
<td>Comparison Treatment</td>
<td>28</td>
<td>1.36</td>
<td>1.03</td>
<td>0-3</td>
</tr>
</tbody>
</table>

*Note: Two points possible per type of historical thinking across two items for a total of six possible points*

*Essay writing.* Students also completed a 30 minute essay writing task with the same documents used for the warrant selection task in order to determine whether or not the patterns in historical thinking observed from the warrant writing task persisted on a task where students needed to choose their own claim and evidence.

In order to analyze the data, students were divided into quartiles and analyzed based on writing quality and evidence of historical thinking. The two students closest to the mean of the quartile for both the treatment and comparison treatment were chosen for analysis. Overall, writing quality was very similar for the treatment and comparison treatment, with the treatment condition \((M = 3.75)\) averaging slightly higher than the comparison treatment \((M = 3.50)\). The performance on measures of quality was high regardless of quartile, with neither condition averaging lower than 3.50 of a possible 4 points. This may be due to the fact that engaging in the
warrant selection and ranking tasks prior to completing the essay task may have served as a scaffold to build knowledge of the possible warrants that could be used to reason about the explosion of the U.S.S. Maine. In addition, semi-structured interview data indicated students found it easier to generate their own argument than provide a warrant for a given claim and evidence. This could be due to the fact that they could select textual elements that they understood or that resonated the most strongly with them, and they did not have to comprehend warrants written by others or follow their reasoning.

In order to gain further insight into performance of students within the different quartiles, I analyzed their pretest scores as well. The students in the first quartile self-evaluated their ability at argument writing at a mean score of 3.75 on a scale of 1-5, with five being extremely effective and one being not at all effective, with a scores ranging from 3 to 4. Treatment condition students evaluated themselves at a mean score of 3.5, while the comparison treatment rated themselves at a mean score of 4.0.

Students in the second quartile self-evaluated their ability at argument writing at a mean score of 3.25, with scores ranging from 3 to 4. The treatment mean was 3.5, while the comparison treatment mean was 3.0. Third quartile students evaluated their argument writing at a mean score of 3.75, with scores ranging from 3 to 4. The treatment group was 3.5, while the comparison treatment was 4.0. Finally, fourth quartile evaluated their argument writing at a mean score of 3.25, with scores ranging from 2 to 4. The treatment mean was 3.0, while comparison treatment mean was 3.50. The pretest scores of this subset of students were very similar, though the comparison treatment evaluated their argument writing ability slightly higher in three of the four quartiles. It is notable that students in the comparison treatment evaluated their argument writing ability more highly than students in the treatment condition prior to the study, considering that there was evidence of the writing in the treatment condition showing higher mean levels of historical thinking in particular.
There were some interesting patterns present in historical thinking when analyzed by quartiles. As a reminder, the quartiles were generated based on the warrant selection scores, which were intended to reflect an ability to differentiate between more effective and less effective warrants. For instance, the mean evidence of historical thinking of the treatment condition in the first quartile (1.5) was higher than it was in quartile two (1.0) and equal to what it was in quartile four (1.5). Students in quartile 3 displayed clearly higher performance ($M = 2.5$). The evidence of historical thinking of the comparison treatment in quartile one ($M = 1.0$) was also slightly higher than the mean performance in quartile two (0.5), and quartile three (0.5), and one point higher than quartile four ($M = 0.0$). Of course, due to the small sample sizes, some of this difference could be random variation.

At first glance, it seems unusual that there was less overall evidence of historical thinking in higher quartiles for the comparison treatment. However, further analysis indicates a pattern in terms of the type of historical thinking displayed in each quartile. The treatment condition shifted from a reliance on sourcing to a greater incorporation of close reading and corroboration. This may be due to the fact that sourcing can be undertaken with a single document, whereas corroboration requires attending to more than one document, adding to the cognitive demand.

The fact that the comparison treatment did not show evidence of historical thinking in the higher quartiles may indicate that while they were able to identify effective warrants in the warrant selection task, they did not yet have ownership of these strategies to employ them in their own writing. The evidence of historical thinking shown in the lower quartiles was sourcing, so it could be evidence of students in the higher quartiles attempting to make more sophisticated arguments, but lacking the heuristics to do so in their own writing. For example, one of the comparison treatment students in the fourth quartile analyzed Document G, a speech given by Senate candidate Albert Beveridge. After quoting the senator, his warrant stated, “In this speech Bridger [sic] admitted that there another motive for the invasion of Cuba...they covered up there
real reason with the main [sic].” Though he is going beyond what is explicitly stated in the text to recognize other possible reasons for the invasion of Cuba other than the explosion of the U.S.S. Maine, he did not attempt to reason about the significance of the source of this document (a Senator seeking office), analyze the specific language the senator used, or corroborate what he said in his speech with other documents. Any of these approaches would have indicated the use of heuristics and strengthened his argument.

Another comparison treatment student analyzed Document E, an unsigned account of a “Reconcentration camp” forwarded via telegram by Fitzhugh Lee, U.S. Consul-General in Cuba. Again, after quoting the atrocities reported in the document, her warrant stated “This was a big factor to why the U.S invaded Cuba. The U.S. saw how terrible and inhuman these acts were, so they played this as a factor to invade Cuba”. Though she recognizes how the reconcentration camps may have been “played” as a reason to invade Cuba, she does not consider the questionable source of this document, analyze specific language used within the text, or corroborate what this document says with another sources indicating outrage at the atrocities in the camps, any of which would have made her warrant stronger. Again, this warrant indicates she is beginning to reason about the event, but not yet using historical thinking heuristics to do so.

The subtypes of thinking reflected in each warrant were also analyzed. In the first quartile, all four students showed evidence of sourcing. Three students (both students in the treatment condition and one in the comparison treatment), used Sourcing C (SoC), which deals with author’s perspective or role as a tool for interpretation of the document. The remaining comparison treatment student used Sourcing A (SoA), examining the source in terms of when it happened, the placement of that event in relation to other events, and how that impacted the interpretation. In addition, one treatment student showed evidence of Close Reading A (ClA), which is close reading of word choice, zooming on a particular word or phrase and its significance for the larger argument.
The example of close reading of word choice from a student in the treatment condition follows. After quoting the title of the New York Journal article, “Destruction of the warship Maine was the work of an enemy”, the student continued:

That title by itself makes people be interested. If the title was "We have an idea of who blew up the USS Maine" people wouldn't buy it or notice due to the fact that it isn't something that people say ooh I want to read that. This article makes people send letters to their congressman or have protests to send their message across [sic].”

In his pretest data this student stated that what makes a written argument effective is “one that shows a strong opinion on the matter and evidence on why they think their view point is correct then stating final conclusion to wrap up what they just said [sic].” In his essay he is going beyond opinion and why their viewpoint is correct and using close reading of the title to warrant a historical claim.

In the second quartile, only one student total showed evidence of sourcing. A student from the comparison treatment showed a warrant using Sourcing C (SoC), the author’s purpose or role. Both students in the treatment condition showed evidence of Close Reading A (ClA), which is close reading of word choice. For example, this student in the treatment condition quotes the word “intolerable” and then warrants his claim through the following statement:

The explosion of the U.S.S Maine lead the U.S to finally realize that things in Cuba weren’t gonna get any better unless they did something about it. The explosion of the Maine was the last straw before the U.S finally decided to invade cuba even though they have had it coming because of the things being held in cuba such as the concentration camps [sic].

In his pretest, this student indicated that what makes a written argument effective is “you can have facts and site documents so it makes your claim a lot more stronger [sic]. Here he is going beyond facts and citing documents to use close reading of the word “intolerable” to support
an argument that the explosion of the Maine was “the last straw” in a series of events leading to
the invasion of Cuba. This recognition of the multiple causality of historical events also reflects a
more sophisticated historical understanding. While this treatment condition student used close
reading effectively, neither student in the comparison treatment employed close reading, and none
of the four students used corroboration in their responses.

In the third quartile, three of the four students showed evidence of sourcing, in addition to
other historical thinking heuristics. Both students in the treatment condition showed evidence of
Sourcing A (SoA), the timing of the event. One of the two comparison treatment students
showed evidence of Sourcing C (SoC), which deals with author’s perspective or role as a tool for
interpretation of the document. Both treatment condition students showed evidence of Close
Reading A (ClA), or close reading of word choice, and one also showed evidence of Close
Reading B (ClB), or close reading of argument, which deals with the way the arguments are
constructed in terms of the claim the author made, evidence provided, etc. Note that the
treatment condition students only received a single point for this, though they showed two
different types of close reading. An example of Close Reading B from a treatment condition
student follows:

If this invasion was about the Maine being destroyed in an accident that was supposedly
caused by Cuba, then why is the President bringing up the fact that trade and business
may be involved? It seems to me like the 258 men that were killed in the accident doesn't
matter as much as the business between Cuba and the U.S.

On his pretest this particular student indicated what made a written argument effective
was “having a convincing reason as to why you wrote that particular argument.” This student
goes beyond showing a reason as to why they wrote the argument and employs close reading of
the evidence the President uses in his argument to analyze and challenge his assertion that Cuba
should be held responsible for the explosion of the Maine. This shows an ability to think
critically about claims made by a person with positional authority, and reflects the critical, questioning stance that is required for historical thinking. Finally, one student employed corroboration, namely Corroboration A (CoA), referring to other documents or knowledge from the document set that also said the same thing in order to strengthen the argument.

In the fourth quartile, neither student in the comparison treatment showed evidence of historical thinking. In the treatment condition, one student showed evidence of Sourcing A (SoA), the timing of the event. He used this heuristic with two different examples, but was only given one point to keep scoring consistent. One student showed evidence of close reading of word choice. In addition this student demonstrated use of Corroboration A (CoA), referring to other documents or knowledge from the document set that also said the same thing to strengthen his claim. After quoting vivid descriptions of the reconcentration camps from two documents including “bloodshed, starvation and horrible miseries” and “women and children thrown on the ground heaped in piles like animals”, the student continues:

This was used to make the people’s emotions riled up so they can invade Cuba without thinking it through. They wanted to rile aside from helping the people as the U.S. says they are also look how will they benefit from this and also to protect the people/industries from the U.S. [sic].

In his pretest, this student said that what makes a written argument effective is “research and being able to prove your argument is right over another person even if it is wrong” which shows a clear conflation between argumentation and persuasion, not surprising considering persuasion has been emphasized over true argumentation with the emphasis of the ACT on persuasive writing. This student does not see argumentation as a process of reasoning from evidence and generating new knowledge, he sees it more about “winning”, a commonly held misconception (Hillocks, 2011). However, his response does show the ability to read across documents, corroborate evidence found in common, and interpret the significance of vivid
language about the horrors of the reconcentration camps as intentionally done to manipulate the opinions of the U.S. public, which displays the ability to engage in reasoning that goes beyond persuasion and shows the ability to warrant a historical argument using heuristics.

**Research question group 2**: Student performance on items designed to reflect specific types of historical thinking and the relationship between skills related to historical argument writing.

**Research question 2a.** What is the relationship between students’ ability to select effective warrants, discriminate between more effective and less effective warrants, and write warrants demonstrating historical reasoning as measured by correlations and patterns in student responses?

**Correlation between warrant selection and warrant ranking.** In order to determine the relationship between students’ ability to select effective warrants and rank warrants in order of effectiveness, a Pearson’s product moment correlation was run to determine the value of the Pearson correlation coefficient. To test for normality I ran descriptive statistics in SPSS and divided skewness and kurtosis by their standard error. The result was not outside the absolute value of 2, so I continued with the analysis. Next I used the graphing function on SPSS to create a scatterplot with Sum of Warrant Selection on one axis and Sum of Warrant Ranking on another. The relationship was not curvilinear.

There is no significant relationship between Warrant Selection and Warrant Ranking scores. Using Cohen’s guidelines for strength of correlation (1988), there was a very small positive correlation between Sum of Warrant Selection and Sum of Warrant Ranking tasks \((r = .12)\). The coefficient of determination was calculated as the square of the correlation coefficient. The coefficient of determination was .014. Expressed as a percentage, performance on the warrant selection task explained only 1.4% of the variance in performance on the Warrant
Ranking task. This lack of a significant correlation affirmed that separate t-tests could be run rather than multivariate analysis to examine these two data sets.

**Correlation between warrant selection and warrant writing.** I also determined the correlation between the Sum of Warrant Selection and Sum of Warrant Writing tasks. After testing for normality I ran descriptive statistics in SPSS and divided skewness and kurtosis by their standard error. The result was not outside the absolute value of 2, so I continued with the analysis. Next I used the graphing function on SPSS to create a scatterplot with Sum of Warrant Selection on one axis and Sum of Warrant Ranking on another and the relationship was not curvilinear.

There is no significant relationship between Sum of Warrant Selection and Sum of Warrant Writing scores. There was a very small positive correlation between Sum of Warrant Selection and Sum of Historical Thinking in the Warrant Writing tasks, $r = .15$. The coefficient of determination was calculated as the square of the correlation coefficient. The coefficient of determination was .025. Expressed as a percentage, performance on the warrant selection task explained only 2.5% of the variance in performance on the Warrant Ranking task.

**Cut scores.** In order to better understand performance on the warrant selection and warrant ranking tasks, and trends within the two groups in performance on the two measures, I created cut scores (Dwyer, 1996). There are formal ways of setting cut scores, including having a panel of judges to set a criteria for performance by “consider(ing) characteristics of test takers, of test questions, of score profiles, or other aspects of performance… with or without prior knowledge of actual test scores obtained by past or present test takers or of the likely failure rate associated with different proposed cut scores” (Dwyer, 1996, p. 36). For instance, raters may estimate the probability a “minimally acceptable” test taker would answer each item individual item correctly, average the estimates for each test item, and use these to determine the cut score.
According to Dwyer, decisions judges make “rest, fundamentally, on their values and experiences” (p. 36). This led Glass (1978) to consider cut scores arbitrary.

The cut scores in this study do not have a robust empirical basis. Rather they are used as a tool for thinking about the performance of different groups of students. I began by setting the minimal criteria for proficiency by estimating reasonable overall performance of a “minimally acceptable” score on both the warrant selection and warrant ranking task, based on initial piloting of items, the complexity of the assessment tasks, and the lack of student familiarity with warrants. Each of the 12 warrant selection items consisted of five choices, of which two or three were effective. There were a total of 30 effective warrants. I categorized students by their performance on selecting effective warrants based on the percentage of effective warrants they selected from those 30 effective warrants, setting the percentage at 60%. The warrant ranking task presented a total of 12 opportunities for students to rank the more effective item in the primary position, looking at the first two items ranked. I categorized students based on their percentage of ranking the more effective warrant in the primary position on those 12 items, setting the percentage at 60%. I then divided scores into three categories starting from the 60% criteria: Advanced, Proficient, and Below Proficient. Advanced indicates superior performance. Proficient indicates demonstration of skills. Below Proficient indicates incomplete demonstration of skills.

The upper of end of the proficient and advanced band reflects categories similar to those often used in classroom grading (80%-100%= Advanced, 60-79%= Proficient, 0-59%= Below Proficient). When the number of items did not allow for a cut score percentage at exactly the cut score, I went to the next lowest possibility (e.g. 75% as a cut score for the warrant selection task). Due to the innovative nature of the assessments and the lack of precedent for how students could be expected to perform, I used these categories because it is similar to how students are categorized on classroom assessments and I wanted to and to reflect on performance as if the
assessment were scored like a traditional classroom assessment. The purpose was to compare performance of students in different ranges of performance, not to draw formal conclusions based on this analysis alone, and I acknowledge the inherently arbitrary nature of the cut scores.

*Cut score analysis for warrant selection task.* The findings for the cut score analysis are included in Table 4.9. I will begin by describing the treatment condition. In the treatment condition, 16% or 4 of the 25 total students were in the Advanced category for selecting effective warrants. 56% or 14 of the 25 total students were in the Proficient category for selecting effective warrants, 28% or 7 out of the 25 total students were in the Below Proficient category for selecting effective warrants. For the comparison treatment, 6.67% or 2 of 30 students were in the Advanced category. 26.67% or 8 out of the total 30 students were in the Proficient category, while 66.67% or 20 of 30 students were in the Below Proficient category. The majority of students were in the proficient category for the treatment condition, whereas the majority of students were in the Below Proficient category for the comparison treatment. Examining the data through the lens of the cut scores reveals what the mean score conceals; there is a larger middle group of proficient students in the treatment condition (56%) versus the 26% in the comparison treatment.
Table 4.9

*Performance by Cut Scores on Warrant Selection Task*

<table>
<thead>
<tr>
<th>Category</th>
<th>Items Correct</th>
<th>Percent Range</th>
<th>$n$ in Treatment</th>
<th>% of Treatment</th>
<th>$n$ in Comparison</th>
<th>% of Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced</td>
<td>24-30</td>
<td>80%-100%</td>
<td>4</td>
<td>16%</td>
<td>2</td>
<td>6.67%</td>
</tr>
<tr>
<td>Proficient</td>
<td>18-23</td>
<td>60%-77%</td>
<td>14</td>
<td>56%</td>
<td>8</td>
<td>26.67%</td>
</tr>
<tr>
<td>Below Prof.</td>
<td>0-17</td>
<td>0%-57%</td>
<td>7</td>
<td>28%</td>
<td>20</td>
<td>66.67%</td>
</tr>
</tbody>
</table>

*Note.* $n$ for Treatment = 25. $n$ for comparison treatment = 30.

Cut score analysis for warrant ranking task. I used the same process to create cut scores for the warrant ranking task, as depicted in Table 4.10. Neither the treatment nor the comparison treatment condition performed well on the warrant ranking task. In the treatment condition, 12% or only 3 of 25 were in the Advanced category. 48% or 12 of 25 were in the Proficient category. 40% or 10 out of 25 were in the Below Proficient category. In the comparison treatment, 10% or only 3 of 30 were in the Advanced category. 36.67% or 11 of 30 were in the Proficient category. 53.33% or 16 of 30 students were in the Below Proficient category.

In the treatment condition, no one category contained a majority of students, though the largest category was proficient at 48% of students. In the comparison treatment, the majority of students (53.33%) were in the Below Proficient category (versus 40% in the treatment condition). Like the warrant selection breakdown by cut scores, the warrant ranking data also indicated a
greater percentage of students in the proficient category for the treatment condition (48% versus 40%).

Finally I created profiles for each student based on their performance on the warrant selection and ranking tasks. Both the comparison treatment and the treatment had only one student that was in the advanced category in both warrant selection and warrant ranking. However, differences begin to emerge as I examined students at the proficient and below proficient levels as illustrated in Figure 4.1. In the treatment condition, there were seven students who were proficient in both warrant selection and ranking, compared to only two in the comparison treatment who were proficient in both categories. At the low end of performance, in

Table 4.10

**Performance by Cut Scores on Warrant Ranking Task**

<table>
<thead>
<tr>
<th>Category</th>
<th>Items Correct</th>
<th>Percent Range</th>
<th>Amount by Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Treatment</td>
</tr>
<tr>
<td>Advanced</td>
<td>9-12</td>
<td>75%-100%</td>
<td>Treatment</td>
</tr>
<tr>
<td>Proficient</td>
<td>7-8</td>
<td>58%-67%</td>
<td>Treatment</td>
</tr>
<tr>
<td>Below Prof.</td>
<td>0-6</td>
<td>0%-50%</td>
<td>Treatment</td>
</tr>
</tbody>
</table>

*Note. n for Treatment = 25. n for comparison treatment = 30.*
the treatment condition, there were only two students who were below proficient in both warrant selection and warrant ranking, compared to 10 students in the comparison treatment.

Figure 4.1. Warrant selection and ranking by category. WS = Warrant Selection; WR = Warrant Ranking. n for treatment = 25; n for comparison treatment = 30.

Cut score analysis for subset selecting multiple warrants. I also examined the subset of 21 students who answered more than one answer for all the warrant ranking tasks. The numbers and percentages in each category were not dramatically different from the larger data set.
indicating the process taken with including the students that only gave one response with the warrant ranking data was not significantly altering the data. This is illustrated in Table 4.11. On the warrant selection task, a narrow majority of students in the treatment group were categorized as proficient (54.55%), while half the students were in the comparison treatment group were proficient (50%).

Table 4.11

*Performance by Cut Scores on Warrant Selection- Subset of Students*

<table>
<thead>
<tr>
<th>Category</th>
<th>Items Correct</th>
<th>Percent Range</th>
<th>Amount by Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Treatment</td>
</tr>
<tr>
<td>Advanced</td>
<td>24-30</td>
<td>80%-100%</td>
<td>3</td>
</tr>
<tr>
<td>Proficient</td>
<td>18-23</td>
<td>60%-77%</td>
<td>6</td>
</tr>
<tr>
<td>Below Prof.</td>
<td>0-17</td>
<td>0%-57%</td>
<td>2</td>
</tr>
</tbody>
</table>

*Note. n for Treatment= 11. n for comparison treatment = 10. This table depicts the warrant selection scores for the subset of 21 students who selected more than one answer on the warrant selection task.*
In terms of the warrant ranking task, the percentages in the three categories of Advanced, Proficient and Below Proficient were similar for both treatment and comparison treatment conditions, never varying by more than 8 percentage points, as indicated in Table 4.12.

Table 4.12

*Performance by Cut Scores on Warrant Ranking Task for Subset of Students*

<table>
<thead>
<tr>
<th>Category</th>
<th>Items Correct</th>
<th>Percent Range</th>
<th>Amount by Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>n in</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Treatment</td>
</tr>
<tr>
<td>Advanced</td>
<td>9-12</td>
<td>75%-100%</td>
<td>1</td>
</tr>
<tr>
<td>Proficient</td>
<td>7-8</td>
<td>58%-67%</td>
<td>3</td>
</tr>
<tr>
<td>Below Prof.</td>
<td>0-6</td>
<td>0%-50%</td>
<td>7</td>
</tr>
</tbody>
</table>

*Note. n for Treatment= 25. n for comparison treatment = 30.*

Based on these findings, students in both groups are struggling in the warrant ranking category. The difference between the treatment and comparison treatment in the warrant selection task is obscured when one examines only the mean differences, whereas it is an accurate representation of the difference in the warrant ranking treatment. When students are categorized
by their performance on both warrant selection and ranking and compared, there are minimal differences apparent, as depicted in Figure 4.2.

Research question 2b. How do students perform on warrant selection items meant to reflect specific types of historical thinking? 2c. How do students perform on writing warrants given claims and evidence meant to stimulate specific types of historical thinking?
Item analysis of thinking types. The warrant selection task contained 60 individual answer choices (five per item on the 12 item assessment), each of which were meant to reflect a different type of historical thinking, namely sourcing, close reading, corroboration, or non-historical. The performance for the treatment and comparison treatment condition on the items linked to each category of thinking is detailed in Table 4.13.

