Talking Physics: Two Case Studies on Short Answers and Self-explanation in Learning Physics

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By
Ryan C. Badeau, M.S.
Graduate Program in Physics

The Ohio State University
2017

Dissertation Committee:
Andrew F. Heckler, Advisor
Lei Bao
Richard Furnstahl
Lin Ding
This thesis explores two case studies into the use of short answers and self-explanation to improve student learning in physics. The first set of experiments focuses on the role of short answer questions in the context of computer-based instruction. Through a series of six experiments, we compare and evaluate the performance of computer-assessed short answer questions versus multiple choice for training conceptual topics in physics, controlling for feedback between the two formats. In addition to finding overall similar improvements on subsequent student performance and retention, we identify unique differences in how students interact with the treatments in terms of time spent on feedback and performance on follow-up short answer assessment. In addition, we identify interactions between the level of interactivity of the training, question format, and student attitudinal views of the respective trainings. The second case study focuses on the use of worked examples in the context of multi-concept physics problems – which we call “synthesis problems.” For this part of the thesis, four experiments were designed to evaluate the effectiveness of two instructional methods employing worked examples on student performance with synthesis problems; these instructional techniques, analogical comparison and self-explanation, have previously been studied primarily in the context of single-concept problems. As such, the work presented here represents a novel focus on extending these two techniques to this class of more complicated physics problem. Across the four experiments, both self-explanation and certain kinds of analogical comparison of worked examples significantly improved student performance on a target synthesis problem, with distinct improvements in recognition of the relevant concepts. More specifically, analogical comparison significantly improved student performance when the comparisons were invoked between worked synthesis exam-
ples. In contrast, similar comparisons between corresponding pairs of worked single-concept examples did not significantly improve performance. On a more complicated synthesis problem, self-explanation was significantly more effective than analogical comparison, potentially due to differences in how successfully students encoded the full structure of the worked examples. Finally, we find that the two techniques can be combined for additional benefit, with the trade-off of slightly more time-on-task.
For Rebecca & Emma.
Acknowledgments

First and foremost, I would like to thank my wife, Rebecca, and our daughter, Emma. Your love and patience have been the bedrock for all of my endeavors. Thank you, Rebecca, for being my sounding-board, my encourager, and my all-around better half.

I thank my family for their love and support. I thank my parents, John and Sheryl, for the inspiration to not only pursue my interests but for their unending encouragement in that pursuit. To my brother, Nick: thank you for being an advocate, confidant, and constant supporter.

I would like to thank Andrew Heckler for serving not only as my advisor, but as a true mentor. I am incredibly grateful for his thoughtful guidance and feedback. I am similarly grateful to Lin Ding, both for his research advice and guidance. In addition, I’d like to thank my other committee members, Richard Furnstahl and Lei Bao, for their own advice and feedback throughout the course of my research.

I would also like to give particular thanks to my collaborators from outside OSU. I am grateful to Andrew Pawl for his insights and contributions with training questions and program structure for several of the studies involving computer-based training. In addition, I would like to thank Richard Catrambone for illuminating conversations in task-analysis. I am grateful to the National Science Foundation for their support (grant no. DRL–1252399).

Finally, I would like to thank my fellow OSU physics education graduate students and post-docs: Daniel White, Nathaniel Amos, Abby Bogdan, Brendon Mikula, Chris Porter, and Bashirah Ibrahim. Thank you for your advice, humor, interest and assistance throughout my research. I would not have learned as much during the last several years without you - nor would I have enjoyed it nearly as much.
VITA

2009 ......................................................... B.S., Physics and Philosophy
    Summa Cum Laude
    Rensselaer Polytechnic Institute
    Troy, NY

2012 ......................................................... M.S., Physics
    Cornell University
    Ithaca, NY

2012-2013 .................................................. Graduate Teaching Assistant
    The Ohio State University
    Columbus, OH

2013-2014 .................................................. Metro Fellowship
    The Ohio State University
    Columbus, OH

2014-Present .............................................. Graduate Research Assistant
    The Ohio State University
    Columbus, OH

Publications


Fields of Study

Major Field: Physics
# Table of Contents

<table>
<thead>
<tr>
<th>Chapters</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I The Effectiveness of Short Answer Questions in Computer-Based Training of Physics Concepts</strong></td>
<td></td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Let’s Talk: Why Invoke Student Short Answers &amp; Explanations?</td>
<td>1</td>
</tr>
<tr>
<td>1.2 A Tale of Two Contexts: Computer-based Instruction and Synthesis Problem Solving</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Thesis Organization</td>
<td>6</td>
</tr>
<tr>
<td><strong>2 Theoretical Background</strong></td>
<td>8</td>
</tr>
<tr>
<td>2.1 Short Answer Formats &amp; Computer-based Instruction</td>
<td>9</td>
</tr>
<tr>
<td>2.2 The Self-Explanation Effect</td>
<td>18</td>
</tr>
<tr>
<td>2.3 Memory &amp; Testing Effects of Short Answer Formats</td>
<td>20</td>
</tr>
<tr>
<td>2.4 Summary</td>
<td>22</td>
</tr>
<tr>
<td><strong>3 Methodology</strong></td>
<td>24</td>
</tr>
<tr>
<td>3.1 Research Questions</td>
<td>24</td>
</tr>
<tr>
<td>3.2 Design of Training Software</td>
<td>29</td>
</tr>
<tr>
<td>3.3 Natural Language Implementation</td>
<td>33</td>
</tr>
<tr>
<td>3.4 Training Questions &amp; Format</td>
<td>35</td>
</tr>
<tr>
<td><strong>4 Study of Short Answer Format &amp; Computer-Based Training in Physics</strong></td>
<td>36</td>
</tr>
<tr>
<td>4.1 Experiment #1: Pilot of Short Answer vs. Multiple Choice (Force &amp; Motion)</td>
<td>36</td>
</tr>
<tr>
<td>4.1.1 Design and Participants</td>
<td>36</td>
</tr>
<tr>
<td>4.1.2 Results</td>
<td>37</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>4.1.3</td>
<td>Discussion</td>
</tr>
<tr>
<td>4.2</td>
<td>Experiment #2: Short Answer vs. Multiple Choice</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Results</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Discussion</td>
</tr>
<tr>
<td>4.3</td>
<td>Experiments #3 and #4: Effects of Follow-up Questions</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Design and Participants</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Results</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Discussion</td>
</tr>
<tr>
<td>4.4</td>
<td>Experiment #5: Effects of Question Variety</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Design and Participants</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Results</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Discussion</td>
</tr>
<tr>
<td>4.5</td>
<td>Experiment #6: Short Answer vs. Multiple Choice (Electric Field &amp; Potential)</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Design and Participants</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Results</td>
</tr>
<tr>
<td>4.5.3</td>
<td>Discussion</td>
</tr>
<tr>
<td>4.6</td>
<td>Summary</td>
</tr>
</tbody>
</table>

II What Works with Worked Examples: Extending Analogical Comparison and Self-Explanation to Synthesis Problems

5 Theoretical Background

5.1 Synthesis Problems                                             106
5.2 Worked Examples                                                107
5.3 Analogical Comparison                                          108
5.4 The Return of Self-Explanation                                 109

6 What Works with Worked Examples & Synthesis Problems?

6.1 Experiment #1: Analogical Comparison - Effects of Type of Worked Examples  113
  6.1.1 Design and Participants                                     113
  6.1.2 Results                                                     117
  6.1.3 Discussion                                                  121
6.2 Experiment #2: Analogical Comparison when Recognition isn’t a Bottleneck  123
  6.2.1 Design and Participants                                     123
  6.2.2 Results                                                     127
  6.2.3 Discussion                                                  132
6.3 Experiment #3: Analogical Comparison vs. Self-Explanation       134
  6.3.1 Design and Participants                                     134
  6.3.2 Results                                                     138
  6.3.3 Discussion                                                  144
6.4 Experiment #4: Analogical Comparison using Self-Explanation     147
  6.4.1 Design and Participants                                     147
  6.4.2 Results                                                     149
  6.4.3 Discussion                                                  156
6.5 Summary                                                         158
7 Conclusion

7.1 Case Study #1: Short Answers in Computer-Based Instruction

7.1.1 Research Questions

7.1.2 Instructional Implications

7.1.3 Limitations and Future Work

7.2 Case Study #2: Worked Examples & Synthesis Problem Solving

7.2.1 Research Questions

7.2.2 Instructional Implications

7.2.3 Limitations and Future Work

Bibliography

Appendices

A Computer-based Training and Test Materials

A.1 FVA Test Questions

A.2 Experiment #1: Force & Motion Training Questions

A.3 Experiments #2-4: Force & Motion Training Questions (Multiple Choice Versions, Initial Question Only)

A.4 Experiment #5: Training Questions (Varied-type)