Table 4.13

<table>
<thead>
<tr>
<th></th>
<th>Total Items</th>
<th>Group</th>
<th>N</th>
<th>M (SD)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sourcing</strong></td>
<td>16</td>
<td>Overall</td>
<td>55</td>
<td>10.00 (2.32)</td>
<td>4-15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treatment</td>
<td>25</td>
<td>10.32 (2.27)</td>
<td>4-15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
<td>30</td>
<td>9.73 (2.36)</td>
<td>6-15</td>
</tr>
<tr>
<td><strong>Close Reading</strong></td>
<td>16</td>
<td>Overall</td>
<td>55</td>
<td>10.16 (2.06)</td>
<td>5-14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treatment</td>
<td>25</td>
<td>10.60 (1.89)</td>
<td>7-14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
<td>30</td>
<td>9.80 (2.16)</td>
<td>5-14</td>
</tr>
<tr>
<td><strong>Corroboration</strong></td>
<td>16</td>
<td>Overall</td>
<td>55</td>
<td>10.13 (2.18)</td>
<td>5-14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treatment</td>
<td>25</td>
<td>10.68 (2.17)</td>
<td>7-14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
<td>30</td>
<td>9.67 (2.12)</td>
<td>5-14</td>
</tr>
<tr>
<td><strong>Non-Historical</strong></td>
<td>12</td>
<td>Overall</td>
<td>55</td>
<td>7.69 (1.95)</td>
<td>3-12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Treatment</td>
<td>25</td>
<td>7.84 (1.52)</td>
<td>5-11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
<td>30</td>
<td>7.57 (2.27)</td>
<td>3-12</td>
</tr>
</tbody>
</table>

*Note. n for Treatment =25. n for Comparison Treatment =30
As the table indicates, mean performance of the treatment group was greater for each type of historical thinking, though none reached the level of significance. The type of thinking that had the greatest mean difference was for items linked to corroboration ($M = 1.01$), followed by items linked to close reading ($M = 0.80$) and sourcing ($M = 0.59$).

Not surprisingly, performance on items linked to non-historical reasoning was most similar for treatment and comparison treatment ($M = 0.27$), because it was not a focus of the instruction that comprised the intervention. To determine whether or not the presence of items reflecting non-historical thinking diluted the overall difference in performance, the data was analyzed with these items removed. The combined mean difference for items linked to sourcing, close reading, and corroboration $M = 2.41$, 95% CI [-0.16 to 4.98], $t(53) = 1.88$, $p = 0.07$ was not statistically significant and did not approach a Bonferroni-adjusted significance level of .005 to account for the increased possibility of type I error associated with nine outcome measures. Therefore there is no significant difference between the treatment condition and the comparison treatment with relation to selecting items only linked to historical thinking.

**Research question 2c.** How do students perform on writing warrants given claims and evidence meant to stimulate specific types of historical thinking?

**Warrant writing task item 1.** The warrant writing items were designed to give students opportunities to engage in historical thinking, without limiting the potential heuristics students could use. The task could stimulate different sorts of historical thinking by the type of evidence it provided. Therefore, in order to interpret the warrant writing data, the prompt given needs to be considered. The prompt for the first warrant writing item is depicted in Figure 4.3.
Directions

There are two paragraphs from two different historical argument essays about the explosion of the U.S.S. Maine below. Each paragraph has a claim and evidence, but it is missing a warrant. Finish each paragraph by writing a warrant in the space below each paragraph that explains how the evidence supports the claim. Be sure to read the claim and evidence carefully. It does not matter whether or not you personally agree with the argument. Your job is to write the best warrant you can. You have the remainder of the class period.

3. The explosion of the U.S.S. Maine did not cause the U.S. to invade Cuba. In his State of the Union Address (Doc D), President McKinley stated there was “serious injury” to the trade and business of Americans who were either working in Cuba, or who had businesses there. He said “reckless destruction” of property on the island had been occurring as the “fight waged for years” between the Spanish and the Cuban rebels. McKinley argued this had left many of these Americans with their “property destroyed and themselves ruined” (Doc D).

Now write a warrant that explains how the evidence supports the claim.

This item cued students to use historical thinking. It targeted only one document, so it did not explicitly cue students to use corroboration, though two students did corroborate this information with an outside source that was not provided in the evidence. The prompt quoted specific words and phrases (e.g. “serious injury”; “reckless destruction”), so students may have chosen to include close reading of the significance of those words in their warrant. The fact that it was a publicly delivered speech by a U.S. President meant students had the opportunity to use sourcing, drawing on the strengths and limitations of a speech as a type of source, or the source as the President and his motivation in giving a public speech. There is also an opportunity for students to cue on the time of the event, the fact that the fight between Spanish and Cuban rebels had been going on for years before the explosion of the Maine.

Table 4.14 presents the scores for Sourcing, Close Reading, Corroboration, and Sum of Historical Thinking for the first item of the warrant writing task.
Table 4.14

*Performance on Warrant Writing Task by Type of Thinking - Item 1*

<table>
<thead>
<tr>
<th>Points Possible</th>
<th>Group</th>
<th>$M$ (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sourcing</td>
<td>Treatment</td>
<td>0.40 (0.50)</td>
</tr>
<tr>
<td></td>
<td>Comparison</td>
<td>0.32 (0.48)</td>
</tr>
<tr>
<td>Close Reading</td>
<td>Treatment</td>
<td>0.16 (0.37)</td>
</tr>
<tr>
<td></td>
<td>Comparison</td>
<td>0.18 (0.39)</td>
</tr>
<tr>
<td>Corroboration</td>
<td>Treatment</td>
<td>0.08 (0.28)</td>
</tr>
<tr>
<td></td>
<td>Comparison</td>
<td>0.00 (0.00)</td>
</tr>
<tr>
<td>Sum of Historical Thinking</td>
<td>Treatment</td>
<td>0.64 (0.57)</td>
</tr>
<tr>
<td></td>
<td>Comparison</td>
<td>0.50 (0.51)</td>
</tr>
</tbody>
</table>

*Note. For treatment, n=25. For Comparison Treatment, n=28*

This data indicates that sourcing was indeed used the most for this prompt, for both treatment and comparison treatment. A representative example of a response from a student in the treatment condition that shows sourcing of the time of event follows:

The explosion of the main did not cause the invasion of Cuba because the lives and property of US citizens were already being destroyed or threatened by the war between the Cuban rebels and Spain. Since one of the united states governments [sic] job is to protect its citizens i [sic] don't think they were just going to sit back and let it happen.

In this case, the student was sourcing based on the time of the event (in this case the explosion of the U.S.S. Maine). This student also uses the source of the document as reflecting
the position of the U.S. government, to reason about why they would be motivated to invade regardless of the explosion of the U.S.S. Maine.

Analysis of the mean performance indicates that the presence each type of historical thinking was relatively low. The treatment condition showed slightly more evidence of sourcing than the comparison treatment (0.40 versus 0.32 respectively). There was little evidence of close reading, and the presence of evidence of close reading was nearly identical for treatment and comparison (0.16 versus 0.18 respectively). Only two students showed any evidence of corroboration. Those were both in the treatment condition (0.08 versus 0.00 respectively). Two additional students in the treatment condition saw the several quotes and two parenthetical citations, and assumed that there was more than one source. They wrote warrants arguing that since “both sources agree” it was a reason to go to war. Looking at the combined evidence of historical thinking, there was a relatively was a minor amount of sourcing, close reading, or corroboration reflected in student warrants (0.64 vs. 0.50 respectively).

In addition, the length of the each warrant for warrant writing task item one was determined, as length can be considered one indicator of warrant quality when triangulated with other information such as evidence of historical thinking. The length of warrants was slightly higher for the treatment condition ($M = 45.32$) than the comparison treatment ($M = 43.50$) on the first task.

Warrant writing task item 2. On the second item, there were greater differences in terms of evidence of historical thinking between the treatment and comparison treatment groups. The prompt for the second warrant writing item is as follows:

The explosion of the U.S.S. Maine did not cause the U.S. to invade Cuba. The New York Times (Doc B) reported that Navy leader Captain Schuley felt the explosion was an “accident”, likely caused by a coal bunker fire that ignited the ship’s ammunition. However, the New York Journal (Doc A) stated that “all” naval officers thought the
Maine was “purposely” destroyed by a Spanish mine. It says the “brutal” Spaniards may have triggered the explosion while the men were sleeping (Doc A).

Now write a warrant that explains how the evidence supports the claim.

This prompt explicitly cues students to corroborate because it contains references to multiple sources. Specifically, it includes sources that are contradictory, so students need to reconcile this difference in order to make a warrant. Sources that vary widely in credibility and reflect different purposes for conveying content (i.e. the New York Times and the New York Journal) prompt students to use sourcing independently or in combination with corroboration. The inclusion of the phrase “Naval leader Captain Schuley” cues students to use sourcing in terms of the perspective of that particular person as a leader who would presumably be knowledgeable, but that must be reconciled with the later statement that “all naval officers thought the Maine was purposely destroyed by a Spanish mine. The inclusion of specific words in quotes (“accident”, “brutal”, “purposely”) encourages students to engage in close reading of word choice. Performance on Item 2 is depicted in Table 4.15.
Table 4.15

Performance on Warrant Writing Task by Type of Thinking - Item 2

<table>
<thead>
<tr>
<th></th>
<th>Highest Possible Total</th>
<th>Group</th>
<th>M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sourcing</td>
<td>1</td>
<td>Treatment</td>
<td>0.60 (0.50)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
<td>0.36 (0.49)</td>
</tr>
<tr>
<td>Close Reading</td>
<td>1</td>
<td>Treatment</td>
<td>0.48 (0.51)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
<td>0.18 (0.39)</td>
</tr>
<tr>
<td>Corroboration</td>
<td>1</td>
<td>Treatment</td>
<td>0.48 (0.51)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
<td>0.32 (0.48)</td>
</tr>
<tr>
<td>Sum of Historical Thinking</td>
<td>3</td>
<td>Treatment</td>
<td>1.56 (1.04)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
<td>0.86 (0.85)</td>
</tr>
</tbody>
</table>

Note. For treatment, n=25. For Comparison Treatment, N=28

As the mean scores indicate, the treatment condition showed higher mean scores of close reading (0.48 vs. 0.18) and sourcing (0.60 vs 0.36) than the treatment condition. Although the difference for close reading $M = 0.30$, 95% CI [0.05 - 0.55], $t(51) = 2.43$, $p = .019$ would have been statistically significant at 95% CI, a Bonferroni-adjusted significance level of .003 was set to account for increased possibility of type I error associated with 17 outcome measures. Therefore there was no significant difference in terms of evidence of close reading in warrant writing task item 2.

Both treatment and comparison treatment showed more evidence of corroboration on this item than they did on the previous item, though the difference between treatment and comparison treatment was small compared to the other types of historical thinking (0.48 vs. 0.32). The mean
sum of historical thinking across the items was also calculated (1.56 versus 0.86). Although the mean difference \( M = 0.84 \), 95\% CI [0.20 – 1.48], \( t(51) = 2.64, p = .011 \) would have been statistically significant at 95\% CI, it was not significant after a Bonferroni-adjusted significance level of .003 was set to account for increased possibility of type I error associated with 17 outcome measures. This meant there was no significant difference in overall use of historical thinking in warrant writing task 2.

There was some evidence of students combining different historical thinking heuristics in their responses. The following response from a student in the treatment condition gives a good example of a student that successfully combines the heuristics of corroboration, sourcing, and close reading in a single response.

While Doc A says it wasn't an accident, Doc B says it was an accident. The New York Times tells the truth about certain events that actually did happen. Unlike the New York Journal they always exaggerate and make up stories to make it more reliable. Doc B provides more evidence to show why because of a fire that could have occurred. Just because Doc A uses the word brutal doesn't mean the evidence back [sic] up the explosion of the U.S.S Maine did cause the U.S to invade Cuba.

The student corroborates and addresses contradictory information, and uses sourcing when she makes the decision to give more weight to the description of events in the New York Times. She also uses close reading, addressing the fact that the New York Journal uses strong language (“brutal”) but does not provide evidence to back up her position.

The length of warrants for warrant writing task item two was also determined. The mean number of words for the treatment (\( M = 53.84 \)) was just over nine words higher than the mean number of words for the comparison treatment (\( M = 44.71 \)). The students in the treatment condition wrote warrants for item two that were considerably longer than the comparison treatment, though this difference was not statistically significant.
Warrant writing task items 1 and 2 combined. The evidence of historical thinking was also totaled across the two items as indicated in the following table. Sourcing had the highest mean score, for both treatment and comparison treatment. There was some evidence of each type of historical thinking for treatment and comparison treatment across the board. As Table 4.16 indicates, both treatment and comparison treatment showed the lowest levels of corroboration compared to the other types of thinking, which is likely influenced by the fact that item 1 did not cue students to corroborate.
Table 4.16

*Performance on Warrant Writing Task by Type of Thinking - Items 1&2*

<table>
<thead>
<tr>
<th>Total Possible</th>
<th>Group</th>
<th>N</th>
<th>M (SD)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sourcing</td>
<td>Overall</td>
<td>53</td>
<td>0.83 (0.67)</td>
<td>0-2</td>
</tr>
<tr>
<td></td>
<td>Treatment</td>
<td>25</td>
<td>1.00 (0.71)</td>
<td>0-2</td>
</tr>
<tr>
<td></td>
<td>Comparison</td>
<td>28</td>
<td>0.68 (0.61)</td>
<td>0-2</td>
</tr>
<tr>
<td>Close Reading</td>
<td>Overall</td>
<td>53</td>
<td>0.49 (0.64)</td>
<td>0-2</td>
</tr>
<tr>
<td></td>
<td>Treatment</td>
<td>25</td>
<td>0.64 (0.70)</td>
<td>0-2</td>
</tr>
<tr>
<td></td>
<td>Comparison</td>
<td>28</td>
<td>0.36 (0.56)</td>
<td>0-2</td>
</tr>
<tr>
<td>Corroboration</td>
<td>Overall</td>
<td>53</td>
<td>0.43 (0.54)</td>
<td>0-2</td>
</tr>
<tr>
<td></td>
<td>Treatment</td>
<td>25</td>
<td>0.56 (0.58)</td>
<td>0-2</td>
</tr>
<tr>
<td></td>
<td>Comparison</td>
<td>28</td>
<td>0.32 (0.48)</td>
<td>0-1</td>
</tr>
<tr>
<td>Sum Hist Thinking</td>
<td>Overall</td>
<td>53</td>
<td>1.75 (1.22)</td>
<td>0-4</td>
</tr>
<tr>
<td></td>
<td>Treatment</td>
<td>25</td>
<td>2.20 (1.29)</td>
<td>0-4</td>
</tr>
<tr>
<td></td>
<td>Comparison</td>
<td>28</td>
<td>1.36 (1.03)</td>
<td>0-3</td>
</tr>
</tbody>
</table>

The combined length of warrants across items 1 and 2 was also determined. The mean length of the treatment condition \( (M = 99.16) \) was nearly eleven words longer than the mean length of the comparison treatment \( (M = 88.21) \), with most of this difference due to the difference
in item 2. Though the treatment condition wrote longer warrants, the difference did not reach the level of statistical significance.

**Essay writing task.** In addition to the warrant writing task, students completed an essay task. Students were given 30 minutes to complete the essay. The essay scores for both treatment and comparison treatment were divided into quartiles based on performance on the warrant selection task for the purpose of analysis. Four essays, two each for treatment and comparison treatment, were selected using the criteria of students whose scores were closest to the mean scores of each quartile, and students who had a complete data set. These essays were first divided into the basic elements of Toulmin’s argument pattern (claim, evidence, and warrant). Evidence of historical thinking present within each warrant (i.e. sourcing, corroboration, and close reading) was coded. A second pass of coding determined the subtypes of historical thinking within each type. I will describe the findings for each quartile in turn, beginning with the first quartile.

The first quartile (25th percentile) was a score of 34 (out of 60), with scores ranging from 25-34 and a mean score of 29.57. The overall quality was determined by combining scores for the same four indicators of writing quality analyzed in the warrant writing task, namely “Addresses Prompt”, “Accurate”, References 2+ Texts” and “Quote”. The mean scores on these indicators of quality were identical for the treatment ($M = 3.50$) and the comparison treatment ($M = 3.50$), out of a possible total of 4 points. The overall level of historical thinking was determined by coding the warrants for evidence of sourcing, corroboration, and close reading. The mean score for the treatment condition ($M = 1.5$) was slightly higher than that of the comparison treatment ($M = 1.0$), out of a possible 3 points.

In quartile one, all four students in the treatment and the comparison treatment showed evidence of sourcing. One student in the treatment condition and no students in the comparison treatment showed evidence of close reading. No students in either the treatment or the comparison treatment showed evidence of corroboration.
The second quartile (50th percentile) was a score of 38, with scores ranging from 25-34 and a mean score of 29.57. The overall quality was determined by combining scores for the same four indicators of writing quality analyzed in the warrant writing task, namely “Addresses Prompt”, “Accurate”, “References 2+ Texts” and “Quote”. The mean scores on these indicators of quality were again identical for the treatment ($M = 3.50$) and the comparison treatment ($M = 3.50$). The overall level of historical thinking was determined by coding the warrants for evidence of sourcing, corroboration, and close reading. The mean score for the treatment condition ($M = 1.00$) was slightly higher than that of the comparison treatment ($M = 0.50$).

In quartile two, no students in the treatment condition and one student in the comparison treatment showed evidence of sourcing. Both students in the treatment condition showed evidence of close reading, while neither of the comparison treatment students displayed close reading. No students in the treatment or comparison treatment conditions showed evidence of corroboration.

The third quartile (75th percentile) was a score of 42, with scores ranging from 39-42 and a mean score of 40.5. The overall quality was determined by combining scores for the same four indicators of writing quality analyzed in the warrant writing task, namely “Addresses Prompt”, “Accurate”, “References 2+ Texts” and “Quote”. The mean scores on these indicators of quality were slightly higher for the treatment condition ($M = 4.00$) than they were for the comparison treatment ($M = 3.50$). The overall level of historical thinking was determined by coding the warrants for evidence of sourcing, corroboration, and close reading. The mean score for the treatment condition ($M = 2.50$) was considerably higher than that of the comparison treatment ($M = 0.50$).

In quartile three, both students in the comparison treatment showed evidence of sourcing, as did one of the two students in the comparison treatment. Both of the students in the treatment condition showed evidence of close reading, while neither of the students in the comparison
treatment did. One of the two students in the treatment condition showed evidence of corroboration, though neither of the students in the comparison treatment did.

The fourth quartile contained scores ranging from 43-51, and the mean was 47.24. The overall quality was determined by combining scores for the same four indicators of writing quality analyzed in the warrant writing task, namely “Addresses Prompt”, “Accurate”, References 2+ Texts” and “Quote”. The mean scores on these indicators of quality were slightly higher for the treatment condition ($M = 4.00$) than they were for the comparison treatment ($M = 3.50$). The overall level of historical thinking was determined by coding the warrants for evidence of sourcing, corroboration, and close reading. The mean score for the treatment condition ($M = 1.50$) was again higher than that of the comparison treatment ($M = 0.00$).

In the fourth quartile, one student in the treatment condition and no students in the comparison treatment showed evidence of sourcing. One student in the treatment condition showed evidence of close reading; neither of the comparison treatment students did. One student in the treatment condition demonstrated corroboration, but neither of the students in the comparison treatment did so.

**Predicate and side analysis for warrant selection task.** In order to determine if “Myside bias” was a factor in student performance, students were asked what side they personally believed about the explosion of the U.S.S. Maine (Wolfe & Britt, 2008). If they agreed with the statement, they were assigned a point. If they disagreed, they were given a zero. For the Predicate and Side “Did Cause”, a bit more than half ($M = 0.61$) of students in the treatment condition believed that the explosion of the U.S.S. Maine caused the U.S. to invade Cuba, as displayed in Table 4.17. Slightly over half of students in the comparison treatment ($M = 0.53$) believed that the explosion of the U.S.S. Maine caused the U.S. to invade Cuba. For the predicate “Should have caused”, treatment and comparison treatment responded nearly identically. For the treatment condition, just over one-third ($M = 0.39$) of the students in the treatment condition believed that the
explosion of the U.S.S. Maine should have caused the U.S. to invade Cuba, which was nearly identical to the comparison treatment ($M = 0.38$).

Table 4.17

Warrant Selection Task Performance by Side and Predicate

<table>
<thead>
<tr>
<th>Items (Points Possible)</th>
<th>Group</th>
<th>$M (SD)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Did Cause</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes, Did Cause</td>
<td>3 (15)</td>
<td>Treatment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.10 (2.18)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.00 (2.13)</td>
</tr>
<tr>
<td>No, Did Not Cause</td>
<td>3 (15)</td>
<td>Treatment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10.14 (2.05)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.58 (2.50)</td>
</tr>
<tr>
<td>Total for Did/Did Not Cause</td>
<td>6 (30)</td>
<td>Treatment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
</tr>
<tr>
<td><strong>Should Have Caused</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes, Should Have Caused</td>
<td>3 (15)</td>
<td>Treatment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
</tr>
<tr>
<td>No, Should Not Have Caused</td>
<td>3 (15)</td>
<td>Treatment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
</tr>
<tr>
<td>Total for Should/Should Not Have Caused</td>
<td>6 (30)</td>
<td>Treatment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Comparison</td>
</tr>
</tbody>
</table>

Note. For treatment, $n=25$. For Comparison Treatment, $N=28$
When examined by predicate and side, there is not a great difference between treatment and comparison treatment. What stands out here is the relatively low performance of both treatment and comparison treatment on items linked to the predicate and side “Should Have Caused” the U.S. to invade Cuba, a predicate that also had the lowest mean levels of agreement ($M = 0.38$ for the treatment condition and $M = 0.39$ for the comparison treatment).

3. Research question group 3: The thinking of students as they complete the assessment tasks and what they report learning from the treatment.

**Research question 3a.** What sort of thinking does a subset of students engage in as they complete the warrant selection, warrant ranking, and warrant writing tasks?