A.5 Experiment #6: Voltage & Potential Questions

B Synthesis Problem Solving Materials

B.1 Experiment #1 Training Materials

B.2 Experiment #2 Training Materials

B.3 Experiment #3 Training Materials

B.4 Experiment #4 Training Materials
List of Figures

Figure | Page
--- | ---
3.1 Schematic illustration of the six short answer experiments. | 25
3.2 Examples of the training interface. | 30
3.3 Example question, student response, and provided feedback from both question formats. | 31
4.1 Experiment #1: Student post-test scores | 38
4.2 Experiment #1: Student post-test scores by median grade split | 39
4.3 Experiment #1: Distributions of training efficiency | 41
4.4 Experiment #1: Effect of false positive feedback during short answer training | 43
4.5 Experiment #1: Proportion of correct responses on training questions | 44
4.6 Experiment #2: Example of an explanation-focused training question | 48
4.7 Experiment #2: Student post-test scores | 49
4.8 Experiment #2: Student post-test scores by median grade split | 50
4.9 Experiment #2: Distributions of training efficiency | 52
4.10 Experiment #2: Effect of false-positive and false-negative feedback during short answer training | 54
4.11 Experiment #2: Proportion of correct responses on training questions | 56
4.12 Experiment #3-4: Illustration of follow-up feedback structure. | 62
4.13 Experiment #3-4: Experiment administration and task sequence | 64
4.14 Experiment #3-4: Mean gain in pre-test to post-test subscore. | 66
4.15 Experiment #3-4: Retention scores. | 69
4.16 Experiment #3-4: Student performance on short answer questions by condition | 73
4.17 Experiment #3-4: Time spent on feedback for each training question by condition. | 77
4.18 Experiment #5: Schematic illustration of training structure and question variations. | 85
4.19 Experiment #5: Experiment administration and task sequence | 87
4.20 Experiment #5: Mean gain in pre-test to post-test subscore. | 89
4.21 Experiment #5: Time spent on feedback for each training question by condition. | 91
4.22 Experiment #6: Student multiple choice post-test scores | 96
4.23 Experiment #6: Student short answer post-test scores | 97
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.24</td>
<td>Experiment #6: Student short answer post-test scores by question</td>
<td>99</td>
</tr>
<tr>
<td>6.1</td>
<td>Schematic illustration of the four experimental designs included in the synthesis study</td>
<td>112</td>
</tr>
<tr>
<td>6.2</td>
<td>Experiment #1: Target synthesis problem and sample worked example problems</td>
<td>114</td>
</tr>
<tr>
<td>6.3</td>
<td>Experiment #1: Experiment Design</td>
<td>115</td>
</tr>
<tr>
<td>6.4</td>
<td>Experiment #1: Distributions of student scores on the target synthesis problem</td>
<td>119</td>
</tr>
<tr>
<td>6.5</td>
<td>Experiment #1: Proportion of students demonstrating recognition of centripetal acceleration and energy conservation</td>
<td>120</td>
</tr>
<tr>
<td>6.6</td>
<td>Experiment #2: Target synthesis problem and sample worked example problems</td>
<td>124</td>
</tr>
<tr>
<td>6.7</td>
<td>Experiment #2: Experiment Design</td>
<td>125</td>
</tr>
<tr>
<td>6.8</td>
<td>Experiment #2: Proportion of students including related item within guide prompt</td>
<td>131</td>
</tr>
<tr>
<td>6.9</td>
<td>Experiment #3: Target synthesis problem and sample worked example problems</td>
<td>135</td>
</tr>
<tr>
<td>6.10</td>
<td>Experiment #3: Experiment Design</td>
<td>137</td>
</tr>
<tr>
<td>6.11</td>
<td>Experiment #3: Proportion of students recognizing physical concepts</td>
<td>140</td>
</tr>
<tr>
<td>6.12</td>
<td>Experiment #3: Proportion of students with correct solution structure on target synthesis problem</td>
<td>142</td>
</tr>
<tr>
<td>6.13</td>
<td>Experiment #4: Experiment Design</td>
<td>148</td>
</tr>
<tr>
<td>6.14</td>
<td>Experiment #4: Proportion of students recognizing physical concepts</td>
<td>152</td>
</tr>
<tr>
<td>6.15</td>
<td>Experiment #4: Proportion of students with correct solution structure on target synthesis problem</td>
<td>153</td>
</tr>
<tr>
<td>6.16</td>
<td>Experiment #4: Time spent on training task by condition</td>
<td>153</td>
</tr>
<tr>
<td>B.1</td>
<td>Experiment #1: Synthesis worked examples</td>
<td>204</td>
</tr>
<tr>
<td>B.2</td>
<td>Experiment #1: Single concept worked examples</td>
<td>205</td>
</tr>
<tr>
<td>B.3</td>
<td>Experiment #1: Mastery &amp; recognition prompts</td>
<td>206</td>
</tr>
<tr>
<td>B.4</td>
<td>Experiment #2: Synthesis worked examples - same variable version</td>
<td>207</td>
</tr>
<tr>
<td>B.5</td>
<td>Experiment #2: Synthesis worked examples - switched variable version</td>
<td>208</td>
</tr>
<tr>
<td>B.6</td>
<td>Experiment #2: Single concept worked examples - same variable version</td>
<td>209</td>
</tr>
<tr>
<td>B.7</td>
<td>Experiment #2: Single concept worked examples - switched variable version</td>
<td>210</td>
</tr>
<tr>
<td>B.8</td>
<td>Experiment #2: Comparison prompts</td>
<td>211</td>
</tr>
<tr>
<td>B.9</td>
<td>Experiment #3: Synthesis worked examples</td>
<td>212</td>
</tr>
<tr>
<td>B.10</td>
<td>Experiment #3: Single concept worked examples</td>
<td>213</td>
</tr>
<tr>
<td>B.11</td>
<td>Experiment #3: Mastery &amp; recognition prompts</td>
<td>214</td>
</tr>
<tr>
<td>B.12</td>
<td>Experiment #3: Combined prompts</td>
<td>215</td>
</tr>
<tr>
<td>B.13</td>
<td>Experiment #4: Presentation of worked example annotations</td>
<td>216</td>
</tr>
<tr>
<td>B.14</td>
<td>Experiment #4: Combined prompts</td>
<td>217</td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Experiment #3-4: Number of students per experimental condition for both</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>experiment administrations (algebra-based and calculus-based).</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>Experiments #3-4: Mean multiple choice post-test score and overtraining</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>score by condition.</td>
<td></td>
</tr>
<tr>
<td>4.3</td>
<td>Experiment #3-4: Overall short answer score by condition (algebra-based</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>and calculus-based).</td>
<td></td>
</tr>
<tr>
<td>4.4</td>
<td>Experiment #3-4: Mean and median time spent (in minutes) on training and</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>corresponding efficiency for each training condition.</td>
<td></td>
</tr>
<tr>
<td>4.5</td>
<td>Experiment #3-4: Student responses to attitudinal exit survey.</td>
<td>80</td>
</tr>
<tr>
<td>4.6</td>
<td>Experiment #5: Mean multiple choice post-test score and over-training score</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>by condition.</td>
<td></td>
</tr>
<tr>
<td>4.7</td>
<td>Experiment #5: Mean and median time spent (in minutes) on training in total</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>and for each session.</td>
<td></td>
</tr>
<tr>
<td>4.8</td>
<td>Experiment #6: Student responses to attitudinal exit survey.</td>
<td>101</td>
</tr>
<tr>
<td>6.1</td>
<td>Experiment #1: Scoring rubric for target synthesis problem.</td>
<td>116</td>
</tr>
<tr>
<td>6.2</td>
<td>Experiment #1: Mean score on target synthesis problem.</td>
<td>118</td>
</tr>
<tr>
<td>6.3</td>
<td>Experiment #2: Scoring rubric for target synthesis problem.</td>
<td>126</td>
</tr>
<tr>
<td>6.4</td>
<td>Experiment #2: Mean score on target synthesis problem.</td>
<td>127</td>
</tr>
<tr>
<td>6.5</td>
<td>Experiment #2: Proportion of students successfully accomplishing key</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td>synthesis problem-solving steps.</td>
<td></td>
</tr>
<tr>
<td>6.6</td>
<td>Experiment #2: Proportion of students demonstrating several specific errors</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>on the target synthesis problem.</td>
<td></td>
</tr>
<tr>
<td>6.7</td>
<td>Experiment #3: Scoring rubric for target synthesis problem.</td>
<td>138</td>
</tr>
<tr>
<td>6.8</td>
<td>Experiment #3: Mean score on target synthesis problem.</td>
<td>139</td>
</tr>
<tr>
<td>6.9</td>
<td>Experiment #4: Mean score on target synthesis problem.</td>
<td>150</td>
</tr>
</tbody>
</table>
Chapter 1
INTRODUCTION