**Think aloud.** Five students engaged in the think aloud tasks. They were Ivan, Jose, Ethan and Sonia (pseudonyms). They were identified by the cooperating teacher Ms. Jones as representing a range of reading and writing ability levels. Their performance on the ACT-PLAN assessment, which they took as sophomores and which was the most recent nationally normed assessment data available, reflects this range as well, as shown in Table 4.18. The following table indicates their scores in comparison to the benchmark score considered on track to be ready to complete college-level work.
Table 4.1

Performance on ACT-PLAN Reading and English versus Benchmark

<table>
<thead>
<tr>
<th>Student</th>
<th>ACT-PLAN Reading (+/− Benchmark)</th>
<th>ACT-PLAN English (+/− Benchmark)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ivan</td>
<td>13 (−5)</td>
<td>13 (−2)</td>
</tr>
<tr>
<td>Jose</td>
<td>15 (−3)</td>
<td>10 (−5)</td>
</tr>
<tr>
<td>Ethan</td>
<td>15 (−3)</td>
<td>16 (+1)</td>
</tr>
<tr>
<td>Sonia</td>
<td>19 (+1)</td>
<td>18 (+3)</td>
</tr>
</tbody>
</table>

Note. ACT-PLAN Reading Benchmark = 18, ACT-PLAN English Benchmark = 15. Benchmark indicates on track to be ready to do college-level work by graduation.

I also examined the pretest data for these four students. Students were asked what makes a written argument effective and the parts of an effective written argument. Then they were asked to self-evaluate their ability at argument writing, and how they felt about argument writing compared to the other things they studied. They were also asked to write down everything they knew about the explosion of the U.S.S. Maine and/or the Spanish American War, though none of the students had significant prior knowledge of these topics. I will list the responses in ascending order of performance on the PLAN assessment.
Ivan, who was five points below the ACT-PLAN Reading Benchmark and two below the Writing Benchmark, said “if you have enough evidence” is what makes a written argument effective. He listed the parts of an effective argument as “claim, evidence, reasoning and conclusion”. Ivan self-evaluated his ability at argument writing as a 5, which seems out of line with how he performed on the ACT-PLAN, but only assigned a 1 to indicate how he felt about argument writing compared to other things he studied.

Jose, who was three points below the ACT-PLAN Reading Benchmark and five below the Writing Benchmark, indicated that what made a written argument effective was “Solid evidence (and) a lot of counter or pre-meditated arguments”. He indicated that the parts of an effective argument were “claim, evidence (and) reasoning”. He self-evaluated his ability at argument writing as a 3 and scored how he felt about argument writing compared to other things he studied as a 3 as well.

Ethan, who was three points below the ACT-PLAN Reading Benchmark, but one above the Writing Benchmark, said that what makes a written argument effective is “using facts, statistics, rhetorical devices, and appealing to logic, emotion, and ethics”. He listed the parts of an effective argument as “1. the intro, 2. claim, 3. research, 4. reasoning (and) 5. Close.” He self-evaluated his ability at argument writing as a 4, and assigned a 4 to indicate how he felt about argument writing compared to the other things he studied.

Finally, Sonia, who was one point above the ACT-PLAN Reading Benchmark and three points above the Writing Benchmark said “examples used” make a written argument effective. She said the parts of an effective argument were “claim, evidence (and) reasoning”. She self-evaluated her ability at argument writing at a 3, and assigned a 2 to how she felt about argument writing compared to other things she studied in school.
Think alouds were conducted in order to determine the type of thinking that students were engaged in during the warrant selection think aloud. The prompt used for the think aloud is depicted in Figure 4.4.

![Image](image.png)

*Figure 4.4. Warrant Selection Task for Think Aloud*
Responses were parsed into comments, or bursts of speech that followed a student reading or interacting with a text, and events, or idea units that represented a unique idea on a particular topic (Wolfe & Goldman, 2005; Manderino, 2011). Relevant idea units were coded with the type of thinking (i.e. sourcing, close reading, or corroboration). I established interrater reliability with the same professor of literacy who is a former AP History teacher. We coded one student together, and then coded a second separately, reaching consensus on the codes we disagreed upon. For the think aloud of the warrant selection task, we chose a sample of 10% of the codes. In this 10% sample of the warrant selection task think aloud, there were a total of 28 idea units, of which we disagreed and reached consensus on two, for an agreement of 92.86 percent. For the 10% sample of the think aloud of the warrant writing task, there were a total of 24 idea units, of which we disagreed and came to consensus on two, for an agreement of 91.67 percent.

Think aloud for warrant selection task. On the think aloud for the warrant selection task, there was a wide variation in the amount of heuristic use by student, in the types of heuristics they used, and in the effectiveness of their heuristic use. This information is depicted in Table 4.19. For instance, sourcing ranged from a low of 0% of the time to a high of 29.4% of the time. Corroboration ranged from a low of 0% to a high of 28.57% of the time. Close reading was a bit more consistent, ranging from a low of 5.88% other to a high of 25.93%. It was also the most consistently effective. Even the least effective of the four students was effective 66.67% of the time.
Table 4.19

*Historical Thinking in Warrant Selection Task Think Aloud*

<table>
<thead>
<tr>
<th>Student</th>
<th>Total Idea Units</th>
<th>HT Idea Units</th>
<th>Type of HT</th>
<th>n</th>
<th>% of Total</th>
<th>% Effective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ivan</td>
<td>28</td>
<td>11</td>
<td>Sourcing</td>
<td>0</td>
<td>0%</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Close Reading</td>
<td>3</td>
<td>10.71%</td>
<td>66.67%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Corroboration</td>
<td>8</td>
<td>28.57%</td>
<td>25.00%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>11</td>
<td>39.29%</td>
<td>36.36%</td>
</tr>
<tr>
<td>Jose</td>
<td>17</td>
<td>6</td>
<td>Sourcing</td>
<td>5</td>
<td>29.41%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Close Reading</td>
<td>1</td>
<td>5.88%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Corroboration</td>
<td>0</td>
<td>0%</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>6</td>
<td>35.29%</td>
<td>16.67%</td>
</tr>
<tr>
<td>Ethan</td>
<td>27</td>
<td>13</td>
<td>Sourcing</td>
<td>4</td>
<td>14.81%</td>
<td>75.00%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Close Reading</td>
<td>7</td>
<td>25.93%</td>
<td>75.00%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Corroboration</td>
<td>2</td>
<td>7.41%</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>13</td>
<td>22.22%</td>
<td>46.15%</td>
</tr>
<tr>
<td>Sonia</td>
<td>16</td>
<td>9</td>
<td>Sourcing</td>
<td>4</td>
<td>25.00%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Close Reading</td>
<td>2</td>
<td>12.50%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Corroboration</td>
<td>3</td>
<td>18.75%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>9</td>
<td>56.25%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Think aloud for warrant writing task. The think aloud for the warrant writing task occurred directly after the think aloud for the warrant selection task. The prompt for the warrant writing task think aloud follows:

The explosion of the Maine did not cause the U.S. invasion of Cuba. For example, the telegram sent by Fitzhugh Lee, the U.S. Consul-General in Cuba (Doc F), recommended that war ships should be stationed near Cuba. These ships should be “ready to move” on short notice to Cuba, to protect Americans in Cuba and their interests. He sent this telegram months before the U.S.S. Maine ever exploded.

Now write a warrant that explains how the evidence supports the claim.

Again, this data is much more meaningful if one considers the way that the claim and evidence cue certain types of historical thinking as a lens to interpret performance. In this instance, the prompt cued sourcing, in particular paying attention to the time of the event in relation to other events. It also cited a specific phrase “ready to move” that cued an opportunity for close reading. Finally, the source of the document added an opportunity for further analysis because students may recognize that this telegram, requesting ships “ready to move” as needed, was sent several months before the explosion of the U.S.S. Maine.

The heuristic use on the warrant writing task was quite different, as shown in Table 4.20. Sourcing was clearly the most often used heuristic overall, ranging from Ivan, who used sourcing for 31.82% of the total idea units, to Jose, who showed less historical thinking overall but did use sourcing 11.76% of the time. The pattern that was the clearest was that students were using...
warrants effectively in their warrant writing think alouds. Fourteen out of 15 idea units were effective. This is in sharp contrast to the warrant ranking task, where only Sonia showed consistently effective use of historical thinking.
Table 4.20

*Historical Thinking in Warrant Writing Task Think Aloud*

<table>
<thead>
<tr>
<th>Student</th>
<th>Total Idea Units</th>
<th>HT Idea Units</th>
<th>Type of HT</th>
<th>n</th>
<th>% of Total</th>
<th>% Effective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ivan</td>
<td>22</td>
<td>9</td>
<td>Sourcing</td>
<td>7</td>
<td>31.82%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Close Reading</td>
<td>2</td>
<td>9.09%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Corroboration</td>
<td>0</td>
<td>0%</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>9</td>
<td>40.91%</td>
<td>100%</td>
</tr>
<tr>
<td>Jose</td>
<td>17</td>
<td>4</td>
<td>Sourcing</td>
<td>2</td>
<td>11.76%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Close Reading</td>
<td>1</td>
<td>5.88%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Corroboration</td>
<td>1</td>
<td>5.88%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>4</td>
<td>23.53%</td>
<td>100%</td>
</tr>
<tr>
<td>Ethan</td>
<td>14</td>
<td>6</td>
<td>Sourcing</td>
<td>4</td>
<td>28.57%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Close Reading</td>
<td>2</td>
<td>14.29%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Corroboration</td>
<td>0</td>
<td>n/a</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>6</td>
<td>42.86%</td>
<td>100%</td>
</tr>
<tr>
<td>Sonia</td>
<td>15</td>
<td>5</td>
<td>Sourcing</td>
<td>3</td>
<td>20.00%</td>
<td>66.67%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Close Reading</td>
<td>2</td>
<td>13.33%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Corroboration</td>
<td>0</td>
<td>n/a</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>5</td>
<td>33.33%</td>
<td>80%</td>
</tr>
</tbody>
</table>
Research Question 3b. How do these students feel the intervention is different from typical classroom instruction and what do they report they have learned?

Semi-structured interview. The semi-structured interview consisted of 10 total items. Items 1 and 2 and 6-9 were coded for historical thinking using the same codes used to develop the instrument design because there was potential for these to include discussion of historical thinking heuristics. However, it is important to note that students were not prompted to discuss historical thinking heuristics, and it was possible to answer these items without mentioning historical thinking heuristics.

Responses were parsed into comments, which were bursts of speech that followed a student reading or interacting with a text, and events, or idea units that represented a unique idea on a particular topic, often reflecting a reading or thinking process (Manderino, 2011; Wolfe & Goldman, 2005). Relevant idea units were coded with the type of thinking (i.e. sourcing, close reading, or corroboration).

I established interrater reliability with the same professor of literacy who is a former AP History teacher. We coded one of the students together, then compared our codes and reached consensus. We then coded one additional student separately. In total there were 33 idea units across the two students and we were in agreement on 29 of them, or 88% of a 50% sample of participants.

Historical thinking in the semi-structured interview. The evidence of historical thinking in the semi-structured interview is presented in Table 4.21. All students showed some evidence
of historical thinking, with a high of 47% of codes for Ivan, and a low of 11% of codes for Jose. Every student showed evidence of sourcing, and three of four students showed evidence of corroboration. None of the students showed evidence of close reading skills in their semi-structured interview responses.
Table 4.21

*Historical Thinking in Semi-structured Interview*

<table>
<thead>
<tr>
<th>Student</th>
<th>Total Idea Units</th>
<th>HT Idea Units</th>
<th>Type of HT</th>
<th>n</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ivan</td>
<td>15</td>
<td>7</td>
<td>Sourcing</td>
<td>4</td>
<td>26.67%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Close Reading</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Corroboration</td>
<td>3</td>
<td>20.00%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>7</td>
<td>46.67%</td>
</tr>
<tr>
<td>Jose</td>
<td>18</td>
<td>2</td>
<td>Sourcing</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Close Reading</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Corroboration</td>
<td>2</td>
<td>11.11%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>2</td>
<td>11.11%</td>
</tr>
<tr>
<td>Ethan</td>
<td>22</td>
<td>5</td>
<td>Sourcing</td>
<td>3</td>
<td>13.64%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Close Reading</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Corroboration</td>
<td>2</td>
<td>9.09%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>5</td>
<td>22.73%</td>
</tr>
<tr>
<td>Sonia</td>
<td>17</td>
<td>2</td>
<td>Sourcing</td>
<td>1</td>
<td>5.88%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Close Reading</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Corroboration</td>
<td>1</td>
<td>5.88%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>2</td>
<td>11.76%</td>
</tr>
</tbody>
</table>

*Note. Includes coding of responses to interview questions 1 & 2, 6-9. Note. HT = Historical Thinking. n = number of idea units for each type of historical thinking (Sourcing, Close Reading, Corroboration and Total). % total= % of total number of idea units. % effective is the percentage of idea units for that type of thinking that were coded as effective.*
What would make this work easier. As part of the semi-structured interview, students were asked what would make this work easier for them. Though there was no clear consensus across respondents, there were some common answers given. Ivan and Sonia said spending more time learning about warrants would be helpful, for a total of three idea units. Ivan and Jose said additional practice with selecting effective warrants would have been helpful, for a total of two idea units. In addition, Jose also said more feedback from the instructor and more opportunities to share his work with his peers would have been valuable. His response included one idea unit about each of these items. Ethan shared that he felt the assessment would have been better if the warrant selection task had only one correct answer. He had a total of three idea units on this topic.

Despite the lack of clear trends in responses, the reasons students gave for their responses were illuminating. For instance, in advocating for more time and more practice, Ivan said, “I think more time to practice would be really good because um like at first when (we learned about argument writing in class), like when I saw the paragraph with reasoning and evidence and how to find it and all that, [sic] after a full semester I think I got good at that just with time”. He felt that a week was not enough time to feel fully confident with using historical thinking as warrants for his argument.

Ethan targeted the design of the assessment as something that could be changed to make the work easier.

Just having one answer, instead of having multiple answers, just picking one. Because I know a lot of people are going to pick all of them, and then they are going to be stuck on the order of importance. So then just having that one answer, and having that be the most important. And that might be more difficult for other people, but for me it would have been easier because at least I know, ‘Okay then these warrants aren’t important...’ [sic].
Ethan’s response helps to contextualize the trend in the data that students in the treatment group were better at selecting right answers, but there was no difference between treatment and comparison treatment in their ability to avoid selecting wrong answers.

Ivan also made a point about the design of the task. He pointed out that having students write their own claim and evidence would make it easier.

I think it would be better if it was like the full paragraph from the person, cause then it would be their own viewpoint, and not someone else’s. It makes it easier when you write your paragraph at the end, with just racking [sic] the same documents they give you just so you know the knowledge … of what happens, and that really helps.

He felt that an initial step that would have been more helpful for him would have been for them to write their own full paragraph rather than just provide a warrant for a given claim and evidence (this student did not participate in the essay task since he was part of the think aloud).

*What students learned this week versus typical instruction.* Students were asked two separate questions, what they learned this week, and how it was similar or different from typical instruction classroom instruction. Their responses often combined both elements into a single response. Though these particular responses were coded for evidence of sourcing, close reading, and corroboration, delving deeper into their responses was helpful to understand both their perception of what new information they learned and how the instruction differed from typical classroom practice.

*Ivan’s response.* While Ivan shared that he had learned about sourcing previously, he had not learned about how sourcing could be used to warrant an argument.

I learned how to use sourcing, like how to connect the evidence and that could make a strong argument, just from sourcing. I didn’t know that before. I mean I knew that, just didn’t use it. We always, we like we used the evidence to explain how the evidence connected to the claim, now we used sourcing, as forwarding your claim [sic].
... So we used sourcing to see if it was reliable or not, but now we’re like, we’re seeing “So this is reliable, so because this person was there, and when was this written, and how far or how close it was, would that change their viewpoint. So it’s always put it just like the sourcing, but then you could use the… I would like to say, “Hey this guy, he was there, he survived, he would know a little more about what happened there, than someone who was like running around ten years later, based on like information they got offline or something like that [sic].”

Ivan went on to articulate how he understood the difference between writing a warrant and “reasoning”, which is the phrase their teacher introduced earlier in the year when they studied argument writing:

Just like the warrant and the reasoning, I thought they would be similar, which they are, just um, I thought it would be like reasoning, connect evidence to the claim, but it actually uses a little bit more. I think it’s pretty cool we can use like sourcing as part of our connecting to evidence and claim [sic].

Along with learning how sourcing could serve as a warrant for his claim, he also referenced corroboration in a similar manner, though he did not refer to it as corroboration. He said he learned “a new way to write warrants, like comparing two documents together to prove something, like that was really a cool thing I learned.” These two examples show how Ivan was beginning to understand how historical thinking heuristics could be used to warrant arguments. While he had exposure to historical thinking skills such as sourcing before, he didn’t realize that these skills could help him warrant his claims.

Jose’s response. On the other hand, Jose did not emphasize historical thinking in his responses to questions about what he learned. He showed relatively low levels of historical thinking in his interview overall, though 11% of his idea units did reflect corroboration. Though he said he learned about warrants and gained “better writing and better thinking skills” as a result
of the instruction, his responses referenced structure more often than historical thinking. He appeared to recognize warrants as a valuable structural element of writing. He stated, “I’ve learned not just to write better but to listen and take into consideration specific details about writing arguments, and structuring my writing, my writing and reading levels to a better level.” However, he did not discuss how historical thinking contained within warrants was a part of what made arguments effective.

*Ethan’s response.* Compared to Jose, Ethan showed a deeper level of engagement with historical thinking heuristics as warrants, and even began to connect their usefulness to his own life. Ethan pointed out that the treatment condition included many more documents than the one or two at a time they typically worked with in history class. The first heuristic that Ethan emphasized in his response was sourcing, specifically when a particular document was written.

Well, I’ve learned that looking at dates is very helpful, I’ve found that out.

Because a lot of these journals, a lot of this stuff was written right after the Maine exploded, and then some were [*sic*] written 10, 15 years after the Maine exploded. And they can do an aspect on, “Okay, how credible is this.” Usually when we write we don’t look at the dates as much.

Along with finding the value in sourcing, Ethan found our work corroborating contradictory information as particularly useful. “I also learned how, how two different stories can be uh misconstrued but still trying to tell the same story, and that can be taken in everyday life too. Like off topic but the stuff that happened in Ferguson, (Missouri), like I learned how you can look at two different sides of a story, and get really good information from each side and make a good argument [*sic*].” Whereas his comments about sourcing focus on historical reading to determine credibility of documents, in discussing corroboration Ethan explicitly connects the process of corroboration to making an argument. He also is able to provide a contemporary context about where this type of thinking could be useful.
Sonia’s response. Of the four students, Sonia seemed to find the treatment least helpful for her learning. In the response to how the instruction was similar or different from what they normally do in class, she said that she felt the term “warrant” confusing.

I’m not really sure. Like the warrants thing confused me a little because we’ve been using reasoning from the beginning of the school year. So being used to that and then having a new term (was) kind of, well confusing, how to use it, when to use it or what it is, in the beginning was kind of confusing [sic].

She also felt that the work we did as part of the treatment, being given a claim and evidence and being asked to warrant it with historical thinking, felt too easy to her compared to generating her own claim.

It was somewhat easier, but it didn’t feel like I was doing much, like it was coming from me, so it was copy and paste. What we did here is kind of different because we usually write our own claims, well like the whole week we usually have our own claims, our own evidence, and then our own like reasoning and concluding sentence, so everything was there for us and we had to use what we were given, and like put it in order and explain our way of putting it, though not our warrants [sic].

Whereas Ivan and Ethan could point to how our discussion of warrants helped crystallize for them how heuristics could be used to further a claim, and Jose felt his writing improved and found warrants helpful as a structural element of writing, Sonia felt like she was being given a different name for something she felt she’d been doing all along. This could well be the case, as Sonia was already above the benchmark in writing prior to the study, though the ACT does not emphasize warrants in their task, and the task is persuasive writing, not writing about text.

Follow up question. At the end of my interview with Ethan, I asked him to elaborate on the reference he made to the events in Ferguson, Missouri. Since including his response would have inflated the amount of historical thinking by deviating from the structure I followed with
other students, I did not include his response in my totals. I include his response here as an
illustration of how one student was able to apply the thinking he learned in the intervention to his
own life.

Yeah like um, you need to look at both opinions of that and the facts of that and take
what you need from those, you know what I mean? Well, the doc, yeah, the documents
of the different, like the New York Journal says *(flipping back to story)*. “Destruction of
the Maine was the work of an enemy”. And then um, and it says “Assistant security
Roosevelt convicted of [sic] - convinced the explosion of the Maine was NOT an
accident, and all this other stuff…
The New York Times (and this was the same year… same day too!). It said “Maine’s
hull will decide. Divers will inspect the ship’s hull to find out whether the explosion was
from the outside or inside.”

So it kind of shows like okay (the U.S. Journal) they’re already set on “this was not
an accident, this was on purpose.” The New York Times is more like, “Okay it might
have been an accident, we don’t have enough information.”

And then the… So it kinda, in seeing that it kind of brought to light, okay everyone has
their own opinion, like even professional journals and newspapers have their own opinion
on… “We, we already know it’s not an accident”, and then the other people are like,
“Well it might have been an accident.” You know, and that brought that to light.

And you can, you can use it in everyday life, because going back to that (Ferguson)
example, you can use it in everyday life because if you want the real facts, you’re going
to have to look through different opinions to get the real facts. And uh, link together
which- which facts in here are true, and see if they match with the facts in the other
newspaper. Cause if they have the same facts, that means it’s not opinion, it’s not
opinion [sic].
While his understanding of the interpretive nature of history is still evolving, Ethan’s response shows that he not only understands corroboration and can explain it in his own language; he also is beginning to understand how this skill can apply to his own life.
V. DISCUSSION AND IMPLICATIONS

The purpose of this study was to determine the characteristics of an effective intervention to promote the use of historical thinking heuristics with high school juniors, in order to develop a local theory of teaching historical argumentative writing to adolescents that can inform work in similar contexts. I compared two promising approaches, one focusing on explicit instruction in warrants, and one focusing on finding and evaluating evidence to support different claims and sides. Because warrants are notoriously difficult to study, being reliant on their claim and evidence partners, several measures were used to help control for differences in performance due to strength of student claims or evidence use or their overall writing ability. My hypothesis was that the historical thinking heuristics treatment would help students better identify effective warrants and better discriminate between more effective and less effective warrants than the evidence-focused treatment.

Overview

The previous chapter included the results of data analysis for the central measures in the study (a pretest, a warrant selection task, a warrant ranking task, a warrant writing task, and an essay writing task, in addition to a semi-structured interview and a think aloud task. In this chapter I will begin with a statement of the overall effect from the most reliable study measures. I will then give a summary and discussion of the findings that may have contributed to this overall effect, organized by the abilities measured in the study (e.g. warrant selection, ranking, and writing). Then I will state the limitations of the study and use these findings to develop the characteristics of an effective intervention and a theory of teaching historical argumentative writing to adolescents that can inform work in similar contexts.
**MANOVA of most reliable study measures.** There was a statistically significant overall difference between treatment and comparison treatment groups on the dependent variables of Warrant Selection: Correct and Warrant Writing: Combined. Follow up univariate ANOVAs showed that there was a statistically significant difference in Warrant Selection: Correct scores between students from different groups, even after a Bonferroni corrected alpha level for two dependent variables. This indicates that there was a significant treatment effect present for the most reliable measures in the study, lending credence (along with the consistently higher mean scores across multiple measures for the treatment condition) to these findings as evidence that the treatment worked at the local level, and can be used to inform the development of a theory of argument writing that will inform instruction in similar contexts.

**Warrant selection.** A follow up ANOVA indicated that there was a statistically significant difference in students in the warrant writing group’s ability to select correct warrants in comparison to the evidence group, even after a Bonferroni correction for two dependent variables. The floor of the range was higher for the treatment condition, and the range of scores was narrower for the treatment as well. However, though the mean scores for overall warrant selection were higher for the treatment condition, the difference after a Bonferroni adjustment was not large enough to remove the possibility of this being due to chance.

The warrant selection task, which included 2-3 valid responses and foils reflecting common errors in historical thinking, approximated the probabilistic nature of historical reasoning better than traditional assessments. Though the mean scores of the treatment group were significantly higher for selecting correct warrants, they were slightly lower for avoiding ineffective warrants (though this difference was not statistically significant). This may indicate students need explicit instruction not only in historical reasoning and writing effective warrants (the focus of this study), but also in discriminating between more and less effective warrants.
Mean performance of the treatment group on the type of historical thinking tagged to each item was greater for each type of historical thinking on the warrant selection task, though none reached the level of statistical significance. Both the treatment and comparison treatment included reading of multiple texts as a part of their instruction. The fact that the heuristic condition students outperformed the evidence condition students (who also read multiple texts) is consistent with Nokes, Dole & Hacker (2007), who found reading multiple texts alone did not boost comprehension unless it was combined with historical heuristic use.

The difference between treatment and comparison treatment was relatively evenly distributed across items tagged to different historical thinking heuristics, though historical thinking typically does not progress evenly (Lee & Ashby, 2000). The mean difference between treatment and comparison treatment may have been less pronounced for sourcing because students had more exposure to sourcing prior to the start of the study. Items linked to corroboration had the greatest level of mean performance, followed by items linked to close reading and sourcing. Corroboration is an important skill because it helps students realize the interpretive nature of history by making multiple perspectives clear and encouraging the formation of an intertext model (Monte-Sano, 2011). Recent interventions emphasizing reading have shown an impact on other heuristics, but not corroboration (Reisman, 2012). Therefore this finding is interesting, though the possibility of it being due to chance variation cannot be ruled out.

To get a better picture of how students used heuristics as they completed the warrant selection task, I used think aloud tasks with a subset of four students. The total number of idea units for each type of historical thinking and their mean scores were calculated. There was a wide variation in the amount of heuristic or close reading use by student, in the types of thinking they used, and in the effectiveness of their thinking. For instance, sourcing ranged from a low of use 0% of the time by Ivan to a high of use 29.41% of the time by Jose. Corroboration ranged from a
low of use 0% of the time by Jose to a high of use 28.57% of the time by Ivan. Close reading was a bit more consistent, ranging from a low of use in 5.88% of idea units by Jose to a high of use in 25.93% of idea units by Ethan. Close reading was also the most consistently effective heuristic across students. Though Ivan’s use of close reading was the least effective overall, he was still effective 66.67% of the time. This was in contrast to the semi-structured interview, where there was no evidence of close reading whatsoever. Since students did not have texts in from of them to refer to during the interview, they would have had to reference specific words or phrases from memory, which would have been challenging.

A semi-structured interview including these same four students was used to learn about how this intervention was different from typical instruction, along with information gathered from the teacher at the beginning of the study. Students reported that instruction typically dealt with one document, or at most at most two, whereas this intervention built toward multiple documents at the same time. Ethan spoke extensively about how helpful he found the examination of the New York Times and the New York Journal’s conflicting articles written immediately after the explosion of the Maine. Seeing these two newspapers giving completely different accounts of the same event helped him realize the power of corroboration to warrant an argument. Though his understanding of history as reflected in the interview was naïve- he felt corroboration would help him find the “real facts” and determine what was “true” and “not opinion” - he took an important step in understanding the role of the heuristic of corroboration in constructing historical arguments.

I also examined the performance on warrant selection items linked to particular predicates and sides. Both treatment and comparison treatment struggled the most to select effective warrants linked to the predicate and side “Yes, the U.S. should have invaded Cuba” following the explosion of the U.S.S. Maine. Students were asked whether or not they personally believed the different claims and sides. Only 39% of students in the treatment
condition and 38% of students in the comparison treatment believed the U.S. should have invaded Cuba. This may indicate that students find it more difficult to warrant claims different from their own beliefs (Wolfe, Britt & Butler, 2009).

Think alouds for the warrant selection task indicated that some students recognized the historical heuristic use that was reflected in a foil and assumed that the item was correct, regardless of whether the reasoning was valid. This may help explain the nearly identical mean score for “Warrant Selection: Not wrong” for treatment and comparison treatment, and the wider range for the treatment condition (scores ranging from 6 to 27) than the comparison treatment (scores ranging from 12 to 29), as contrasted with the statistically significant higher performance of the treatment condition on the Warrant Selection: Correct measure. Because students were just being introduced to these heuristics functioning as warrants, they were not used to tasks requiring them to consider, for example, the degree of conclusions drawn based on limited evidence as a lens for interpreting the effectiveness of a warrant. The short length of this study prevented me from going in depth on issues of degree in evaluating warrants.

Students reported that the way they were assessed was different from their traditional assessments. Ethan argued that having to choose more than one warrant on the warrant selection task was confusing for him and would likely be for many of his classmates as well. He argued that it would have been better to select a single warrant. However, the traditional “one correct answer” format does not mirror the probabilistic reasoning and questioning stance that historians need to take when examining documents. Ethan’s response may have been evidence of a less developed understanding of what it means to think historically. Though students would be taking the online PARCC assessment later that school year, which includes many multiple select and evidence-based selected response items, according to Ethan they were not yet comfortable with this format.

**Warrant ranking.** There was no statistically significant difference between mean student performance on the warrant ranking task. Though the mean performance was a bit higher
and the range (5-10 versus 3-11 for the comparison treatment) was a bit narrower, both groups struggled. Writing effective warrants was emphasized over making nuanced distinctions between warrants. The nature of the foils on the items (e.g. warrants that were accurate, but the degree was too extreme, or that were accurate but irrelevant, or in which the interpretations drawn from the use of one heuristic conflicted with another) meant that students could not take a warrant at face value, much as historians do not take historical accounts at face value. In addition, the warrant ranking task did not recalibrate students prior to ranking the items. Students only ranked the items they selected, and more than half the students selected only one warrant as effective on or more items. This seriously limited any conclusions that could be drawn from this data alone.

**Cut scores.** I created cut scores to uncover patterns in student performance when looking at both warrant selection and warrant ranking. Without removing the students who only selected one answer choice, the treatment condition had a greater number of students who were proficient in both warrant selection and warrant ranking, and a much smaller number of students who were below proficient in both categories. This may indicate that while the treatment did not build students’ historical thinking enough to reach the advanced category, it was effective in moving students out of the below proficient category.

This pattern weakened when examining the smaller subset of 21 students who selected more than one answer for warrant ranking. The treatment condition did not have greater numbers of students who were proficient in both warrant selection and warrant ranking, though there remained fewer students in the below proficient categories. The sample size of this group may have been too small to see any measurable difference, or there could have been an underlying characteristic of students who select one warrant versus students who select more than one, such as a less developed epistemology of history as having one right answer, that led to less variation.

**Warrant writing.** When scores for historical thinking and quality were added into a Warrant Selection: Combined and included with Warrant Selection: Correct Scores, the overall
effect was significant according to a MANOVA analysis. Follow up testing indicated that the warrant writing task alone did not reach the level of significance, though it did contribute to the overall effect.

There was no statistically significant difference in measures of quality or historical thinking written by students in the treatment condition and the comparison treatment after statistical corrections for multiple statistical tests on the same data set. Though students in the treatment condition had higher overall mean levels of historical thinking than the comparison treatment, the data did not eliminate the possibility of it being a chance finding.

While reading and writing require some of the same cognitive skills, they also differ considerably (Shanahan, 1984). There is a still a marked difference between the skills required to read and evaluate warrants and the ability to write one’s own. Lack of a significant relationship between the ability to select effective warrants and write effective warrants may have been impacted by the complexity of the warrant selection task. Based on the pretest data, how students feel about argument writing and evaluating their ability to write arguments is moderately related to their ability to state the parts of an argument, and how effective students report being at argument writing is strongly related to how they feel about argument writing compared to other subjects they study in school.

There was a moderate relationship between the historical thinking students showed in their arguments and the pretest “Parts of an argument” measure, $r = .29$. In addition, there was a strong relationship between historical thinking and the quality of student responses. This is not surprising as the quality measures (i.e. addressing the prompt, accuracy, including a direct quote, and including explicit references to two or more texts) may enable historical thinking.

The presence of each type of historical thinking was relatively low on item one for both treatment and comparison treatment. One possible reason could be due to confusion around the claim. Some students wrote a warrant arguing that the US would not invade Cuba at all, because
it would put American citizens at risk. However the claim they were supposed to warrant was, “The explosion of the U.S.S. Maine did not cause the U.S. to invade Cuba”. They were not asked to about whether the U.S. would invade, so students did not get credit for their responses. This may be related to the fact that even college students have difficulty recalling the main predicate of simple, 2-clause arguments such as “should intervene” or “is right to intervene” accurately (Britt et al, 2008). While skilled readers recall predicates more accurately than less-skilled readers, it is an issue for both groups, and not attending to predicates makes it difficult to attend to warrants (Britt et al., 2008). The length of warrants was similar for treatment and comparison treatment in item one, with the treatment condition being just under two words longer than the comparison treatment.

Sourcing being the heuristic used the most often overall in students’ argument writing was not surprising, given that sourcing may become a habitual practice with students such as these who regularly write about text as part of Document Based Questions (Reisman, 2012). The average student in the treatment condition used sourcing at least once across the two warrant writing items. Close reading was the second-most exhibited form of thinking in students’ argument writing. Close reading is also a relatively discrete, tangible action students can take, often emphasizing specific words in a single text (Reisman, 2012). Sourcing and close reading are related in the sense that knowing about the source allows students to interpret the choices that particular author made in language. The nature of the task, which forced students to write a warrant and cued them with particular arrangements of textual evidence, may have artificially inflated the percentage of students with heuristic use present in their warrants for comparison treatment students, as opposed to just prompting students to write an argument.

The students in the treatment condition had a higher mean corroboration score. Unlike sourcing or close reading, Corroboration (and contextualization, which this study did not address), may be considered more sophisticated strategies because they involve thinking about
multiple texts (Reisman, 2012) and may be related to forming an intertext model of the relationships across texts (Monte-Sano, 2011). The mean corroboration scores for the warrant selection data had the largest difference between treatment and comparison treatment, although it was not statistically significant. The evidence of corroboration was most apparent on item 2, when the stimulus material telegraphed corroboration by presenting multiple agreeing and conflicting texts in the evidence students were given. Only a handful of students from the treatment condition attempted to use corroboration in the first item which did not telegraph corroboration and only two used it effectively. The responses for item two were also over nine words longer on average for the treatment than the comparison treatment, indicating students were writing more extensive and potentially higher-quality warrants.

Though I attempted to telegraph close reading on both warrant writing items by quoting specific words and phrases that showed evidence of tone or could be read rhetorically, only the second item showed evidence of close reading. Perhaps the presence of the conflicting sources made students more aware of the documents as authored, and cued them to think more critically about the words and phrases that the author chose. It may be also be possible that for the first item, students used sourcing and stopped there, thinking interpreting the source was sufficient to warrant their argument. They may not have seen the need to evaluate language use as part of their warrant.

The think aloud data gave some additional insight into the use of heuristics. Though there was a range in types of thinking, sourcing was clearly the most often used heuristic, ranging from a high of use in 31.82% of idea units by Ivan to a low of use in 11.76% percent of idea units by Jose. However, a range of historical thinking was demonstrated. The strongest pattern was that students were using historical thinking warrants effectively in their writing. 14 of the 15 total idea units linked to historical thinking were effective, which was much more consistently effective than the thinking they displayed in the warrant selection and ranking think aloud.
There are several possible reasons why warrant use was particularly effective on the warrant writing task. When students generate their own warrants, they can simply determine how the textual evidence helps support the claim, whereas on the warrant selection task, they have to interpret and evaluate someone else’s reasoning about how the texts are connected. In addition, the stimulus material is different on both items. Though more items would be needed to draw firm conclusions, it is interesting that this percentage of effectiveness for students in the think aloud of the writing task was greater than on the overall writing task for the remainder of the students. It may be that there was something about the process of explaining ones thinking through a think aloud that supports the generation of effective warrants, and this could be an area for additional research.

This study provided the claim and evidence for students on the warrant writing task in order to control for these in analyzing student warrants, because the quality of warrants are often dependent on the clarity of the claim and the suitability of evidence selected. However it could be considered more challenging to write a warrant for a claim and evidence that one did not choose because when students are writing their own, they can stick to parts of the text that they understood or found most compelling. In his semi-structured interview, Ivan argued that he would have found it more helpful to generate his own arguments. Ivan’s ACT-PLAN Reading score is significantly below the ACT-PLAN benchmark indicating being on track for college. That score may indicate that his struggles as a reader make warranting an instructor-provided claim and evidence particularly challenging for him. The treatment condition’s mean performance on other measures of quality was slightly higher in the areas of “Addresses the prompt”, “Cites 2 or more texts”, and “Contains a quote”, but was slightly lower in terms of “Accuracy”. This may be related Voss and Wiley’s (1999) finding that while more complex inferential tasks may help students develop a situation model, recall-focused tasks may lead to higher scores on factual recall tasks because students are devoting their cognitive energy to recall. The comparison
treatment group spent considerable time finding evidence from text to support claims they were
given, which meant they were focusing on what was stated in the text. The treatment condition
spent little time finding evidence in text, and more time warranting claims using historical
thinking heuristics and close reading. In addition, less accurate statements in the treatment
condition may be related to the nascent stages of knowledge transformation, where students are
beginning to go beyond reporting what the text says and bringing their own reasoning to bear on
what they have read (Scardamalia & Bereiter, 1987).

Both groups scored low on the measures of including a quote and referencing two texts.
Though these are reasonable measures of quality in student writing, quotes and (in the case of the
second response) multiple texts were already provided in the prompt, which contained the claim
and evidence. The warrant statement did not necessarily need a direct quote as that evidence was
already established earlier in the argument. Students were more likely to use heuristics like
sourcing or close reading that focused on the source or specific language from a single text, and
these heuristics do not require reading multiple texts (Reisman, 2012).

Despite the complexities of interpreting the warrant writing data, when these Warrant
Writing: Quality and Warrant Writing: Historical Thinking measures were added to form the
more reliable Warrant Writing: Combined variable and included with the Warrant
Selection: Correct variable, they contributed to an overall significant effect. Though follow up
ANOVA indicated the difference for writing alone was not significant, it is sufficient impact to
influence the building of theory.

Student ability to write warrants was also assessed on a time-bound essay writing task
with the same documents to determine whether or not the patterns in historical thinking observed
from the warrant writing task continued on a task where students needed to choose their own
claim and evidence. To aid in the finding of patterns, students were divided into quartiles based
on performance on the Warrant Selection task.
While the amount of historical thinking did not increase significantly for the higher quartiles, there was evidence of a shift to more sophisticated types of historical thinking in the treatment condition, with students showing greater levels of close reading and corroboration. Students in the lower quartiles relied almost entirely on sourcing when they used historical thinking heuristics. These students were used to being prompted to identify the source as part of the Document Based Questions they answered in history class, so sourcing may be somewhat of a routine behavior (Reisman, 2012). However, according to Ms. Jones the reasoning they did after identifying the source was limited.

This is an indication that the treatment may have an effect on student writing even when the claim, side and evidence are not provided, and that the use of the warrant selection assessment score to form quartiles may be differentiating between students with more and less argument knowledge or skill with argument writing. For instance, a treatment student in the third quartile used Sourcing A (SoA), or sourcing of when an event took place, to indicate “even before (the explosion of the Maine) took place there were many conflicts”. He stated that the U.S. was already “looking for a way to go to war”. He used Close Reading A (ClA), or close reading of specific words or phrases, to point out that the president states “The reasons to go to war”. This combination of sourcing and close reading indicates that he is thinking deeply about both the source and the language used.

When asked about their argument writing ability on the pretest, students in the comparison treatment happened to evaluate their argument writing ability more highly than the treatment condition, thought this difference was not statistically significant. Students were also asked in the pretest an open response question about what makes a written argument effective. When looking at the treatment and comparison treatment in combination, there was a qualitative difference between the responses of students in the first and fourth quartiles, though there was great variation in student responses. For instance, one student in the first quartile said, “a strong
opening sentence and a strong claim makes a written argument effective”, while another said “by getting the listeners attention by tone of voice. Having eye contact and using a good speed.” In contrast, one typical response from a fourth quartile student read, “having a claim, evidence, reasoning and a conclusion.”

**Additional example of transfer.** I will share an extended example from Ethan during the think aloud, which was not included in the main analysis of the think-aloud data set because it came after the protocol. In the semi-structured interview, Ethan shared how helpful he found the instruction on using the corroboration heuristic to examine the New York Times and the New York Journal’s conflicting articles written immediately after the explosion of the Maine. Seeing these two newspapers giving completely different accounts of the same event helped him realize the power of corroboration to warrant an argument. As we were wrapping up our interview, Ethan elaborated about how he found the use of corroboration helpful in his own life. He explained that he and his father were discussing the news coverage of the shooting of Michael Brown in Ferguson, Missouri together. Ethan pointed out how CNN and Fox News covered the Michael Brown shooting differently. He said CNN’s coverage treated Michael Brown as a “victim”, while Fox News portrayed him as a “troublemaker”. Ethan made a parallel to our study of the U.S.S. Maine. He said he described to his father the connection he saw to how the New York Times and the New York Journal gave conflicting accounts about what caused the Maine explosion. Ethan said that based on what he’d learned from the study, he assumed neither CNN nor Fox News was telling the “full story”, but felt that he needed to watch both so he could figure out what really happened. Though his understanding of history as reflected in the interview was naïve- he felt corroboration would help him find the “real facts” and determine what was “true” and “not opinion” - he took an important step in understanding the role of the heuristic of corroboration in constructing historical arguments. This is a clear account of how studying
heuristics can help students make sense of the conflicting media reports and complex issues they encounter in their daily lives.

**Limitations**

This study had several limitations. First of all, though a diverse setting was chosen intentionally, this was a sample of convenience. Students were randomly assigned to condition, but they were the students of one teacher at one school, and the assignment happened within the boundaries of students’ existing class periods. When the most reliable measures were combined, there was a statistically significant overall effect, and a follow-up ANOVA showed that the Warrant Selection: Correct score was statistically significant. I ran several t-tests to explore what else may have contributed to this difference, though the small sample sizes made it difficult to achieve significant results. However, consistently higher mean scores across measures are promising enough to demonstrate that the design works in a local context and can be used to generate theory to inform teaching in similar contexts.

The short duration of the study was another limitation. The fact that there was an overall significant effect from the most reliable measures and t-tests indicated students in the treatment condition exhibited higher (though not statistically significant) mean scores across individual measures does not necessarily indicate that this finding would be consistent or lasting if reassessed regularly. A power analysis gave insight into the additional numbers of students that would be necessary in order to yield a significant effect on the t-tests. Ethan self-reported one instance of transferring the use of the corroboration heuristic to viewing media reports of the shooting of Michael Brown in Ferguson, Missouri and the essay writing task indicated that students were able to transfer the use of historical thinking heuristics to a task when not given claim, side and evidence. It showed that some comparison treatment students were able to use historical thinking heuristics as warrants as well, though higher quartile students appeared to use
less-sophisticated warrants than treatment condition students. There is also no way to know if students exhibited these behaviors when they read documents outside the context of the study.

Design-based research typically includes several cycles of developing, prototyping, teaching, assessing, and analyzing data as part of an iterative process to refine theory, and the short length of this study is a limitation in that regard. However, if the desired effect instructional effect is achieved, early iterations of the process can be used to define the characteristics of an effective intervention and generate theory. In this instance, there are positive indications that the treatment impacted student ability to select correct warrants reflecting different types of historical thinking, and led to higher mean performance on measures of warrant writing. These findings will inform a line of research that continues with design-based research at the middle school level to further refine a theory of how the ability to write arguments warranted by historical thinking can be developed in adolescents.

The focus on specific types of historical thinking, namely sourcing, corroboration and close reading, limited the scope of my findings. These areas were targeted because they were the focus of the treatment, so they made the most sense as outcome measures. However these do not capture the full dimension of historical thinking. Historical reasoning is multidimensional. The focus on sourcing, corroboration and close reading could have obscured other effects that the intervention had on historical reasoning that did not fit in these categories (Monte-Sano & De La Paz, 2010).

There were also limitations due to the innovative nature of the measures in this study. There were not existing validated measures to draw upon. Therefore, I emphasized construct validity, ensuring the items reflected the underlying historical thinking heuristics. Though the instruments were piloted with a handful of students in a similar context, and designed with feedback from historical experts, they were not examined statistically for reliability or internal
consistency. Intercorrelations indicated some relationship between key variables, though the strength of the relationship varied and was low for some variables.

While it provided interesting data about the student performance in selecting different types of historical warrants, the design of the warrant selection task complicated the interpretation of the data. Items designed to reflect certain types of historical thinking such as sourcing included foils that did reflect use of sourcing, but went beyond the scope of what was warranted by the evidence. Students who selected these foils had their answers marked as incorrect, but it was impossible to know whether or not their selection was due to a misunderstanding of sourcing itself or the issue of degree, because the other possible answer choices reflected different types of thinking. Though the assessment was designed so that it would be possible to study all foils tied to sourcing and look for patterns, the limited number of items made it difficult to draw concrete conclusions.