This thesis represents two distinct areas of research. In Volume I, we report on experiments that investigate the relative effectiveness of short answer questions combined with immediate feedback in the context of computer-based training of physics concepts. Volume II examines a series of studies with worked examples that measure and compare the effectiveness of analogical comparison and self-explanation in order to improve student performance on multi-concept physics problems.

1.1 Let’s Talk: Why Invoke Student Short Answers & Explanations?

“What I cannot create, I do not understand.”

Those words were famously etched in chalk on Richard Feynman’s blackboard at the time of his death. Given that the quote was followed by an invocation to “Know how to solve every problem that has been solved,” Feynman’s intention seems clear: to understand something, one needs to be able to recreate it. It is not enough to simply be told the answer; understanding implicitly involves individualized construction.

This very idea has been traditionally formalized in terms of a constructivist theory of learning, as developed by Piaget, Vygotsky, and others (Piaget, 1950; Vygotsky, 1978). Constructivism emphasizes the role of the learner in building their own knowledge through a continual process of revision, connecting existing knowledge and conceptual schemas to new knowledge and experiences. In contrast to the transmissionist view – which asserts that
knowledge is something that can be transferred in independent, concise chunks from teacher to learner – the constructivist view argues that understanding is developed by the learner and relies heavily on previous knowledge. As such, there is no easy way out - learning demands active engagement on the part of the learner.

In that spirit, the following two case studies included in this thesis examine the use of short answer formats designed to place the impetus on the student to recall physics-related concepts and explain physics examples. The two areas of research are distinct in context and implementation; consequently, each case study will focus on its own set of core research questions. However, taken together the case studies presented here can be thought of as two instantiations of a single broad but underlying goal: to develop and assess effective treatments that provide opportunities for students to talk about physics concepts – to explain them in their own words – in the hope of subsequently improving student conceptual understanding and problem solving.

In the subsequent sections, we briefly review more specific, guiding motivations for each case study and outline the overall organization of the thesis.