These issues are not limited to this instrument. The ACT “Reading Between the Lines” study (Ferguson, 2006) showed that there was no difference in student performance between items intended to reflect different thinking skills (e.g. supporting details, main idea, sequential, comparison contrast and cause and effect relationships, and generalizations and conclusions). Rather, differences in student performance appeared to be related to the complexity of the texts that students read as a part of the task. Because of the multifaceted nature of reading behavior, it is impossible to say in any reliable way that the reason students missed a particular item was due to the linked skill it was designed to reflect or a myriad of other factors, such as academic vocabulary, reading fluency or student motivation (Shanahan, 2014). If anything, the design of my instrument was valuable in that it reflected the full scope of a particular heuristic. I determined it was most worthwhile to think about the assessment more globally as reflective of students’ historical thinking as they interacted with a complex text. For this reason, all of the items tagged to different types of historical thinking were combined to create one warrant
selection and one “Warrant Selection: Correct” score, to increase the reliability of these measures. I now think about the differences in thinking among warrant types as formative information to triangulate with evidence from other data sources, such as student writing.

The conclusions that could be drawn from the warrant ranking task were limited by the unfortunate design decision to have students rank whatever warrants they selected as effective, rather than recalibrating students and having them rank all five warrants. Since several students selected only one warrant as effective on at least one item, the pool of students who completed the task as intended was too small to draw firm conclusions. However, the fact that there was no statistically significant difference between treatment and comparison treatment, and the fact that only the selection of correct warrants was statistically stronger and not the avoidance of incorrect warrants, indicates that the treatment did not impact student ability to differentiate between degrees of warrant effectiveness, something that was not explicitly taught in the study.

In addition, there were issues of measurement validity in the warrant writing task due to the small number of items. The study was initially designed to have eight items, but the scope of the measurement had to be reduced due to unforeseen circumstances involving student absences. Because there were only two items, there is not sufficient data to know if the variations in student performance by type of historical thinking are related to consistent patterns in student thinking, or are unique to that item. For that reason, when doing my MANOVA analysis I combined the scores for historical thinking and warrant quality across the two items to increase the reliability of the warrant writing task findings. The pattern of higher mean scores on this more reliable combined measure and the more sophisticated heuristic use at higher quartiles on the essay writing task may indicate that there is a difference beyond any idiosyncrasies in instrument design.

The impact of prior instruction also limits the conclusions that can be drawn as a result of this study. The fact that students had some exposure to historical thinking heuristics may have
diluted the impact of the treatment. Though their work was often limited to single documents and was focused on sourcing, Ms. Jones did previously teach some lessons on historical thinking. There is no way to know whether or not the affect would have been greater with a population of students who have not been exposed to historical thinking heuristics. In addition, Ms. Jones had been teaching argument writing to students prior to the study. The pretest finding that 57.8% of the students could identify the parts of an argument (though I did accept related terms such as “opinion” for claim, “facts” for evidence, and “reasoning” or “elaboration” for warrant), indicated that the majority of students had prior knowledge of a writing component functionally similar to a warrant. Class discussion and follow up questioning of Ms. Jones indicated this reasoning tended toward general rules or elaboration and not the discipline-specific warrants called for in history (Bruner, 1960). Nonetheless, prior instruction in argument writing may have influenced student performance and may have resulted in higher comparison treatment scores than if the study had been done with students with less exposure to argument writing.

While the typical instruction at Northwest High did not include using heuristics as warrants, it also did not reflect traditional textbook-based history instruction. The comparison treatment had some similarity to typical instruction at that school, with students reading historical texts and gathering evidence to support claims, with the exception being that students were using multiple texts and finding the best evidence to support different sides of a claim. Though teachers had not received professional development in teaching argument writing according to Ms. Jones, the fact that they did not emphasize the textbook and had eliminated multiple choice assessments from their history curriculum reflects a more progressive approach to instruction and assessment.

Three of the four students in the semi-structured interview clearly stated that they learned something from the treatment. Sonia was the only student that said she only learned a limited amount. Sonia said the instruction felt like “copy and paste” to her. Since Sonia is above the ACT-PLAN benchmark in reading, she may have found the process of warranting a claim easier
than students like Ivan, who had the lowest summative reading performance of the four students. As a more skilled reader and writer, Sonia may have had an easier time comprehending the claim and evidence and determining the relationship between them. The act of isolating and studying different historical thinking heuristics separately each day may have felt limiting to her, which may indicates the need for differentiation based on student literacy levels. Differentiation was not a part of this study.

Sonia also found the term “warrant” confusing, and said she was distracted by the fact that there was a different word now for what she felt she had been referring to as “reasoning” since the beginning of the year. However, since she demonstrated effective use of warrants that displayed historical thinking in her semi-structured interview, it could be that the confusion she felt did not interfere with her ability to do the tasks, or it could be that she was already able to do this work prior to the study.

**Implications for Research and Instruction**

Despite its limitations, this study has several implications for further research and classroom instruction. When the most reliable measures were combined, there was a statistically significant improvement in the combined measures of selecting correct warrants and writing effective warrants. When t-tests were used to analyze performance on additional measures, mean performance was consistently higher for the treatment condition focusing on warrants, though this difference was not statistically significant. The short duration of the study, the small sample size, and the Bonferroni correction due to multiple statistical tests on the same data set made significant results hard to attain. However, there are many promising trends that should be explored further in future research.

**Implications for research: Assessment design and analysis.** Due to the innovative nature of the instruments in this study, it is not surprising that refinements will need to be made. Assessments of historical thinking are notoriously difficult to design and require multiple
iterations in order to ensure they are historically accurate, that they have a design that effectively targets the underlying constructs, and that they provide data that is valuable for teachers (Breakstone, 2014). In order to draw stronger conclusions that are applicable to a larger population, this study should be replicated with several changes. There should be a larger sample size to lend more statistical power to the results of the study. According to power analyses, even with statistical corrections, adding approximately 35-60 students to each condition (depending on the measure) would yield statistically significant results. Design decisions could be made to avoid the need for these corrections, meaning a smaller sample size could be possible.

Intercorrelations showed a relatively small relationship between measurements in this study. Further piloting and refinement of instruments should be able to strengthen the relationship between measures of similar constructs prior to the next study.

I found it surprising that there was not a stronger relationship between the warrant selection and the warrant writing tasks. My assumption in developing the study was that the process of selecting an effective warrant would be clearly related to a student’s ability to write an effective warrant, and would be less demanding than warrant writing. However, selecting an effective warrant requires students to comprehend the text, the claim and evidence, and the warrants of another, and requires them to analyze the reasoning within those warrants. The warrant writing task in which a claim and evidence were given requires students only to understand the text and the claim and evidence given, not the warrants of another or the reasoning they contain. Though students must generate their own warrants, they can rely on the portions of the text they understand to do so.

In terms of the warrant selection task, students should be asked to provide an explanation in writing for why they chose each answer they did, to get a clearer picture if their response was due to the type of historical thinking, the foil, or some other reason. The warrant ranking task should be refined so that students are recalibrated and asked to rank all items by order of
effectiveness, not just the ones they selected. The warrant writing task should be expanded to include several items reflecting each type of historical thinking, so that issues of measurement validity can be addressed without needing to combine scores into an overall measure. By expanding the think aloud to include a representative sample of students at different reading ability levels in both the treatment and the comparison treatment condition, I would be able to make better inferences about the thinking that each group undergoes while completing the instruments.

After making these changes to assessment design, analysis could also be refined to take into account performance across assessment measures. Looking across performance on warrant selection, ranking, writing, and essay writing tasks could help inform instruction for teachers. For instance, assessment data may reveal profiles of students that can select certain types of effective warrants, but struggle with warrant ranking and show common errors in their warrant writing. This could help teachers design the instruction that will best support their areas of need as learners. Though this analysis was intended to occur the current study, variations in student attendance made such analysis challenging.

This study also contributes to an emerging body of work about innovative assessments of historical thinking. Though Northwest High recently abandoned the use of multiple choice tests altogether, they continue to dominate assessments in history (Martin, Maldonado, Schneider & Smith, 2011). However, multiple choice assessments cannot effectively measure the complexity of historical thinking. While innovative formative assessments of historical thinking are beginning to be developed and popularized (e.g. Breakstone, Smith & Wineburg, 2013), these assessments focus on heuristics as used in reading and to not the ability to evaluate or write effective warrants. While some assessments contain innovative features such as arranging documents by date likely written can assess historical thinking separate from expressive language use, these innovative writing assessments still contain common confounds and it is difficult to
know if the quality of student reasoning reflected in writing is due to their reading, thinking, or writing ability. My assessment design has the potential to help teachers make better inferences about the specific strengths and areas for growth within students’ historical thinking and writing skills by helping assess thinking about warrants separately from writing them.

This study can also inform thinking about the design of items meant to stimulate specific types of historical thinking for both researchers and classroom teachers. The warrant writing tasks were designed to telegraph certain types of historical thinking on the part of students. For instance, the second warrant writing task item successfully stimulated more corroboration on the part of students by including multiple documents from different sources. Further study of what assessment items tend to successfully cue certain types of historical thinking can help provide information about how to develop assessments that allow students to engage in specific types of thinking, and analysis can be refined to track when students use warrants not telegraphed by the stimulus material.

As my literature review indicated, the lack of attention paid to warrants is not unique to the history classroom. For instance, argumentative writing is also a key part of science and English classrooms as well. A disciplinary approach that foregrounds the kinds of warrants used in scientific argumentation and literary analysis should be applied to study warrant use and design interventions for these disciplines as well.

**Implications for instruction: Characteristics of intervention and theory.** If we want students to be prepared for the demands of college and career, they must be taught the skills they will need to be successful. This research addresses an area of disciplinary literacy instruction that is largely ignored by practitioners and researchers alike, namely instruction in warranting historical arguments. When the most reliable measures in this study were combined using MANOVA, there was a statistically significant overall effect. Means for the treatment condition were consistently higher, though not statistically significant, on additional t-tests. These findings
begin to make a case for changing the way argumentative writing is taught in history classes in similar contexts. Therefore, I will briefly lay out the characteristics of an effective intervention, as I develop a theory of adolescent argumentative writing instruction informed by this cycle of design-based research. These characteristics may change as new cycles are completed and as the theory evolves.

**Backwards design.** As a design-based research study, it is informative to first consider how the study prototype was developed, in order to apply this process to further prototyping and to identify elements that may inform the planning of historical argumentative writing instruction. While the idea of backwards design for writing is not new (e.g. Vermont Writing Collaborative, 2008; Wiggins and McTighe, 2005), such work rarely emphasizes the specialized literacy practices of a particular discipline. In designing this study, I had to plan backwards from the central understanding I wanted students to have about the explosion of the U.S.S. Maine, the historical thinking heuristics they would use to reason about the documents, and the writing skills I wanted to focus on to help them demonstrate this understanding in writing, namely warrants. I determined the argument I wanted students to be able to make at the end of the study and wrote a question (“Did the explosion of the U.S.S. Maine cause the U.S. to invade Cuba?”) to focus their thinking. I wrote an essay response to the question myself to better understand the demands of what I was asking of students, being aware of the heuristics I used in my own argument. This process helped me select particular passages within texts for modeling and reading closely with students that were challenging, provided opportunities to engage in historical thinking, and contained essential information to answer the question (Vermont Writing Collaborative, 2008). These sections became the places in the text where I embedded my warrant-generating questions, targeting whatever heuristic was relevant for that particular section and thereby helping students understand when to engage with particular types of thinking. Similarly, the degree of intentionality and backwards planning that the prototype development required mirrors many of
the processes teachers need to engage in to plan this sort of instruction. This process is very
different from how teachers typically plan, and from how I used to plan as a classroom teacher.
Though the process itself is important, it is by no means a finished product and will be continued
to be refined through further cycles of design-based research.

**Careful selection of texts.** The selection of texts is essential for this instructional
approach. As I planned for this study, I read and reread documents closely myself as a historian
would, annotating them and paying careful attention to what made them complex, including both
universal elements of qualitative complexity and those unique to historical text. For instance,
along with understanding how authors chose specific words to create a particular tone, I had to
take into account who the author was and what their motivation might be, based on my
knowledge of the time period and how this document corroborated or contradicted other sources.
By reading texts closely and understanding the complexities and the opportunities for historical
thinking they present, teachers can identify those parts of a text to read closely with students and
target with warrant-generating questions. In this study, consultation with history experts and
feedback from teachers who took the instruments themselves and piloted them with students was
essential in tuning my plans for students to engage with text. Collaborative planning could make
this approach more manageable for classroom teachers.

While including texts that have a balance of sources, perspectives and evidence for both
sides of an argument is necessary, it is by no means sufficient. As Wolfe, Britt and Butler (2009)
indicate, students attend to information that fits their particular side, and tend to ignore
information that does not. Students typically do not use strategies for comprehending multiple
texts without being taught them (Monte-Sano & De La Paz, 2012; Wineburg, 2001) and just
telling students to use historians’ strategies is generally not enough to improve their use (Britt &
Aglinskas; Reisman, 2012; Shanahan & Shanahan; 2008). Explicitly modeling reading across
texts (in this case, through the modeling of corroboration and the use of warrant-generating
questions) is often necessary to help students do this demanding thinking. Students in this study indicated that typical classroom instruction and assignments dealt with one document at a time, whereas this intervention moved from a single text, to paired texts, to reading across multiple documents. In the semi-structured interview, Ethan spoke extensively about how using corroboration to examine the New York Times and the New York Journal’s conflicting accounts of the same event immediately following the explosion of the Maine helped him realize the power of corroborating to warrant an argument. Students in the treatment condition had higher mean performance on warrant selection items linked to corroboration, and showed more evidence of corroboration in their warrants, though the difference was not statistically significant.

**Close reading and heuristic use.** Researchers have suggested that close reading is not only an important approach for historians in its own right, but may enable sourcing and corroboration by helping readers to cue on specific language or structural decisions that reveal authors’ perspectives or biases (Reisman, 2012). Despite the emphasis on close reading associated with the shift to the Common Core State Standards, the history teachers at Northwest High had not taught students how to read closely in history class, and had received no training in it themselves. Based on conversation during instruction, students in the study associated close reading with something they do in English class and did not see those skills as relevant for history. Refinements to this study design would include explicit instruction in how close reading works in the discipline of history that intentionally builds on the knowledge students have from their experience in English classrooms.

In future study designs I will be more intentional about helping students understand how close reading works in combination with other historical thinking heuristics. When historical thinking heuristics are taught, they tend to be taught in isolation, just as generic comprehension strategies such as questioning and making connections were initially studied and taught in isolation before researchers began studying and advocating for their use in combination. Like
expert readers who seamlessly combine comprehension strategies, historians regularly leverage several heuristics in the course of making an argument. This study began by introducing heuristics in isolation, before briefly introducing how they can be combined. I underestimated the need to be explicit about why and how to combine heuristics. I assumed that a natural extension of, for instance, looking closely at how an author used language to describe the explosion of the U.S.S. Maine in one event, would be to reflect on the significance of that language use given that document’s source and the author’s particular perspective or interests. While some students did use heuristics in combination, there was a tendency for many students to use them independently from one another, missing opportunities to strengthen their arguments by combining heuristics to warrant their claims.

Reisman (2012) made the point that a limited form of sourcing, namely identifying who wrote the document and when, seems to be a routine for many students, as they have been taught to do this as a first step any time they read a historical document. Prior to the study sourcing was used as an act students were instructed to do when they first pick up a text that is finished once they identify the author. In the essay writing task, students in the lower quartiles in both treatment and comparison treatment relied heavily on sourcing. However, in the treatment condition, students in the third and fourth quartiles shifted to using close reading and to a lesser extent, corroboration. This may indicate students in the comparison treatment are still using sourcing as a routine, while students in the treatment condition using more sophisticated forms of reasoning about texts as a result of the intervention. There was even some evidence of sourcing being combined with other heuristics. For example, a third quartile student in the treatment condition used sourcing in combination with close reading when she analyzed the date a speech was given and the language the President used to warrant her argument that the explosion of the Maine was only an excuse to go war.
The use of warrant generating questions (Freeman, 2011, Toulmin, 1958) tailored to specific types of historical thinking can break the routine of only thinking about sourcing when one first picks up a document. This approach forces students to engage with what they know about, for instance, who wrote a particular text, at the very point in the text when it is most useful to do so. The warrant generating questions prompt students to reason about the significance of the source information to the larger argument they are constructing. Embedding the use of heuristics to warrant emerging claims while building understanding of content through reading complex texts is an intricate yet promising approach to teaching historical argumentative writing.

As I refine the intervention design moving forward, I will select arguments and texts that offer opportunities to combine warrants in powerful ways. I will also find clear instances where a conclusion warranted by one heuristic would be in conflict with a conclusion warranted through the use of a different heuristic, in order to foreground the importance of combining heuristics with students. Such work should also lead to less reliance on sourcing and more balance in heuristic use.

Though the mean performance on corroboration was higher for the treatment condition, the difference was not statistically significant. It is likely that students need additional scaffolds to support the collection and analysis of evidence within and across texts, beyond warrant generating questions and sentence frames for corroboration. Part of the rationale for this is based on the performance of the comparison treatment. The comparison group performed similarly on some measures, through an explicit focus on gathering evidence within and across texts. This implies that support gathering evidence would be valuable for the treatment group as well.

Teachers can help students create organizers to gather and organize evidence that supports their emerging arguments, expanding on the simple T-Charts used in this study to more involved charts capturing the key events and the relationship between them (Shanahan T. & Shanahan, S., 2008).
In addition, the corroboration students did tended to be focused on evidence contained in texts, and not interpretation of that evidence. Charting of evidence can include employing heuristics to analyze evidence from texts using heuristics, and then synthesizing across texts. For instance, students can examine how a construct is implicitly defined in each of three texts by looking closely at the words and phrases authors use. After this has been done, they can use corroboration to think more deeply about the similarities and differences between these definitions across texts.

**Scaffolding student writing.** Writing instruction is an area where many history teachers feel they need support. As Ms. Jones indicated was the case at Northwest High, most history teachers have not received any formal training in argument writing, and tend to assign argument writing but not explicitly teach it. The treatment condition of this study provided students with sentence frames (Graff, 2006) based on warrant generating questions that also included historical thinking heuristics to help students generate their written warrants, along with modeling and peer feedback.

In the think aloud task, two students requested to use the sentence frames they had learned that week. It was clear students found them valuable. Though the mean performance was higher for students in the treatment condition, it may have been significantly more so if students had enough time to internalize the use of these sentence frames so they could more easily transfer them to a task. Sentence frames can be updated to incorporate common combinations of heuristics, such as sourcing with corroboration. In connection with the idea of combining heuristics, modeling and comparing how an argument is weakened or strengthened by combining heuristics would be a powerful way to help students incorporate this into their own argumentative writing.

**Differentiating between more and less effective warrants.** Follow up ANOVAs showed that there was a statistically significant difference in students’ ability to identify correct warrants.
However there was no difference in students’ abilities to avoid incorrect warrants. The focus of instruction was on generating effective warrants, but not choosing from a list of warrants to determine which was the most effective. The warrant ranking data also indicated no statistically significant difference between treatment and comparison treatment groups. Therefore, students would benefit from instruction and practice differentiating between possible warrants for a particular claim. This could take the form of sorts where students physically order warrants and discuss why they ordered them the way they did, or activities where students select one or more warrants from a series of possible warrants and argue why they feel these are valid. This skill building work could then be applied to closely reading historical documents and analyzing how arguments are warranted, considering the implications of the choices the author made on the effectiveness of the argument.

**Opportunities to discuss warrants and receive feedback.** Before students are ready to write, they need opportunities to talk about their emerging thinking. In the think aloud data, students reported that more feedback from teachers and peers would be helpful. Just as historians refine their emerging ideas by discussing them and getting feedback in writing through peer review, students need discussion and feedback on their own writing. Both the treatment and comparison treatment received periodic feedback from peers, but due to the short duration of the study the amount of feedback was limited. Discussing ones’ writing and receiving feedback on the strength of ones’ argument could be an opportunity to help develop students’ identities as readers and writers as members of a community doing historical inquiry. Norms of discourse and values of the discipline of history, implied in the heuristics that historians use but not explicitly taught to students in this study, should be shared with students to support this work.
Analyzing student writing to influence instruction. In giving feedback on student writing, teachers should begin by being aware of warrants as something to pay attention to in writing. In fact, I was not explicitly aware of warrants as a classroom teacher until I began my doctoral study. An initial examination of student argumentative writing can help teachers understand whether or not students are using claims, evidence and warrants in their writing. If they are using warrants, teachers can determine what the nature of these warrants is (e.g. based on non-historical thinking, relying on heavily on one sort of historical thinking or another, serving more as an elaboration than an explanation of how the evidence supports the claim, etc.). Once specific heuristics are taught, teachers can hold students accountable for using these tools as they read through examining student annotations, and can look for evidence of heuristic use in student writing, making the newly introduced heuristic the focus of feedback. Rubrics or checklists for looking at student work can expand over the course of the year, as students develop skill with additional heuristics. Examining claim and evidence alongside warrants can help teachers think about whether struggles with warrants are due to a lack of skill in using heuristics, or whether broad claims or insufficient evidence are also contributing to student difficulty in warranting their arguments.

One of the challenges of this study is that it was designed without a depth of knowledge about what these students already knew and could do in terms of heuristic use and argumentative writing. Study redesign will begin with an analysis of student writing prior to the subsequent lesson prototype development to determine the best entry point for instruction in warrants, building on the types of reasoning students are already demonstrating. This will also provide valuable pre-post data as subsequent cycles of design are undertaken.
Conclusion

The results of this study are important for three reasons. First of all, the research addressed an area of advanced literacy instruction that is recognized as increasingly valuable for students in the 21st century, namely historical argumentative writing. The skills of reading and evaluating multiple complex historical texts reflect challenges that are similar to those students face when reading documents online, and are essential for engaged citizenship (Wineburg & Reisman, 2014). This study specifically targeted the writing of warrants, an essential yet understudied aspect of argumentation. Warrants are hard to research because they are reliant on their claim and evidence partners. They are hard to separate from students’ overall writing and thinking ability and are not always explicitly stated. In history, warrants often reflect the specific ways of thinking about documents and historical evidence that historians use. This study was designed to help students use historical thinking heuristics and close reading to warrant their claims.

Second, this study used an innovative approach to assessment design that provided information about students’ use of the historical thinking heuristics of sourcing and corroboration along with close reading as they selected, ranked, and wrote their own warrants. Past assessments have not been able to provide this sort of information. The warrant selection and ranking tasks included answer choices that reflected different types of historical thinking, and foils that reflected common errors. The multiple select nature of the assessment reflected the probabilistic nature of historical thinking. Measures related to selecting correct warrants, avoiding incorrect warrants, and ranking warrants provided information about student ability to differentiate between more effective and less effective warrants. By providing the claim and evidence, the warrant writing task made it possible to isolate students’ ability to write a warrant from their ability to generate a claim or select evidence, while the essay writing task showed how students could apply this skill when generating their own claim and evidence. The think alouds and semi-
structured interview provided insight into the thinking students were doing as they completed the assessment and their own experience of the instruction.