1.2 A Tale of Two Contexts: Computer-based Instruction and Synthesis Problem Solving

(a) Short Answers in Computer-based Instruction

The first case study in this thesis focuses on the role of short answer questions during computer-based training of physics conceptual knowledge. More specifically, this series of experiments compared short answer and multiple choice training formats under conditions where the feedback provided to the students was identical. In order to conduct this comparison, we developed short answer computer-based training software to automatically assess students’ natural language responses and corresponding training materials focused on two conceptual topics in introductory physics: 1) one-dimension force and motion and 2) electric field and potential.

This comparison of short answer and multiple choice formats was motivated by both
practical and theoretical factors. From a practical standpoint, short answer questions rep-resent a potentially powerful and distinct question format for use in the development of instructional software. For example, short answer questions may be particularly useful in computer-based training programs focused on student practice with multiple representations, such as the essential skills framework (Mikula and Heckler, 2017a). Natural language would provide another representation to supplement more typical multiple-choice, numerical calculation, and graphical question formats. However, there are important empirical questions regarding the implementation of short answers in such a setting. In particular, are there measurable attitudinal differences depending on the format? Do students find the short answer format useful or frustrating? Is the short answer format less effective than the multiple-choice counterpart given its corresponding trade-offs (imperfect answer interpretation and thus possibly incorrect feedback to the student)?

From a theoretical standpoint, previous research suggests multiple motivations for this investigation. First, prior work has shown that intelligent computer tutors based on natural language can be very effective instructional tools (Graesser et al., 2005; Rus et al., 2016). Second, previous research has demonstrated that the very act of self-explanation – that is, having the student actively explain key steps or concepts in their own words – can be an effective instructional tool (Chi et al., 1989; Chi, 2000). Moreover, there is some research suggesting that self-explanation can still be effective when terms are selected (via a glossary) rather than spontaneously generated by the student (Aleven and Koedinger, 2002). Finally, research into testing effects have demonstrated that differences in the cognitive processing necessary to answer a question (i.e. retrieval vs. recognition) can potentially influence the retention of the related material (Smith and Karpicke, 2014; Kang et al., 2007).

Our studies complement this prior work by explicitly comparing two stem-equivalent question formats with the same feedback (short answer vs. multiple choice) that not only control for content, but also the interactivity of the training. Motivated by the above, we consider four research questions:

- Controlling for feedback and content, are short answer computer-based training for-
mats more effective than multiple choice counterparts (in terms of overall student performance and efficiency)?

- Controlling for feedback and content, does short answer computer-based training provide retention gains compared to multiple choice?

- Does an increase in the interactivity of feedback during training improve the relative performance of short answer formats?

- Does an increase in the variety of questions during training improve the relative performance of short answer formats?

(b) Worked Examples & Synthesis Problem Solving

The second case study presented in this thesis represents a novel attempt to extend previously developed instructional interventions using worked examples to a specific subclass of physics problem – synthesis problems – in order to improve subsequent student performance on problem solving. In contrast to traditional single-concept textbook examples, synthesis problems are multi-concept physics problems, requiring the application of more than one major physics concept, often from disparate parts of the teaching timeline (Ding et al., 2011). As a result, synthesis problems provide unique difficulties for students in the recognition and application of multiple concepts (Ding et al., 2010, 2009; White et al., 2014). Not only do students need to be able to recognize and correctly apply the individual physical concepts, they need to figure out how to combine them - both conceptually and mathematically - in order to arrive at a solution.

Given the unique challenges represented by synthesis problems, there are several important and currently unknown questions regarding how best to scaffold student use of worked examples in order to improve subsequent problem solving ability. For example, in terms of the provided worked examples themselves, is it better to have students study sets of component single-concept examples or full-fledged examples of synthesis problems? On the one hand, the former would represent a significant reduction in cognitive load and examples more akin to those they have traditionally seen in class; on the other hand, the latter would
imply a higher degree of structural similarity to the target synthesis problem we would like students to be able to solve.

In addition, what kind of questions should we ask to assist students in extracting the structure of worked examples? Previous work has suggested that open-ended prompts for self-explanation can be useful for studying worked examples, but the vast majority of that work has been done in the context of single-concept problems (Chi et al., 1989; Atkinson et al., 2003). The effectiveness of such an approach may be limited by the increased complexity characteristic of synthesis problems.

Given that previous studies have identified recognition as a key bottleneck for synthesis-problems (White et al., 2014), we hypothesized that guided prompts invoking analogical comparisons between pairs of worked examples might assist students in extracting the underlying structure of a target synthesis problem. Rather than focus on surface features of a single example, scaffolded comparisons of two base examples may help students to hone in on common facets in the joint application of those concepts – effectively averaging out surface feature “noise” to better understand the underlying solution structure. These techniques have been used successfully in other domains – to help learners understand business negotiation techniques, for example (Gentner et al., 2003) – but not in the context of multi-concept physics problems.

As such, our experiments in the context of synthesis problems represent a novel investigation comparing the relative effectiveness of different treatments with worked examples. In that pursuit, we investigate three main research questions:

- Given the increased complexity of synthesis problems, is it more effective to invoke comparisons between worked examples that break down the target synthesis problem into its single-concept parts, or worked examples that include the concepts in combination?