The assessment design makes an important contribution to a growing field of innovative formative assessments of historical thinking. Taken as a whole the assessments in this study can help teachers determine whether or not the issue is the thinking skills underlying the generation of a warrant, or the process of putting this thinking in their own writing that is causing students’ difficulty, allowing teachers to target student areas of need. It can also reveal specific types of historical thinking where students may need more support.

Third, this study produced a significant overall effect on student warrant selection and writing ability. Messnick (1992) explains the principle of consequential validity, that the validity of a claim is based on its consequences, meaning the change it can produce in a system. When the most reliable measures in this study were combined, there was a significant overall effect for the treatment condition emphasizing instruction in heuristic use on the selection of correct warrants and the writing of effective warrants. Though issues in instrument design limited the impact of subsequent t-tests for individual measures, the means were consistently higher for the treatment condition. Post-hoc power analyses indicated that a moderately larger sample size may yield a significant treatment effect on these individual measures. Given the design-based nature of this study, there is sufficient consequential validity to use this prototype to define the characteristics of effective instruction in historical argumentative writing and generate theory that can inform work in similar contexts.

As a result of this design-based research, I made many refinements to my understanding of the characteristics of an effective instructional intervention. Some of these refinements will hinge on looking at student work at the start of the unit to identify what skills students demonstrate in warranting arguments. It will include characterizing student use of claim, evidence, and warrant in order to refine the instructional prototype to build upon student strengths
and address misconceptions. Planning instruction alongside the classroom teacher based on what students have and have not been taught will address confusion or redundancies about similar terms such as reasoning and warrant. This approach will also make it possible to bridge how close reading works in the discipline of history to knowledge students may have about close reading from English class. In addition, an ongoing cycle of examining student writing and warrant use over time and documenting changes can inform decisions about when to remove scaffolding and what refinements should be made to the intervention.

In terms of the intervention itself, I will be more explicit about how heuristics can be used in combination, and how combining heuristics such as sourcing and corroboration can strengthen an argument. This should result in students being more capable at selecting effective warrants and writing warrants supported by multiple heuristics. Choosing text sets for instruction where heuristics can clearly be used in combination to strengthen a warrant, or where they can lead to conflicting interpretations of text, may help build student skill in using warrants in combination. This should help increase balance across the types of heuristics writers employ in their warrants. Including additional supports for reading across texts such as evidence organizers may help students use corroboration more effectively to go deeper in their analysis. For instance, having evidence from multiple texts recorded in one organizer may enable students to more easily reflect on the sources of that evidence. They may be able to more easily use the source information to reason about how each author may be using language intentionally to have a particular impact on the reader.

Providing students with additional instruction in differentiating between more and less effective warrants for a particular claim is another refinement I have made to the characteristics of an effective intervention, based on the fact that students in the treatment condition were no better at avoiding incorrect warrants. Practice sorting and selecting warrants and discussing the decisions one makes may lead to improvements in warrant selection and ranking. Reading
historical documents that contain arguments and explicitly stated warrants, evaluating their
effectiveness and thinking about how warrants might be improved through the use of heuristics
may help strengthen student writing as well. Students will become adept at analyzing and
improving warrants and can apply this to their own writing.

Finally, expanding student opportunities to discuss their writing and receive feedback
from peers and teachers on warrants may improve their writing. The same norms of the
discipline can be used to shape the manner of discussion and feedback in order to apprentice
students into a community of scholars doing historical inquiry. Future work can more explicitly
address the epistemology underlying the heuristics that historians use, or how such beliefs align
or do not align with students own beliefs and identities as readers and writers. This approach fills
a real need because history teachers often do not provide explicit writing instruction, yet they are
the most qualified to teach students how to read and reason about historical texts and write
historical arguments, as called for by the Common Core State Standards.

The characteristics of this instructional intervention are grounded in a theory of teaching
historical argumentative writing that differs significantly from traditional history instruction.
Rather than focus on a textbook as an “unbiased” source of information, this study views history
as constructed, written by human beings for a particular purpose (Stahl & Shanahan, 2004).
Content area literacy approaches which treat literacy as the same set of generic strategies spread
across subject areas are often unsuccessful because they fail to take into account the unique
demands of reading, writing and thinking in a particular discipline. History teachers have been
rightfully resistant when such content area literacy approaches devalue content at the expense of
decontextualized reading skills. This theory of teaching argumentative writing promises to build
content knowledge and advanced literacy skill and strategy use concurrently through reading and
reasoning about historical texts and controversies, which may be more appealing to history
teachers and more engaging for students than either traditional or content-area literacy-based instruction.

This approach calls for history teachers to leverage their disciplinary expertise to develop units centered on these historical controversies, backwards planned from the understandings about content and the argumentative writing they want students to be able to do at the end of the unit. By reading texts carefully and doing the writing historical argumentative writing themselves, teachers engage in the work of the discipline in order to identify the best opportunities to model and scaffold the use of heuristics to warrant historical arguments. Such an approach, over time, will enable students to become metadiscursive, leveraging the literacy practices of the dominant Discourse of historical argumentation in their own lives.

The demands of 21st century citizenship require higher levels of literacy than ever before. This line of research is ultimately important because providing students with advanced literacy skills like historical argumentative writing is an issue of social justice. As the recent election cycle demonstrates, there is no shortage of competing and conflicting claims, all of which politicians and pundits assert are supported by mounds of evidence. Students who recognize the importance of warrants and can evaluate whether or not the evidence provided in an argument truly supports the claim will have greater agency in their lives. They will be less likely to be manipulated by sweeping promises based on dubious evidence or faulty reasoning, because they will know to corroborate reports with other sources of evidence. They will be able to see beyond rhetoric by analyzing who writes a particular document and their interest in presenting things a certain way, and make better decisions about what policies to support, what health insurance to choose, or how to spend or invest their money. In addition, while this research begins with students learning historical thinking heuristics, it need not end there. While the shorthand mental tools experts use to reason about multiple documents are powerful, they do not replace students’ own ideas and reasoning about their own lives. Combining the ability to warrant arguments with
historical heuristics and students’ existing knowledge and skills can enable them to be more fully informed and empowered citizens in an ever-changing society.
VI. REFERENCES


Purchased by William Randolph Hearst in 1895, the Journal published investigative and human interest stories that used a highly emotional writing style and included banner headlines and graphic images.

Source: Excerpt from New York Journal and Advertiser, February 17, 1898

DESTRUCTION OF THE WAR SHIP MAINE WAS THE WORK OF AN ENEMY

Assistant Secretary Roosevelt Convinced the Explosion of the War Ship Was Not an Accident.

The Journal Offers $50,000 Reward for the Conviction of the Criminals Who Sent 258 American Sailors to Their Death. Naval Officers All Agree That the Ship Was Destroyed on Purpose.

NAVAL OFFICERS THINK THE MAINE WAS DESTROYED BY A SPANISH MINE.

George Bryson, the Journal’s special reporter at Havana, writes that it is the secret opinion of many people in Havana that the war ship Maine was destroyed by a mine and 258 men were killed on purpose by the Spanish. This is the opinion of several American naval authorities. The Spaniards, it is believed, arranged to have the Maine drop anchor over a harbor mine. Wires connected the mine to the magazine of the ship. If this is true, the brutal nature of the Spaniards will be shown by the fact that they waited to explode the mine until all the men had gone to sleep.

Spanish officials are protesting too much that they did not do it. Our government has ordered an investigation. This newspaper has sent divers to Havana to report on the condition of the wreck. This newspaper is also offering a $50,000 reward for exclusive
APPENDIX A (continued)

evidence that will convict whoever is responsible. Assistant Secretary of the Navy Theodore Roosevelt says he is convinced that the destruction of the Maine in Havana Harbor was not an accident. The suspicion that the Maine was purposely blown up grows stronger every hour. Not a single fact to the contrary has been produced.
APPENDIX A (continued)

Document B: New York Times

Established in 1851, the New York Times provided investigative coverage of local New York issues and events, as well as national and international news.


MAINE’S HULL WILL DECIDE

Divers Will Inspect the Ship’s Hull to Find Out Whether the Explosion Was from the Outside or Inside.


It has been a busy day for the Navy Department. The war ship Maine was destroyed in Havana Harbor last night. Officials in Washington and Havana have been sending cables all night long. Secretary Long was asked whether he thought this was the work of the enemy. He replied: “I do not. I am influenced by the fact that Captain Sigsbee has not yet reported to the Navy Department. It seems he is waiting to write a full report. So long as he has not made a decision, I certainly cannot. I should think from the signs, however, that there was an accident – that the magazine exploded. How that came about I do not know. For the present, at least, no other war ship will be sent to Havana.”

Captain Schuley, who knows a great deal about war ships, did not entertain the idea that the Maine had been destroyed on purpose. He said that fires would sometimes start in the coal bunkers, and he told of such a fire on board another war ship that started very close to the magazine. The fire became so hot that the heat blistered the steel wall between the fire and the ammunition before the bunkers and magazine were flooded with water to stop the fire. He did not believe that the Spanish or Cubans in Havana had either the information or the
APPENDIX A (continued)

equipment necessary to blow up the magazine, while the Maine was under guard.

Document C: Awake United States

This song was rushed into print between the sinking of the Maine on February 16, 1898 and the declaration of war on April 25, 1898.

Source: Marie Elizabeth Lamb, Awake United States! (New Orleans, LA, 1898).

Eagle soar on high, and sound the battle cry!

1. How proudly sailed the warship Maine,
a Nation’s pride, without a stain!
A wreck she lies, her sailors slain.
By two-faced butchers, paid by Spain!

Refrain:

Eagle soar on high,
And sound the battle cry
Wave the starry flag!
In mud it shall not drag!

2. Why does the breeze such sad thoughts bring,
Like murmuring seas the echoes sing?
Why do clouds thus backward roll.
Like wave on wave, on rock on shoal!

3. Awake! Thy Stars and Stripes unfurl,
And shot and shell and revenge hurl!
Though clouds gather, they will go,
and sunlight follow after woe.

Refrain:

Awake! it is no dream;
Do you hear the sailors scream?
Comrades will you go?
Avenge the cruel blow!

And crush their marble heart!
Document D: President McKinley’s State of the Union Address

President McKinley went before Congress to ask for a declaration of war against Spain.

Source: Excerpt from President William McKinley's War Message to Congress, April 11, 1898.

The reasons to go to war are these:

First, in the cause of humanity and to put an end to the bloodshed, starvation, and horrible miseries that are now there.

Second, we owe it to our citizens in Cuba to give them protection for life and property which no government there can or will give.

Third, the right to get involved may be justified by the very serious injury to the trade and business of our people, and by the reckless destruction of property and ruin of the island.

Fourth, and this is very important . . . With such a fight waged for years in an island so near us and with which our people have such trade and business relations; when the lives and liberty of our citizens are in constant danger and their property destroyed and themselves ruined; where our trading vessels might be seized by warships of a foreign nation, all these and others are a constant threat to our peace...

I have already sent to Congress the report on the destruction of the battleship Maine. The destruction of that noble ship has filled the national heart with horror.

The destruction of the Maine, by whatever cause, is an obvious sign that things in Cuba are intolerable. The Spanish government cannot assure safety and security to the American Navy in the harbor of Havana on a mission of peace, and rightfully there . . .
Document E: Reconcentration Camps

By the late 1800s, the Spanish were losing control of their colony, Cuba. Concerned about guerilla warfare in the countryside, they moved rural Cubans to "Reconcentration" camps where the Spanish claimed they would be better able to protect them. However, people around the world saw newspaper reports that described horrible conditions in the camps for the Cuban people, who were called “reconcentrados.” This account was forwarded to Washington, D.C., by Fitzhugh Lee, U.S. Consul-General in Havana, who said its author was “a man of integrity and character.”

Source: Excerpt from telegram sent by Fitzhugh Lee, U.S. Consul-General in Cuba, November 27, 1897. Havana, Cuba.

SIR:

We will tell you what we saw with our own eyes: 460 women and children thrown on the ground, heaped in piles like animals, some in a dying condition, others sick and others dead, dirty and helpless.

Among the many deaths we witnessed, there was one scene impossible to forget. There was a young girl of 18 years, whom we found seemingly lifeless on the ground; on her right-hand side was the body of a young mother, cold and rigid, but with her young child still alive clinging to her dead breast; on her left-hand side was also the corpse of a dead woman holding her son in a dead embrace.

Bodies were piled up, dead and alive, so that it was impossible to take one step without walking over them; it was very dirty, there was little light, air, and water; there was not enough food to live on.

From all this evidence, we think that the number of deaths among the reconcentrados has amounted to 77 percent.
Document F: “Prepared to Move”

Lee was appointed the U.S. Consul-General in Havana, Cuba in 1896 by President Grover Cleveland. He wrote this letter to the Assistant Secretary of State in the U.S. almost three months before the Maine explosion. The consul-general was a government official living in a foreign city charged with overseeing the protection of U.S. citizens and promoting trade. He would make periodic reports to his superiors in the U.S. Department of State.

Source: Excerpt from telegram sent by Fitzhugh Lee, U.S. Consul-General in Cuba, December 3, 1897. Havana, Cuba.

SIR:

I still think that at least two war ships should be at Key West, prepared to move here at short notice, and that more of them should be sent to Dry Tortugas, and a coal station be established there.

We should do this to protect the Americans on the island and their properties, both of which are objects of the greatest concern to our Government.

FITZHUGH LEE,
Consul-General.
APPENDIX A (continued)

Document G: March of the Flag

Beveridge gave this speech while he was campaigning to become a senator for Indiana. The speech helped him win the election and made him one of the leading advocates of American expansion.

Source: Excerpt from Albert J. Beveridge’s Senate campaign speech, Sept. 16, 1898, Indiana.

Fellow citizens,

It is a noble land that God has given us, a land that can feed and clothe the world. It is a mighty people that he has planted on this soil. It is a glorious history our God has given his chosen people, a history of soldiers who carried the flag across the blazing deserts and through the hostile mountains, even to the gates of sunset, a history of a people who overran a continent in half a century.

The opposition tells us that we should not govern a people without their consent. I answer, the rule of liberty that government gets its authority from the consent of the governed, applies only to those who are capable of self-government. I answer, We govern the Indians without their consent, we govern our territories without their consent, we govern our children without their consent.

They ask us how we will govern these new lands. I answer: If England can govern foreign lands, so can America. If Germany can govern foreign lands, so can America. What does all this mean for every one of us? It means opportunity for all the glorious young manhood of the republic — the most ambitious, impatient, militant manhood the world has ever seen. It means that the resources and the commerce of these immensely rich lands will be increased.

In Cuba alone, there are 15 million acres of uncut forest. There are mines of iron. There are millions of acres still unexplored.

It means new employment and better wages for every working man in the Union. . . .

Ah! as our commerce spreads, the flag of liberty will circle the globe. And, as they salute the flag, ignorant peoples will know that the voice of Liberty is speaking, at last, for them; that civilization is dawning, at last, for them — Liberty and Civilization, those children of Christ’s gospel.

Fellow Americans, we are God’s chosen people.
Document H: Textbook

The following is an excerpt from a U.S. History textbook written in 2006 covering the period leading up to and including the explosion of the U.S.S. Maine and the decision to invade Cuba.


War Fever Escalates
In 1896, Spain responded to the Cuban revolt by sending General Valeriano Weyler to Cuba to restore order. Weyler tried to crush the rebellion by herding the entire rural population of central and western Cuba into barbed wire concentration camps. Here civilians could not give aid to rebels. An estimated 300,000 Cubans filled these camps, where thousands died from hunger and disease.

Headline Wars
Weyler's actions fueled a war over newspaper circulation that had developed between the American newspaper tycoons William Randolph Hearst and Joseph Pulitzer. To lure readers, Hearst's New York Journal and Pulitzer's New York World printed exaggerated accounts ... of “Butcher” Weyler's brutality. Their sensational style of writing, which exaggerates the news to lure and enrage readers, became known as yellow journalism.

When President William McKinley took office in 1897, demands for American intervention in Cuba were on the rise. Preferring to avoid war with Spain, McKinley tried diplomatic means to resolve the crisis. At first, his efforts appeared to succeed. Spain recalled General Weyler, modified the policy regarding concentration camps, and offered Cuba limited self-government.

In February 1898, however, the New York Journal published a private letter written by Enrique Dupuy de Lôme, the Spanish minister to the United States. The de Lôme letter criticized President McKinley, calling him “weak” and “a bidder for the admiration of the crowd.” ... Americans were angry over the insult to their president.

The U.S.S. Maine Explodes
Only a few days after the publication of the de Lôme letter, American resentment toward Spain turned to outrage. Early in 1898, President McKinley had ordered the U.S.S. Maine to Cuba to bring home American citizens in danger from the fighting and to protect American property. On February 15, 1898, the ship blew up in the harbor of Havana. More than 260 men were killed. At the time, no one really knew why the ship exploded; however, American newspapers claimed that the Spanish had blown up the ship.

War with Spain Erupts
Now there was no holding back the forces that wanted war. “Remember the Maine!” became the rallying cry for U.S. intervention in Cuba. It made no difference that the Spanish government agreed, on April 9, to almost everything the United States demanded, including a six-month cease-fire.
### APPENDIX B: INSTRUMENTATION TEXT CHARACTERISTICS

Table 1: Sources for Warrant Selection, Warrant Ranking, and Warrant Writing Tasks

<table>
<thead>
<tr>
<th>Doc.</th>
<th>Source Information</th>
<th>Source Type</th>
<th>Lexile</th>
<th>Mean Sentence Length</th>
<th>Mean Log Word Frequency</th>
<th>Word Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Maine's Hull will Decide, New York Times (Feb. 17, 1898): 1.</td>
<td>Newspaper Article</td>
<td>1190L</td>
<td>22.54</td>
<td>3.74</td>
<td>293</td>
</tr>
<tr>
<td>C</td>
<td>Lamb, Marie Elizabeth. Awake United States! New Orleans, LA, 1898.</td>
<td>Song</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>F</td>
<td>Lee, Fitzhugh. Fitzhugh Lee, US Consulate-General in Cuba, to Assistant Secretary of State Day, 3 December, 1897. In Message from the President of the United States, transmitting, in response to the resolution of the House of Representatives, Dated February 14, 1898. Calling for information in respect to the condition of the reconcentrados in Cuba, the state of the war and the country, and the prospects of projected autonomy in that island. Washington, DC: Government Printing Office, 1898. 11-12.</td>
<td>Telegram</td>
<td>1660L</td>
<td>42.5</td>
<td>3.73</td>
<td>85</td>
</tr>
</tbody>
</table>
Table 1: Source Types for Warrant Selection, Warrant Ranking, and Warrant Writing Tasks

<table>
<thead>
<tr>
<th>Doc.</th>
<th>Source Information</th>
<th>Type of Historical Source</th>
<th>Lexile</th>
<th>Type</th>
<th>Mode</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>&quot;Destruction of the War Ship Maine was the Work of an Enemy,&quot; New York Journal and Advertiser (Feb. 17, 1898): 1.</td>
<td>Primary</td>
<td>1250L</td>
<td>Corroborated with C; Contradicted with B, H</td>
<td>Written</td>
<td>Newspaper article suggesting sinking of Maine was due to a mine placed by the Spaniards</td>
</tr>
<tr>
<td>B</td>
<td>Maine's Hull will Decide, New York Times (Feb. 17, 1898): 1.</td>
<td>Primary</td>
<td>1190L</td>
<td>Corroborated with H; Contradicted with A, C</td>
<td>Written</td>
<td>Newspaper article indicating the sinking of the Maine was likely an accident</td>
</tr>
<tr>
<td>C</td>
<td>Lamb, Marie Elizabeth. Awake United States! New Orleans, LA, 1898.</td>
<td>Primary</td>
<td>N/A</td>
<td>Corroborated with A; Contradicted with B</td>
<td>Written</td>
<td>Song printed between the sinking of the Maine and the declaration of war on Spain blaming Spain for its sinking and calling for the US to attack Spain to avenge this act.</td>
</tr>
<tr>
<td>E</td>
<td>Lee, Fitzhugh. Fitzhugh Lee, US Consulate-General in Cuba, to Assistant Secretary of State Day, 27 November, 1897. In Message from the President of the United States, transmitting, in response to the resolution of the House of Representatives, Dated February 14, 1898, Calling for information in respect to the condition of the reconcentrados in Cuba, the state of the war and the country, and the prospects of projected autonomy in that island. Washington, DC: Government Printing Office, 1898. 10-11.</td>
<td>Primary</td>
<td>1050L</td>
<td>Corroborated with D, H</td>
<td>Written</td>
<td>Unsigned account included in telegram sent by a US official living in Cuba describing death and despair in Cuban “reconcentration” camps</td>
</tr>
<tr>
<td>F</td>
<td>Lee, Fitzhugh. Fitzhugh Lee, US Consulate-General in Cuba, to Assistant Secretary of State Day, 3 December, 1897. In Message from the President of the United States, transmitting, in response to the resolution of the House of Representatives, Dated February 14, 1898, Calling for information in respect to the condition of the reconcentrados in Cuba, the state of the war and the country, and the prospects of projected autonomy in that island. Washington, DC: Government Printing Office, 1898. 10-11.</td>
<td>Primary</td>
<td>1660L</td>
<td>Unique</td>
<td>Written</td>
<td>Telegram sent by US official living in Cuba three months before US invasion of Cuba calling for the US to station two war ships at Key West “prepared to move” to Cuba on short notice, and to send more ships to the Dry Tortugas and establish a cola station</td>
</tr>
<tr>
<td></td>
<td>Office, 1898. 11-12.</td>
<td></td>
<td></td>
<td>there to protect the safety and property of Americans living in Cuba</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>Beveridge, Albert J. &quot;March of the Flag.&quot; (September 16, 1898, Indiana). In The Meaning of the Times, and Other Speeches. Indianapolis: Bobbs-Merrill, 1908. 47-57.</td>
<td>Primary</td>
<td>940L</td>
<td>Unique</td>
<td>Written</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Excerpt of Albert J. Beveridge speech asserting U.S. should expand into new lands in general and Cuba in particular</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>The Americans: Reconstruction to the 21st Century, McDougall Littell, 2006, Ch. 10, p. 348</td>
<td>Tertiary</td>
<td>1090L</td>
<td>Corroborated with B, E; Contradicted with A, C</td>
<td>Written</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Textbook account stating the Maine sent to return American citizens in danger and protect US property in Cuba. It stated that no one knew why the ship exploded though US papers made it seem it was the Spanish who were to blame.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A. The Spanish American War

This contemporary textbook excerpt gives an overview of the events leading up to the explosion of the U.S.S. Maine and the invasion of Cuba.