- How does the focus of the prompts influence analogical comparison (i.e. prompts involving holistic structure vs. prompts for fine-grain applications of the individual concepts)?
• How does analogical comparison across a pair of worked examples compare with self-explanation of each worked example independently?

1.3 Thesis Organization

As previously described, the thesis is divided into two volumes. Over the course of three chapters, Volume I details our case study into the effectiveness of short answer questions in the context of computer-based training in physics. To this, Chapter 2 provides a representative overview of the existing literature. Chapter 3 describes the overall methodology of the related experiments, including a description of the research goals of each particular study and the overall design of our computer-based training software. Chapter 4 includes the design, results, and discussion of six different experiments with short answer questions formats.

The second volume of the thesis includes Chapter 5 & 6 and focuses on the use of worked examples in improving student performance on synthesis problems. Chapter 5 provides further background research specifically related to worked examples and techniques such as analogical comparison that have been previously used to invoke effective student discussion and explanation in other learning domains. Chapter 6 includes the design, results, and discussion of our four experimental investigations.

Finally, Chapter 7 discusses several general conclusions from the two overall research areas, their pedagogical implications, and opportunities for future research.
Volume I

The Effectiveness of Short Answer Questions in Computer-Based Training of Physics Concepts
Chapter 2

Theoretical Background

The first volume of this thesis is devoted to the study of short answer, free-response question formats in the context of computer-based instruction in physics. As such, the goal of this chapter is to provide a brief overview of related research and establish a foundation for discussion of the studies presented in the subsequent chapters.

Although no universal definition exists, short answer questions are typically defined by the type of response invoked, namely a response that recalls knowledge not stated explicitly in the question, given in natural language, with a length that can vary between a short, few-word phrase to several sentences (Burrows et al., 2015). Perhaps not surprisingly, research studying the effectiveness and instructional implications of short answer questions has a long history, extended across multiple domains and methods of inquiry. For the purposes of this overview, we have separated the review into three relevant research trajectories: computer-based tutors, studies of the self-explanation effect, and educational assessment research. In some cases these research findings are intricately intertwined (for instance, self-explanation has been directly implemented in several computer-based tutors); others, such as comparisons of retention effects due to short-answer and multiple choice questions, are most often discussed in contexts relating to testing effects and memory retrieval. To this, the intention of this overview is to tour these different research communities and construct useful bridges between them to motivate and frame the subsequent studies. Finally, although it will not be a focus of this review or the experiments discussed here, it is worthwhile to note that there is a vast and active computer science literature specifically devoted to the study
and improvement of natural language processing techniques in support of the short-answer format (Burrows et al., 2015).

2.1 Short Answer Formats & Computer-based Instruction

Research into the development of computer-based instruction systems dates back to the 1960’s (Ma et al., 2014; VanLehn, 2011). Since then, vast and varied classes of computer tutors have been developed from different cognitive frameworks with different methods and levels of student-computer interaction. Consequently, it is useful to separate this diverse field into two rough categories commonly used in the literature: computer-aided instruction systems (also referred to as computer-based training or computer assisted learning systems) and intelligent tutoring systems. Although the two categories often share overlapping elements, distinctions can be drawn between them based on the information gathered from the student, the manner in which they model the students performance on the task, and whether they adapt instruction or feedback according to the information gathered (Ma et al., 2014).

As such, the following sections provide a brief overview of these general classes of computer-based instruction, important findings, and in turn, discuss specific results related to the study of short answer questions within these two classes of computer-based instruction.

(a) Computer-based training systems

On the one hand, computer-based training systems tend to be answer-based, follow a preset curriculum or question script, and provide no adaptation of instruction or feedback. Online homework systems such as Mastering Physics traditionally fall under this category. On the other hand, intelligent tutoring systems tend to gather more detailed information about a student’s thought process and method when solving a problem (as compared to simple answer-based methods), model student performance or identify a student’s state in comparison to an expert model of the subject domain, and adapt tutoring functions
Accordingly through variations in feedback or the selection of subsequent questions.

As a result, computer-based training systems tend to be more common than their intelligent tutor counterparts. Consequently, these systems have been the focus of a massive amount of research in the past four decades - so vast in fact, that sufficient summation of the state of the field is increasingly represented through second-order meta-analyses (Tamim et al., 2011; Kulik and Kulik, 1991). For example, the second-order meta-analysis conducted by Tamim et al. (2011) included over 25 meta-analyses encompassing over 1,055 primary studies in order to evaluate the effect of implementing computer-based instruction in a classroom setting. The take-away message? On average, computer-based instruction represented an effect size of 0.35 over non-technological based instruction methods (Tamim et al., 2011). The analysis by Kulik and Kulik (1991) came to a similar result (an effect size of 0.30).