American power and economic interests around the world were growing by the end of the 19th century. Still, the United States remained reluctant to risk war with other powers to acquire colonies. That changed, however, in 1898, when America went to war against Spain. The United States acquired colonies and became a world power.

Spain was an imperial nation in decline. Its formerly vast empire had dwindled to a small number of possessions, including the Philippine Islands in the Pacific and the Caribbean islands of Puerto Rico and Cuba.

Cubans Rebel Against Spanish Rule: By 1897, American businessmen had invested $50 million in sugar cane plantations and other ventures in Cuba. They saw nearby Cuba as a growing market for American products. However, the island was very unstable. Yearning for freedom, the Cubans repeatedly rebelled against Spanish forces. In response, Spanish General Valeriano Weyler devised a plan to deprive the rebels of food and recruits. He herded the rural population into Reconcentration camps, where tens of thousands died from disease and starvation. The brutality of Spanish tactics deepened American affection and sympathy for the rebels. But other Americans, especially business people, were worried about U.S. economic interests in Cuba and hoped Spain would quickly put down the rebellion.

The Yellow Press Inflames Opinion: Rival newspaper publishers Joseph Pulitzer and William Randolph Hearst heightened the public’s dislike of the Spanish government. To boost their readership, Pulitzer’s New York World, Hearst’s New York Journal, and other similar newspapers pasted sensational headlines and pictures on their front pages. Their stories exaggerated Spanish atrocities and compared Cuban rebels to the patriots of the American Revolution.

President William McKinley warned the Spanish to quickly establish peace, or the United States would take whatever steps it “should deem necessary to procure this result”.
Spain recalled General Weyler and offered the Cuban rebels some reforms. But the rebels insisted on independence, which Spain refused to grant. McKinley ordered the battleship Maine to Havana harbor to protect American citizens in Cuba.

Then, in February 1898, the Journal published a private letter written by Enrique Dupuy de Lome, Spain’s ambassador to Washington, D.C. The letter called McKinley a weak and stupid politician. Hearst published the letter under the sensational headline, “Worst Insult to the United States in Its History”.

**The Maine Blows Up:** Soon after the Journal published de Lome’s letter, the Maine exploded in Havana harbor. Of the 350 officers and crew on board at the time, 266 died. The Yellow Press promptly accused Spain of blowing up the battleship. But President McKinley did not ask Congress to declare war just yet. Instead, he ordered a special naval board of inquiry to investigate the cause of the explosion. Even after the investigation, there remained a great deal of uncertainty as to exactly what happened and who was to blame. Still, most people blamed Spain for the explosion.

**The Nation Goes to War:** War fever gripped the nation. In newspapers, speeches, and songs, patriotic Americans implored their fellow citizens to “Remember the Maine!” In response to American demands, Spain agreed to abolish the Reconcentration camps and make other concessions, but it was too little too late. On April 11, 1898, McKinley asked Congress for the authority to use force against Spain to end the fighting in Cuba “in the name of humanity, in the name of civilization, on behalf of endangered American interests”. Eight days later, Congress enacted four resolutions that amounted to a declaration of war on Spain. The navy quickly blockaded Cuban ports, and McKinley called for more than 100,000 volunteers to join the army. In response, Spain declared war on the United States.
B. Not an Accident

On February 15, 1898, an explosion ripped through the American battleship Maine, anchored in Havana harbor, sinking the ship and killing 260 sailors. Americans responded with outrage, assuming that Spain, which controlled Cuba as a colony, had sunk the ship, even though there was not sufficient evidence to indicate who was responsible. Publishers such as William Randolph Hearst and Joseph Pulitzer used their many newspapers to stir public opinion over the sinking of the Maine into a frenzy and sell more papers.

Source: New York World, 2/17/1898

(Special to the World)

NOT AN ACCIDENT, CAPT. SIGSBEE SAYS

Intimates in a Suppressed dispatch to Long that the Disaster Was Due to an Enemy-Is Now Investigating and Not Prepared to Speak Authoritatively.

WASHINGTON, Feb. 16.—A suppressed cable dispatch received by Secretary Long from Capt. Sigsbee announced the Captain’s conclusion, after a hasty examination, that the disaster to the Maine was not caused by accident.

He expressed the belief that whether the explosion originated from without or within, it was made possible by an enemy.

He requested that this intimation (hint) of his suspicions be considered confidential until he could conduct a more extended investigation.

This dispatch was laid before the President, at whose suggestion Assistant Secretary Day cabled Consul-General Lee to make whatever examination was possible himself and render assistance to Capt. Sigsbee.

On the same dispatch Capt. Sigsbee said that not more than one hour prior to the explosion the magazines and boilers had been carefully inspected, thus, in his judgment, precluding (preventing) the possibility of accident.
C. The Maine Disaster

A great deal of the American public’s outrage over the explosion of the U.S.S. Maine on February 15, 1898 was generated by media coverage of the incident. While sensational stories in newspapers like those published by William Randolph Hearst and Joseph Pulitzer were very influential in causing Americans to blame Spain for the explosion, the highly respected New York Times cautiously reported on February 17, 1898, that there “was no evidence to prove or disprove treachery” as a factor in the sinking of the battleship.

Source: New York Times, February 17, 1898

THE MAINE DISASTER

Capt. Sigsbee Reports the Number of Dead as 253 and of Survivors as 96.

ONLY THEORY AS TO THE CAUSE OF DISASTER

All the Information at Hand Tends to Indicate That the Loss Was Due to an Accident.

Most of the Rescued Men Sent to Key West on the Ward Line Steamer Olivette.

Nothing has been learned of the cause of the loss of the battle ship Maine. She is a burned and broken wreck, resting on the bottom of Havana Harbor, and two officers and 251 sailors have perished. There is no evidence to prove or disapprove treachery. Naval men tell of many ways in which the disaster could have been caused by accident which could not have been guarded against.

The Spanish authorities in Havana and Madrid have profusely expressed regret and sympathy. The people of Havana are reported to have done all they could to help the survivors and to show their sorrow for the dead. The newspapers of Madrid reflect in their utterances the course of the Government.

An investigation of the condition of the vessel will be made immediately, and until that has been done nobody can know whether it will be possible to raise her to show
whether the explosion was within or from without. The officers are reticent (hesitant to reveal their thoughts). Expressions by some of them indicate their opinion that there was an accident to one of the dynamo engines. Prominent members of both houses of Congress express suspicions that the Maine was destroyed by foul play, but say they will await evidence.

Document modified from original retrieved online from http://historymatters.gmu.edu/d/5473

**D. Monroe Doctrine**

In 1823, President James Monroe made a bold foreign policy speech to Congress that signified a departure from past U.S. isolationism. The principles he laid out in the speech would become known as the “Monroe Doctrine” and would influence policy decisions thereafter.

**Source:** Excerpt from President James Monroe’s Seventh Annual Message to Congress, December 2, 1823.

The American continents are free and independent. They are not to be considered as subjects for future colonization by Europe. The citizens of the United States respect and value the liberty and happiness of their fellow men in Europe. In the wars of the European powers, in matters relating to themselves, we have never taken any part. It is only when our rights are invaded that we resent injuries, or prepare for our defense. With the events in our half of the world, we are more immediately concerned.

We owe it, therefore, to the good relations existing between the United States and the European powers, to declare, that we should consider any attempt on their part to extend their system to any portion of this half of the world as dangerous to our peace and safety.

With the existing colonies or dependencies of any European power we have not interfered, and shall not interfere. But with the governments who have
declared and maintained their independence, and whose independence we have acknowledged, we would view any oppressive interference by any European power as a clear sign of an unfriendly attitude towards the United States . . .

Some of the language and phrasing in this document has been modified from the original. The original document can be found on the HTM website, www.historalthinkingmatters.org.
Questionnaire Prior to the U.S.S. Maine Unit

Before we begin the next unit, we want to know what you already know about writing arguments and how effective you feel you are at writing them. We also want to know what knowledge you already have about the topics we will be studying this unit. There are no right or wrong answers, so please be open and honest.

* Required

1. Name *

_________________________________________

2. Period
   Mark only one oval.
   ☐ 4/5
   ☐ 5/6
   ☐ 9
   ☐ 10

3. 1. What makes a written argument effective? *

   _______________________________________
   _______________________________________
   _______________________________________
   _______________________________________
   _______________________________________

4. 2. Name the parts of an effective written argument. *

   _______________________________________
   _______________________________________
   _______________________________________
   _______________________________________
   _______________________________________
5. **Self-evaluate your ability at argument writing.**

1 is the lowest rating, not at all effective. 5 is the highest rating, extremely effective. 
*Mark only one oval.*

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not at all effective</td>
<td></td>
<td></td>
<td></td>
<td>Extremely effective</td>
</tr>
</tbody>
</table>

6. **How do you feel about argument writing, compared to other things you study in school?**

1 is the lowest rating, meaning argument writing is one of your least favorite things to study in school. 5 is the highest rating, meaning argument writing is one of your most favorite things to study in school. 
*Mark only one oval.*

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Least favorite</td>
<td></td>
<td></td>
<td></td>
<td>Most favorite</td>
</tr>
</tbody>
</table>

7. **Please write down everything you know about the U.S.S. Maine and/or the Spanish American War.**

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________
APPENDIX E: TEACHER QUESTIONNAIRE

Teacher Questionnaire- Prior to Study

These questions are to better understand how you think about and teach argument writing prior to the study. There are no right or wrong answers so please be open and honest in your responses.

Name*

1. What do you think makes a written argument effective?*

2. What do you think are the parts of an effective written argument? *

3. Please briefly describe how you have taught argument writing in your U.S. History classes this year.*

4. How many written arguments have students produced this year in your class?*
   Please also give some information about the length and requirements of the assignments.

5. What can you tell me about other classes where students might receive instruction in argument writing? *
   Any description of what they may be learning in these classes would also be helpful.

6. Please define a warrant in your own words. *

7. Have you taught students about warrants this year? If so, please describe what you have taught them.*
   If you have taught students about a similar concept but used a different term, please share that as well.

Please provide me with a few representative samples of student writing at high, medium and low levels.

Please provide me with three samples for each category (high, medium and low).
## APPENDIX F: SCHEME FOR WARRANTS, SUBTYPES AND FOILS

### Types of warrants (e.g. sourcing good/bad) and subtypes (e.g. close reading of word choice)

<table>
<thead>
<tr>
<th>Type 1</th>
<th>Sourcing</th>
<th>Subtypes of Sourcing good/bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>1y</td>
<td>Sourcing good</td>
<td>SoA. Time of event (___ happened before the ____), which means ___</td>
</tr>
<tr>
<td>1n</td>
<td>Sourcing bad</td>
<td>SoB. Strengths and limitations of that source type (This is a diary, which often means ___)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SoC. Perspective of that particular person (___ was a ____ which means he was likely to believe ___)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SoD. Combination with corrob: ‘While ____ said this, ____ said this)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type 2</th>
<th>Corroboration</th>
<th>Subtypes of Corroboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>2y</td>
<td>Corrob. good</td>
<td>CoA. Corroborates w/ source (vote counting) (Three sources said ___)</td>
</tr>
<tr>
<td>2n</td>
<td>Corrob. bad</td>
<td>CoB. Corroborates &amp; addresses contradictory information (while ___ said, A+B said ___)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CoC. Used in comb. with sourcing (While ____ was private doc, ____ was public, while ___ was speech to ___)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type 3</th>
<th>Close Read. Hist.</th>
<th>Subtypes of Close Reading-Historical</th>
</tr>
</thead>
<tbody>
<tr>
<td>3y</td>
<td>Cl. Read Hist.</td>
<td>CIA. close reading of word choice (the choice of ____ word is significant because ___)</td>
</tr>
<tr>
<td>3n</td>
<td>Cl. Read Hist.</td>
<td>CIB. close reading of argument (the claim that ____ means that ___)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CIC. close reading of word order/structure (the fact that he listed ____ first means ___)</td>
</tr>
</tbody>
</table>

### Other: Subtypes of Non-Historical

- NHy: Non-Hist. good
- NHn: Non-Hist. bad

### Subtypes of wrong answers

- 1. Inaccurate
- 2. Accurate but irrelevant (e.g. yes that is accurate, but doesn’t support claim)
- 3. Relevant but degree (e.g. over/under generalize based on evidence)
- 4. Misuse heuristic (e.g. thought doc closest in time to event always best)
- 5. Conflict w/other heuristic (e.g. sourced effectively, but didn’t corroborate - if they had would have seen this source was outlier)
APPENDIX G: WARRANT SELECTION AND RANKING TASK

Note: A sample item is included here. The entire form is accessible at http://goo.gl/forms/sJE7WVF7zz.

Warrant Selection and Ranking: The Explosion of the U.S.S. Maine

Directions: Your job is to select effective warrants, then rank the warrants you select in order from most to least effective. Warrants are statements in an argument that explain how the evidence supports the claim.

This assessment is split into four sections of three problems each. The claim changes each section, so pay careful attention to the claim so you can see whether or not the warrant effectively explains how the evidence supports the claim.

Each problem consists of two parts:

- In PART A, the evidence is listed, with a reminder of the claim. You click to select the effective warrants that effectively explain how the evidence supports the claim. There may be more than one right answer. DO NOT take into account whether or not YOU PERSONALLY agree, only judge how effective the warrants are.

- In PART B, rank ONLY the the effective warrants you selected in PART A from most effective to least effective.

* Required

1. First and Last Name *

2. Period *
   Please click to select your class period. Mark only one oval.
   
   - 4/5
   - 6/8
   - 9
   - 10

Section 1 Claim: YES, the explosion of the U.S.S. Maine DID CAUSE the U.S. to invade Cuba

Select effective warrants that explain how the evidence supports the claim, "Yes, the explosion of the U.S.S. Maine DID CAUSE the U.S. to invade Cuba". There may be more than one correct answer.
APPENDIX G (continued)

3. 1. Part A. Evidence: A popular song titled "Awake, United States" (Doc.C) was written soon after the sinking of the Maine. It includes the lines "Soldiers slain, by two-faced butchers paid by Spain" and "Awake! Thy Stars and Stripes Unfurl, and shot and shell and revenge hurt". 

Claim reminder: "Yes, the explosion of the U.S.S. Maine DID CAUSE the U.S. to invade Cuba". 
Check all that apply.

☐ A. Since this song was written just after the sinking of the Maine, it shows the public was eager for revenge. This probably influenced the decision to invade Cuba. 

☐ B. Doc. A (the New York Journal) and Doc. H (Textbook) also describe the angry response to the Maine. This shows "Awake, United States" wasn't just an isolated song. It was another sign of how upset the public was, which influenced the decision to invade Cuba.

☐ C. This song uses the word "slain", which makes it seem like U.S. soldiers were murdered on the Maine. Song lyrics don't exaggerate things. They give factual evidence of what actually happens in the world. This evidence that soldiers were murdered shows how the Maine explosion did cause the U.S. to invade Cuba.

☐ D. Document A (the New York Journal) also suggests Spain was to blame for the Maine explosion. Since a song and the New York Journal agree, this proves Spain was to blame. This caused the U.S. to invade Cuba.

☐ E. Popular songs usually make a lot of money. The fact that this song probably made a lot of money shows how the explosion of the U.S.S. Maine caused the U.S. to invade Cuba.

4. 1. Part B. Now, rank ONLY the items you selected above from most to least effective, beginning with 1, the MOST EFFECTIVE.

Warrant Writing Task

* Required

Name (First and Last) *

Period *
- 4/5
- 6/8
- 9
- 10

Directions

There are two paragraphs from two different historical argument essays about the explosion of the U.S.S. Maine below. Each paragraph has a claim and evidence, but it is missing a warrant. Finish each paragraph by writing a warrant in the space below each paragraph that explains how the evidence supports the claim. Be sure to read the claim and evidence carefully. It does not matter whether or not you personally agree with the argument. Your job is to write the best warrant you can. You have the remainder of the class period.

The explosion of the U.S.S. Maine did not cause the U.S. to invade Cuba. In his State of the Union Address (Doc D), President McKinley stated there was “serious injury” to the trade and business of Americans who were either working in Cuba, or who had businesses there. He said “reckless destruction” of property on the island had been occurring as the “fight waged for years” between the Spanish and the Cuban rebels. McKinley argued this had left many of these Americans with their “property destroyed and themselves ruined” (Doc D). *

Now write a warrant that explains how the evidence supports the claim.

The explosion of the U.S.S. Maine did not cause the U.S. to invade Cuba. The New York Times (Doc B) reported that Navy leader Captain Schuley felt the explosion was an “accident”, likely caused by a coal bunker fire that ignited the ship’s ammunition. However, The New York Journal (Doc A) stated that “all” naval officers thought the Maine was “purposely” destroyed by a Spanish mine. It says the “brutal” Spaniards may have triggered the explosion while the men were sleeping (Doc A). *

Now write a warrant that explains how the evidence supports the claim.
APPENDIX I: EXAMPLE INSTRUCTION SCRIPT AND GRAPHIC ORGANIZER FOR COMPARISON TREATMENT

A. Instruction Script for Comparison Treatment: 12/3, Day 4 of study

(Note: T= Teacher, Q= Question)

T: The previous two days we have been collecting evidence for the Question “Did the Explosion of the U.S.S. Maine Cause the U.S. to invade Cuba? We divided our evidence on a T-Chart. We wrote a response explaining which claim had the most evidence to support it.

Today we are going to be looking at a new document, from a new perspective. The question you are going to be answering is “SHOULD the U.S. have invaded Cuba following the explosion of the U.S.S. Maine?

T: Based on this question, what could our possible claims be?

● Yes, the U.S. SHOULD have invaded Cuba following the explosion of the U.S.S. Maine
● No, the U.S should NOT have invaded Cuba following the explosion of the U.S.S Maine

T: *As we read, we may refine our claim to adjust the certainly or degree to which we feel the U.S. should or should not have invaded Cuba (i.e. definitely should have, probably should have, absolutely should not have, etc.)

T: *However, for this portion of the study, we will simply think in terms of whether the U.S. SHOULD or SHOULD NOT have invaded Cuba

T: This question is trickier. It looks like it is asking for your opinion. But actually it is asking you to make an argument. When it says “should” it is not asking DID the explosion cause the U.S. to invade. It is asking SHOULD the U.S. have invaded.

T: You can believe that in fact the explosion DID NOT cause the U.S. to invade, but you think it SHOULD HAVE caused them to invade.

T: Or you can believe that the explosion DID cause the US to invade, but the US SHOULD NOT have invaded, that it was a bad thing.

Q: What other possibilities are there?

If you are writing a historical argument, either you can start with a question and find evidence determine what your claim should be, or you can generate your own claim based on evidence.

T: You can check whether or not your evidence fully supports my claim. Even though this question asks “SHOULD”, you still want to be calm, cool and collected as you weigh this evidence. You want to be like a judge, weighing the evidence, and deciding what side is strongest.
APPENDIX I (continued)

Just like we discussed on Monday, you are going to test to see if your evidence supports their claims. You want to make sure that you don’t over-generalize or ‘under-generalize’. You want to be sure you have enough evidence to prove your claim.

T: Today we will be reading a new document, from The New York Times, called “The Maine Disaster”. Source: New York Times, 17 February 1898, to see what evidence it gives us to address our question, “Should the U.S. have invaded Cuba following the explosion of the U.S.S. Maine?”

(I DO) T: Read the document through once, underlining anything you think is important and circling anything you think is confusing. Teacher models this with the first paragraph.

T: Now, I am going to read through a second time, and find anything you think might be evidence to help us address our research question.

Teacher models how to use the T-Chart organizer in the same way as yesterday

Teacher makes T chart on overhead

- Yes the U.S. should have invaded following the explosion of the U.S.S. Maine  |  No the U.S. should not have invaded Cuba following the explosion of the U.S.S. Maine.
- Teacher explains how there is a difference between evidence of “did” something cause something else, or ‘should” the U.S. have invaded.
- If you are looking for evidence of did, you are looking for a cause and effect relationship. Now it is no longer asking you to find evidence or whether or not the U.S.S. Maine caused the explosion. You are looking for evidence of whether the explosion SHOULD have caused the U.S. To in. If you are looking for “should” you are looking for justification of the decision to invade (reasons that that decision was the right or wrong thing to do).
- Teacher models one example. Teacher models a direct quote. Teacher mentions that evidence could also be a paraphrase.

(WE DO) T: Now let’s do one together.

- Teacher states claim again

  Q: Who has another piece of evidence? Calls on student. Student reads their evidence.

- Teacher asks students to explain what side of the argument that evidence supports.

(YOU DO) Teacher directs students they have rest of period to read and gather evidence on their T-Chart 10 minutes before the end of the lesson,

T: Turn to neighbor and share your evidence with them. First, give your neighbor feedback on their evidence. Is the evidence they chose evidence that says why they SHOULD or SHOULD NOT have invaded? Does the evidence give a specific quote or paraphrase? Is it clear what the evidence is saying, without going back to the passage?
APPENDIX I (continued)

T: Once you have done that, “Give one, Get one”. Give one piece of evidence to your neighbor that they did not have. Get one piece of evidence that you did not have. Add them to the T chart.

Exit Ticket: Based on the document you read today, which claim do you think has the most evidence to support it? Yes, the U.S. should have Invaded Cuba following the Explosion of the U.S.S. Maine, or No, the U.S. should not have invaded Cuba following the explosion of the U.S.S. Maine. Give evidence to support your position.

B. Graphic Organizer accompanying Example Lesson for Comparison Treatment

12-3, Day 4 of Study

Evidence Organizer

Question: SHOULD the U.S. have invaded Cuba following the explosion of the U.S.S. Maine?

1. Find evidence from the document that supports each claim. Remember, evidence can be direct quotes, facts and figures, or paraphrase of the text.

<table>
<thead>
<tr>
<th>Yes, the U.S. should have invaded Cuba following the explosion of the U.S.S. Maine</th>
<th>No, the U.S. should not have invaded Cuba following the explosion of the U.S.S. Maine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. After gathering the evidence, what claim has the most evidence to support it? Explain, using evidence to support your position.
A. Instruction Script for Treatment Condition: 12/3, Day 4 of study

(Note: T= Teacher, Q= Question)

● T: So today, we are going to be talking about writing arguments again, and focusing on warrants again, but THIS time, we are going to write claims, evidence and warrants that address a slightly DIFFERENT question.