In addition to overall effectiveness, there have been several (mostly) generalizable findings. For example, computer-based training tends to improve student attitudes towards course instruction and reduces the amount of time students need for learning (Kulik et al., 1983). In addition, computer-based training implementations that provide instructional feedback typically demonstrate higher effect sizes, around 0.5 (Hattie and Timperley, 2007). However, despite this vast background and key findings, there are still important remaining questions. In particular, the prevalence of computers in education has shifted research towards assessment of relative benefits due to different types of computer-based training, rather than comparisons between technology-based and non-technological settings (Tamim et al., 2011).

(b) Examples of Short Answer Implementations in Computer-based training

There have been a comparatively small number of studies in the sciences that have explored the use of short-answer formats in these contexts. In particular, within physics, Nakamura et al. (2016) studied short answers using an interactive and multimedia-based tutoring system - the Pathway Active Learning Environment (PALE). The PALE system combined short answer, natural language questions with feedback in the form of pre-recorded video.
explanations (Nakamura et al., 2016). Although the system did not provide response specific feedback during training, the logs of student responses \( n \approx 150 \) to multiple different physics questions were recorded for subsequent analysis. The corresponding questions all dealt with Newton’s Laws, but varied significantly based on the physical context and the instructional activities asked of the student: a ball and track, a train crash, or a bowling ball and ice skater, etc.

Post-training classification and assessment of student responses was done through a machine learning software focused on the extraction of features from text known as LightSIDE (Nakamura et al., 2016). Human-computer coding agreement was calculated for 9 different questions, with matching agreements ranging from 48% to 89% depending on the question. Noting the large difference in success, the authors argued, “This...illustrates what we believe to be the key to writing questions and activities amenable to this type of automated analysis: questions should have a limited and small number of reasonable responses, and those responses should connect as directly as possible to distinct ways of thinking about the physical system.”

Similar experiments analyzing the post-classification of student responses (as in the work of Nakamura et al. (2016)) have been conducted in other science-related domains. One study explored short answer questions related to evolution (Nehm et al., 2012) and another study, acid-base chemistry (Haudek et al., 2012). These studies also reported promising results in terms of high inter-rater reliability between computer and human coding, but they did not assess the instructional effectiveness of the short answer questions themselves.

In contrast, studies by Jordan et al. (2013) have explored student engagement and interaction with an online short answer assessment system with real-time coding of student responses. These studies involved a variety of general physical science questions (Jordan, 2012; Jordan et al., 2013). In addition to comparing computer and human coding of student responses (with high agreement, depending once again on the nature of the question), student interactions with the short answer format were analyzed in terms of the length and structure of their responses. Although not universal, many students tended to answer in short, incomplete phrases (often with spelling mistakes). Corresponding one-on-one
interviews indicated that students would often acknowledge the role of computer as assessor and that students phrased their answers with that consideration in mind. In addition, student behavior tended to depend on the setting of the assessment question (students were more likely to display gaming behaviors such as leaving their response blank during formative, rather than summative settings) and the applicability of feedback (Jordan et al., 2013).

(b) Intelligent tutoring systems

Overall, intelligent tutoring systems (ITS) also have a successful history. A comprehensive meta-analysis by VanLehn (2011) which studied the use of intelligent tutoring systems across STEM related fields found that intelligent tutoring systems have on aggregate demonstrated significant learning gains versus no-tutoring control conditions, both in laboratory settings and when implemented in classrooms, with an aggregate effect size of $d=0.76$.\footnote{Interestingly, VanLehn’s meta-analysis also suggests that the effect size for human tutoring is actually much lower ($d=0.79$) than the 2-sigma outlier reported by Bloom (1984), and consequently comparable with the gains represented by intelligent tutoring system (VanLehn, 2011).} This finding is in general agreement with more recent meta-analyses that explored the effectiveness of intelligent tutoring systems on college students across content domains (Steenbergen-Hu and Cooper, 2014) and additional grade levels and domains (Steenbergen-Hu and Cooper, 2013; Ma et al., 2014). In each analysis, intelligent tutoring systems were found to provide significant, moderate positive effects on aggregate, varying based on whether the control conditions were taken as computer-assisted instruction, homework-assignments, classroom instruction, or no-treatment controls.

Many of the strengths of intelligent tutoring systems derive from the same advantages of computer-based training methods: the ability to offer individual self-paced practice, provide immediate feedback, guide self-assessment, identify areas of difficulty both for the student and the teacher, and scale to large class sizes (Ma et al., 2014; VanLehn, 2011). Part of the particular success of intelligent tutors comes from refining these advantages to provide further learning gains, such as finer step-based or sub-step feedback (VanLehn, 2011). In other cases, intelligent tutors are able to add completely new advantages to the
mix such as adaptive and individualized task-selection. Either via intricate domain models, as in the case of the Newtonian mechanics learning progression in DeepTutor (Rus et al., 2014a,b), or the mastery models used in ALEKS (Hardy, 2004), intelligent tutors select tasks based on student performance and models of student understanding. The result is a highly individualized sequence of tasks and feedback interactions.