● T: Instead of answering DID the explosion of the U.S.S. Maine cause the U.S. to invade Cuba, we’re going to be answering SHOULD the explosion have caused the U.S. to invade. So this is more of a value judgment, but we still want to keep our heads clear, and look for evidence to support whatever side we decide to argue.

● Q: But because it’s SHOULD, now the kind of evidence will be different. How will it be different?

Responses you are looking for:

○ Should, not looking for whether or not it caused it. Could have caused it or not. Now worried about SHOULD it have caused it.

○ Looking for if there’s evidence that the U.S. should have invaded. Was it the right thing to do? Was it reasonable? Did the evidence support that

● T: Today we are going to look at a different document. This is another news article. Whereas the one yesterday was from the “World”, a paper famous for exaggerated, over the top stories, this one is from The New York Times, called The Maine Disaster. (Source: New York Times, 17 February 1898). The New York Times is well known for more balanced, objective reporting. It is one of the world’s most respected newspapers and is still influential to this day. It is a newspaper that people trust.

● T: That’s right. Yesterday we focused on close reading of structure- what the author puts first or last, what words he chooses, and how that shows his perspective or how that will influence readers. Yesterday, we read an article where (describe article). That is only one possible focus for close reading. There are other areas as well:

○ Close reading of word choice

○ Close reading of argument

○ Close reading of word order/structure

T: Today we will be focusing first on the close reading of word choice

Have students get out their Close Reading handout. (Refer to the word choice section of the close reading handout. Point out the questions that accompany close reading for word choice.)
APPENDIX J (continued)

(I DO) T: I’m noticing the source, the New York Times, 17 February 1898. Just like our document yesterday, this document is written immediately after the U.S.S. Maine explosion. So I know that the New York Times is typically a more reliable source, and I know that it was written at exactly the same time as the New York World article, so it will be interesting to see how they are similar or different.

- T: So as I’m reading the document, I’m thinking about the question, and what my claim should be. I’m going to do my first read and mark what I thought was important and circled what I thought was confusing.
- T: When I read like a historian, one of the things I always do is pay attention to the specific words that the author chooses.
  - I ask myself, “What language does the author use to convince the reader (words, images, symbols, etc.)?”
  - How do the words they choose show the author’s perspective?

T: Read aloud the document, including the information at the top.

T: I’m noticing this quote: “Only theory as to the cause of the disaster.”

- T: In this case I underlined the word theory. Only **Theory** as to the cause of the disaster. Use of the word theory. A theory is like an explanation for something I am trying to understand, but it isn’t a fact. It’s implying that there is no hard evidence at this point.
- Q: How might the fact that this is a theory only support the idea of whether or not the US should invade Cuba?
  * (Possible Student Responses)
    - Shouldn’t invade, because not enough hard evidence
    - Only a theory

T: I’m going to read the rest of the document aloud. I’m not going to annotate, but I want you to underline what you think is important and circle what you think is confusing.

- T: I’ll model for you my thinking about how I can choose a claim and then write a warrant. So I’m thinking my claim should be “The explosion of the Maine **SHOULD NOT** have caused the U.S. to invade Cuba. (Model this on the overhead transparency-
  - I’m going to write my claim: “The U.S. should not have invaded Cuba following the explosion of the U.S.S. Maine. Then I’m going to write my evidence, “only theory as to the cause of the disaster”. Now I’m going to write my warrant, which explains how the evidence proves my claim.
    - T: Refer to sentence frames for close reading of word choice
      - I’m going to use my stems to help me structure my warrant:
        - the choice of ____ word/phrase is significant because it shows that ___
APPENDIX J (continued)

T: Write warrant: The choice of the word Theory is significant because it shows that there isn’t hard evidence as to what caused the explosion. Therefore, the U.S. shouldn’t invade Cuba since they don’t know if Spain was to blame.”

Does anyone have any questions about how I wrote this warrant using close reading?

(WE DO) Now with this same claim, what other evidence can we find through close reading of word choice? What other words in this document did the author use that could help us make the claim that the US should not invade Cuba?

-Students provide responses.

- Q: Who can use a sentence frame to write a warrant that explains how that word choice supports our claim that the U.S. should not invade Cuba?

_DISTRIBUTE Close reading warrant practice handout_

(YOU DO) T: Now you try It: Decide what side of the argument you want to argue: Yes the U.S. should have invaded Cuba, or no the U.S. should not have invaded Cuba... That will be your claim.

T: Then find a specific words or phrases in the text that the author chooses. Pick a word or phrase that reveals something important that supports your argument, or a word that shows what the author is trying to do by writing this, or how they feel about the event.

T: Finally, write a warrant that explains how the evidence supports the claim. Use the sentence frames from your close reading sheet to help you.

Exit ticket: When you are finished, write your claim, evidence and warrant and turn them in before the end of class.
B. Graphic Organizer accompanying Example Lesson for Treatment Condition

12-3, Day 4 of Study

Warrant Organizer

Now write your argument below!
APPENDIX K: THINK ALOUD PROTOCOL


Student will be audio and video recorded to capture what student says from over the shoulder to capture how they are using the documents without videotaping their face to ensure student privacy.

Think Aloud Protocol Analysis

1. First, the researcher will model what a think aloud entails, with an example related to writing, but not to the topic in question:

   a. I am going to give you an example of a think aloud. This is not an example of what you should say. It is an example of thinking aloud while you go. If I were going to do a think aloud about all the reasons I love the Cincinnati Bengals football team I might say,

      i. “Hmm so I’m thinking about all the reasons now. I think I’ll start with a topic sentence that gives an overview how much I love the Bengals. Though they are not the greatest team in the world, the Bengals have always been my favorite football team. Now I’m going to think about examples to show how I love them... What might be a good example...?"

2. Then the researcher will read the Think Aloud Interview Initial Instructions: “You read a series of historical sources concerning the explosion of the U.S.S. Maine and the U.S. invasion of Cuba. I have the sources here so you can refer to them.

5. Researcher will ask student to think aloud their answers to two questions on the warrant selection and ranking task.

   a. What are you thinking now? What are you doing first?
   b. What warrant(s) did you choose? Why did you choose this?
   c. Why did you not choose this warrant?
   d. How would you rank your warrants? Why would you rank them this way?

6. Researcher will ask student to think aloud as they write a warrant for warrant generation task

   a. What are you thinking now? What are you doing first?
   b. Why did you write this?
   c. Keep going
APPENDIX L: SEMI-STRUCTURED INTERVIEW PROTOCOL

Semi-structured Interview Questions

Following the think-aloud, the researcher will ask the following semi-structured interview questions about targeted parts of the student’s writing:

1. Do you think the explosion of the U.S.S. Maine caused the U.S. to invade Cuba? Why or why not?

2. Do you think the U.S. should have invaded Cuba following the explosion of the U.S.S. Maine? Why or why not?

3. What are warrants?

4. What makes a good warrant?

5. What role do warrants play in an argument?

6. What have you learned this week?

7. Tell me about the warrant selection task. How did it go?

8. Talk to me about the warrants that you wrote. How do you think you did?

9. How is this week similar or different from what you normally do in U.S. History class?

10. What would make this work easier? Would more time to practice or more feedback make it easier?
VIII. VITA

Ryan McCarty
Curriculum Vitae

HOME ADDRESS

149 Wendell Rd.
Shutesbury, MA 01072
H: (413) 256-6024
C: (617) 620-8645
rmccarty2875@gmail.com

EDUCATION

Fall 2006- University of Illinois-Chicago
Ph.D. Literacy, Language and Culture
  - Defense successfully completed; will graduate May 8, 2016
Literacy, Language and Culture, Master’s Program, 2006-2007

June 2001 National-Louis University
M.Ed. in Curriculum and Instruction
  - Creating Classroom Community in a Diverse Environment, teacher research project

June 1998 Indiana University, Bloomington
B.S. in Secondary Education
  - Certification: Type 09 Standard Secondary (Illinois)
    - Endorsements: English, Reading, U.S. History

UNIVERSITY EXPERIENCE

University of Illinois at Chicago

2011-2013 Graduate Assistant (Instructor of Record)

Courses Taught

CI 414, Middle and High School Literacy (Undergraduate-3 semesters)

CI504, Secondary Literacy (Graduate-3 semesters)
Research Projects

2014-2015  University of Illinois at Chicago, Chicago, IL
Principal Investigator: “Heuristics as warrants: Leveraging sourcing and corroboration heuristics as warrants in the historical argumentative writing of 11th grade students” (Design-Based Research)

2012-2013  University of Illinois at Chicago, Chicago, IL
Principal Investigator: A selective meta-analysis of historical argumentation (Meta-Analysis)

Spring 2013  University of Illinois at Chicago, Chicago IL. Research Consultant:
“Roles of engagement: Examining the use of role-playing controversy in the teaching and learning of argument writing among 9th grade students” (Quasi-Experimental/Mixed methods)

2008-2009  University of Illinois at Chicago, Chicago IL. Co-Principal Investigator (with Marilyn Johnston, University of Illinois Urbana-Champaign):
Integrating reading and writing workshop and inquiry in social studies for social justice (Collaborative Action Research)

Spring 2008  University of Illinois at Chicago, Chicago, IL. Research Assistant (PI: William Teale and Taffy Raphael, University of Illinois at Chicago):
In the hands of children: Innovative handheld technology for early literacy (Evaluation Research)

2007-2008  University of Illinois at Chicago, Chicago IL. Research Assistant and School Liaison (PI: Taffy Raphael, University of Illinois at Chicago):
Partnership READ: A standards-based reform school-university partnership in Chicago Public Schools. (Design-Based Research)


RESEARCH CONFERENCE PRESENTATIONS: REFEREED


SERVICE

Reviewer for the Literacy Research Association Annual Conference (2016)


Session Chair for Urban Education at the International Congress of Qualitative Inquiry (2010)

Reviewer for the World Congress on Reading (2008)

Reviewer for the National Reading Conference Annual Meeting (2008)

LITERACY LEADERSHIP EXPERIENCE (NONPROFIT)
2014- **Achievement Network**, Springfield MA. Director of School Support.

2013-2014 **Academy for Urban School Leadership** (AUSL), Chicago, IL. Literacy Coordinator.

2012-2013: **Academy for Urban School Leadership** (AUSL), Chicago, IL. High School Instructional Coach: Curriculum Team.


MIDDLE AND SECONDARY TEACHING

2009-2010: Revere Elementary, Chicago Public Schools, Chicago, IL. Middle School Reading Coach/Lead Literacy Teacher

2008-2009: Revere Elementary, Chicago Public Schools, Chicago, IL. 7th Grade Reading/7th and 8th Grade Social Studies Teacher

2006-2007: Willowbrook High School, DuPage Dist. 88, Villa Park, IL. 11th Grade Reading/ESL III & Advanced Reading Strategies Teacher

2005-2006: Hinsdale South High School, Hinsdale Dist. 86, Darien, IL. English II, English III, American Literature, Remedial English II (summer)


1998-2003: Park Jr. High School, La Grange Dist. 102, La Grange Park, IL. Regular and Inclusion Eighth Grade Reading/Language Arts

SELECTED CONSULTANT WORK

*Blogger*, The Teaching Channel. I write a well-received blog about Common Core literacy instruction. 2013-Present.

*Disciplinary Literacy: Preparing All Students for College and Career*. Presentation to the Center for Labor and Community Research, Chicago IL. July, 2012. Strategic planning work with Austin Coming Together community organization and presentation to CLCR to inform their work at Austin Polytechnical Institute.
Presentation and follow-up support to UIC Education Extension Faculty in aligning their existing coursework to Common Core State Standards.

ICTS Basic Skills (Teacher Licensure) test evaluation and website review, University of Illinois at Chicago. June, 2011
In-depth evaluation of the ICTS Basic Skills Practice Test and an analysis of ICTS preparation websites for the UIC College of Education. I also made recommendations for preparation materials for students.

Analysis of the Computer-Based IL300 ICTS Basic Skills (Teacher Licensure) Test, University of Illinois at Chicago, October, 2011.
Took actual ICTS Basic Skills tests and reported experience as a learner. Wrote up and shared insights about the experience with UIC faculty to support preparation of UIC students for teacher licensure.

**SELECTED COMMITTEES/LEADERSHIP**

Advanced Teacher Selection Committee, Springfield Empowerment Zone Partnership 2015

National Literacy Mindtrust, The Achievement Network 2014-

English Language Arts Team Lead, The Achievement Network 2014-

Assessment Committee Liaison, The Achievement Network 2014-

Instructional Leadership Team, Orr Academy HS 2012-2014

English Department Chair, Phillips Academy HS 2010-2011

Assessment Committee Member, Revere Elementary 2008-2009

Standards Based Change Committee, Revere Elementary 2008-2009

Diversity Committee, Willowbrook H.S. 2006-2007

Literacy Team Representative, Evanston Township H.S. 2003-2005

SELECTED HONORS AND ACHIEVEMENTS

UnboundED Standards Institute (Leadership Pathway) 2016
Emerging Education Policy Scholar, Thomas B. Fordham Institute And American Enterprise Institute 2015
Education Policy 101 Course-Recitation Cohort, Thomas B. Fordham Institute and 50 CAN 2015
Teachers College Reading and Writing Project Coaching Institute 2013
Tuition Waiver Recipient for PhD Program, UIC 2009-2010, Spring, 2011
Rochelle Lee Teacher Award Recipient 2009-2010
Oppenheimer Family Foundation Teacher Incentive Grant 2009
Who’s Who Among American Teachers 2005
Tenured Teacher Recognition Award, District 102 2003
Teacher Appreciation Honoree- Davis Memorial AME 2003
Larry Stilgebauer Technology Award- Exemplary Use 2003
Phi Delta Kappa Education Honorary 1998
Kappa Delta Pi Education Honorary 1998

ACADEMIC AWARDS/ACHIEVEMENTS

4.0 GPA, Ph.D. Program, University of Illinois at Chicago
4.0 Graduate GPA, National Louis University
3.5 Undergraduate GPA, Indiana University
Indiana University Dean’s List for High Academic Achievement, 1996-98
National Merit Commended Scholar, 1994
PROFESSIONAL ORGANIZATIONS

Literacy Research Association

International Literacy Association

National Council of Teachers of English

National Council of the Social Sciences
Approval Notice

Initial Review (Response to Modifications)

November 4, 2014

Ryan McCarty, MEd
Curriculum and Instruction
149 Wendell Rd
Shutesbury, MA 01072
Phone: (312) 413-7321 / Fax: (312) 996-7773

RE: Protocol # 2014-0953
“Heuristics as Warrants: Leveraging Sourcing and Corroboration Heuristics as Warrants in the Historical Argumentative Writing of 11th Grade Students”

Dear Mr. McCarty:

Your Initial Review application (Response to Modifications) was reviewed and approved by the Expedited review process on November 3, 2014. You may now begin your research.

Please note the following information about your approved research protocol:

Please remember to submit Spanish and Polish translations of all recruitment/consent documents that will be used to obtain permission from parents/guardians who do not speak English as a primary language. Translations must be accompanied by an Amendment form when submitted to the UIC IRB.

Approved Subject Enrollment #: 141

Additional Determinations for Research Involving Minors: The Board determined that this research satisfies 45CFR46.404, research not involving greater than minimal risk. Therefore, in accordance with 45CFR46.408, the IRB determined that only one parent’s/legal guardian’s permission/signature is needed. Wards of the State may not be enrolled unless the IRB grants specific approval and assures inclusion of additional
protections in the research required under 45CFR46.409. If you wish to enroll Wards of the State contact OPRS and refer to the tip sheet.

**Performance Sites:**
UIC, West and East Leyden High Schools, District 212 - IL

**Sponsor:**
None

**Research Protocol:**

a) Heuristics as Warrants: Leveraging Sourcing and Corroboration Heuristics as Warrants in the Historical Argumentative Writing of 11th Grade Students, Dissertation Proposal; Version 1.0; 09/15/2014

**Recruitment Materials:**

a) Teacher Recruitment Follow Up Document; Version 1; 10/23/2014
b) Parental Permission Cover Letter; Version 1.0; 10/24/2014
c) Recruitment Script for Student Subjects; Version 1; 10/24/2014
d) Email to Teacher; Version 2; 10/24/2014
e) Student Assent Cover Letter; Version 1.0; 10/24/2014

**Informed Consent:**

a) Teacher Consent Form; Version 2; 10/24/2014

**Assent:**

a) Student Assent Form; Version 2; 10/24/2014

**Parental Permission:**

a) Parental Permission Form; Version 2; 10/24/2014

Your research meets the criteria for expedited review as defined in 45 CFR 46.110(b)(1) under the following specific categories:

(5) Research involving materials (data, documents, records, or specimens) that have been collected, or will be collected solely for nonresearch purposes (such as medical treatment or diagnosis), (6) Collection of data from voice, video, digital, or image recordings made for research purposes, (7) Research on individual or group characteristics or behavior (including but not limited to research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

**Please note the Review History of this submission:**

<table>
<thead>
<tr>
<th>Receipt Date</th>
<th>Submission Type</th>
<th>Review Process</th>
<th>Review Date</th>
<th>Review Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/07/2014</td>
<td>Initial Review</td>
<td>Expedited</td>
<td>10/09/2014</td>
<td>Modifications Required</td>
</tr>
<tr>
<td>10/30/2014</td>
<td>Response To Modifications</td>
<td>Expedited</td>
<td>11/03/2014</td>
<td>Approved</td>
</tr>
</tbody>
</table>
Please remember to:

➔ Use your **research protocol number** (2014-0953) on any documents or correspondence with the IRB concerning your research protocol.

➔ Review and comply with all requirements on the enclosure:

"**UIC Investigator Responsibilities, Protection of Human Research Subjects**"

(http://tigger.uic.edu/depts/ovcr/research/protocolreview/irb/policies/0924.pdf)

Please note that the UIC IRB has the prerogative and authority to ask further questions, seek additional information, require further modifications, or monitor the conduct of your research and the consent process.

Please be aware that if the scope of work in the grant/project changes, the protocol must be amended and approved by the UIC IRB before the initiation of the change.

We wish you the best as you conduct your research. If you have any questions or need further help, please contact OPRS at (312) 996-1711 or me at (312) 996-2014. Please send any correspondence about this protocol to OPRS at 203 AOB, M/C 672.

Sincerely,

Sandra Costello
Assistant Director, IRB # 2
Office for the Protection of Research Subjects

Enclosures:

1. **UIC Investigator Responsibilities, Protection of Human Research Subjects**
2. **Informed Consent Document:**
   a) Teacher Consent Form; Version 2; 10/24/2014
3. **Assent Document:**
   a) Student Assent Form; Version 2; 10/24/2014
4. **Parental Permission:**
   a) Parental Permission Form; Version 2; 10/24/2014
5. **Recruiting Materials:**
   a) Teacher Recruitment Follow Up Document; Version 1; 10/23/2014
   b) Parental Permission Cover Letter; Version 1.0; 10/24/2014
   c) Recruitment Script for Student Subjects; Version 1; 10/24/2014
   d) Email to Teacher; Version 2; 10/24/2014
   e) Student Assent Cover Letter; Version 1.0; 10/24/2014
cc: Kimberly Lawless, Curriculum and Instruction, M/C 147  
Cynthia R. Shanahan (faculty advisor), Curriculum and Instruction, M/C 147

University of Illinois  
At Chicago

Office for the Protection of Research Subjects (OPRS)  
Office of the Vice Chancellor for Research (MC 672)  
203 Administrative Office Building  
1737 West Polk Street  
Chicago, Illinois 60612-7227

Approval Notice  
Continuing Review

November 2, 2015

Ryan McCarty, MEd  
Curriculum and Instruction  
149 Wendell Rd  
Shutesbury, MA 01072  
Phone: (312) 413-7321 / Fax: (312) 996-7773

RE: Protocol # 2014-0953  
“Heuristics as Warrants: Leveraging Sourcing and Corroboration Heuristics as Warrants in the Historical Argumentative Writing of 11th Grade Students”

Dear Mr. McCarty:

Your Continuing Review was reviewed and approved by the Expedited review process on October 30, 2015. You may now continue your research.

Please note the following information about your approved research protocol:

Approved Subject Enrollment #: 141 (data analysis from 89 subjects)  
Additional Determinations for Research Involving Minors: The Board determined that this research satisfies 45CFR46.404, research not involving greater than minimal risk.  
Performance Sites: UIC, West and East Leyden High Schools, District
Sponsor: None
PAF#: Not applicable

Research Protocol:

a) Heuristics as Warrants: Leveraging Sourcing and Corroboration Heuristics as Warrants in the Historical Argumentative Writing of 11th Grade Students, Dissertation Proposal; Version 1.0; 09/15/2014

Recruitment Material:

a) N/A – Limited to data analysis only

Informed Consent:

a) N/A – Limited to data analysis only

Your research meets the criteria for expedited review as defined in 45 CFR 46.110(b)(1) under the following specific categories:
(5) Research involving materials (data, documents, records, or specimens) that have been collected, or will be collected solely for nonresearch purposes (such as medical treatment or diagnosis).
(6) Collection of data from voice, video, digital, or image recordings made for research purposes.
(7) Research on individual or group characteristics or behavior (including but not limited to research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

Please note the Review History of this submission:

<table>
<thead>
<tr>
<th>Receipt Date</th>
<th>Submission Type</th>
<th>Review Process</th>
<th>Review Date</th>
<th>Review Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/21/2015</td>
<td>Continuing Review</td>
<td>Expedited</td>
<td>10/30/2015</td>
<td>Approved</td>
</tr>
</tbody>
</table>

Please remember to:

→ Use your research protocol number (2014-0953) on any documents or correspondence with the IRB concerning your research protocol.

→ Review and comply with all requirements on the OPRS website under: "UIC Investigator Responsibilities, Protection of Human Research Subjects" ([http://tigger.uic.edu/depts/ovcr/research/protocolreview/irb/policies/0924.pdf](http://tigger.uic.edu/depts/ovcr/research/protocolreview/irb/policies/0924.pdf))

Please note that the UIC IRB has the right to seek additional information, require further modifications, or monitor the conduct of your research and the consent process.

Please be aware that if the scope of work in the grant/project changes, the protocol must be amended and approved by the UIC IRB before the initiation of the change.
We wish you the best as you conduct your research. If you have any questions or need further help, please contact OPRS at (312) 996-1711 or me at (312) 996-9299. Please send any correspondence about this protocol to OPRS at 203 AOB, M/C 672.

Sincerely,

Anna Bernadska, M.A.
IRB Coordinator, IRB # 2
Office for the Protection of Research

Subjects

Enclosure: None

cc: Danny B. Martin, Curriculum and Instruction, M/C 147
    Cynthia R. Shanahan, Faculty Sponsor, Curriculum and Instruction, M/C 147