However, it should be noted there is an important caveat to the claim that intelligent tutors are successful due to their use of finer grain size feedback. Through a careful meta-analysis, VanLehn demonstrated that the benefit from additional granularity seems to plateau after reaching problem step-based grain sizes (VanLehn, 2008, 2011) Even though the grain size in interaction with a human tutor is theoretically unbounded and some computer tutors are already capable of sub-step interactions, the typical effect size from both methods are comparable with those of step-based intelligent tutoring systems. The research implication of this “interaction plateau” is that one cannot simply presume to go to ever finer scales of interaction and expect ever-increasing learning gains.

As such, developers of intelligent tutors have sought to take advantage of affective manipulations to further support student learning. Student attitudes, beliefs, and emotional states can have significant impacts on the effectiveness of learning interventions (D’Mello and Calvo, 2011). In particular, student motivation and attitudes have important consequences for gaming behaviors. In the context of computer-based instruction, ”gaming the system” encompasses counter-productive behaviors that allow the student to take advantage of the computer system to avoid activities intended to promote learning. For example, following the path of least resistance, students will sometimes proceed rapidly through a sequence of hints until they reach the bottom-out stage, where they are ultimately told the answer, or presented with enough of the setup that arriving at the answer is trivial (Baker et al., 2004). This behavior tends to correlate more with student than problem (Muldner et al., 2011), so it is often not as simple as changing the hints to a particularly challenging task. Moreover, gaming behavior is not simply restricted to bottoming-out the hint procedure. Other unintended behaviors include those students who rarely use hints, but frequently guess; and those students who could benefit from additional scaffolding or
program features but never apply them (Muldner and Burleson, 2010).

A study by Baker et al. investigated how students’ self-reported beliefs and attitudes correlated with gaming behavior in two different intelligent tutoring systems: ASSISTments and Cognitive Tutors (Baker et al., 2008). Increased gaming behavior tended to correlate with dislike of the subject matter, a lack of self-motivation, frustration, and a dislike of computers or the learning environment; students performance orientation (performance vs. learning goals) was not significantly correlated with gaming behavior (Baker et al., 2008). In a subsequent work, students affective states were monitored while students worked with one of three different intelligent tutoring systems: AutoTutor, The Incredible Machine (a Rube Goldberg puzzle simulation), and a mathematics ITS, Aplusix II (Baker et al., 2010). Out of multiple affective states identified and coded – boredom, frustration, confusion, delight, engaged – boredom was the only state found to significantly correlate with subsequent gaming behavior.

Boredom was also found to be significantly negatively correlated with learning in a different study with AutoTutor (Craig et al., 2004). In that study, students were monitored while they studied information related to computer hardware; learning gains were measured via pre/post-test performance gains. Interestingly enough, confusion was found to significantly and positively correlate with improved learning gains (Craig et al., 2004). From this result, the authors argued that confusion as an affective state was a consequence of useful cognitive disequilibrium.

The idea that confusion, used appropriately, could be utilized to induce cognitive disequilibrium (and subsequent resolution) was tested in a series of recent experiments in a tutor on scientific reasoning (D’Mello and Graesser, 2013; D’Mello et al., 2014). In one of these experiments, confusion was induced by invoking a contradiction between two natural-language pedagogical agents, and monitored by a combination of self-reporting and forced-choice task performance (D’Mello et al., 2014). Overall, post-test and transfer performance was statistically higher versus control when the contradictions induced confusion, but there was no improvement over control when confusion was not induced by the tutor (D’Mello et al., 2014).
Examples of short-answer implementations in intelligent computer tutors

Intelligent tutors utilizing short answer formats are so prominent that they compose their own sub-category: conversational computer tutors. The flagship example of this category of physics ITS is AutoTutor (Graesser et al., 2005, 2001). Although later variations of AutoTutor would branch out into other content domains, the goal of the original version was to tutor students in conceptual Newtonian physics. AutoTutor operates based on the idea of expectation and misconception tailored dialogue (Graesser et al., 2005). In practice, this works as follows: AutoTutor prompts a student to answer a conceptual question (for example, describing why a rear-end collision causes car passengers to suffer neck injuries) and then through short answer mixed-initiative dialogue turns, helps students to refine their answer until it matches an ideal, expert response; particular attention is given to targeting specific misconceptions that might be expected within a student's answer (Graesser et al., 2005). The effectiveness of AutoTutor as an instructional tool has been evaluated in multiple studies in comparison to textbook reading, no-training control, and human training, often with the result that training with AutoTutor resulted in moderate-to-large learning gains over the non-human controls (D’Mello and Graesser, 2012; Graesser et al., 2001).

In order to evaluate the natural language responses of students, AutoTutor relies on a two-stage natural language processing technique (D’Mello and Graesser, 2013, 2012). First, a student response is classified into a corresponding language category (assertion, various question types, short responses, metacognitive expression, and so on) based on punctuation and a syntactic parser. After categorization, student input is matched to any expected responses and misconceptions, as well as standard question and answer phrases, through a mix of natural language processing techniques. Although various incarnations of AutoTutor have used multiple different techniques, the workhorse has been the statistical technique of Latent Semantic Analysis (Graesser et al., 2005, 2001). Finally, after processing the natural language input, a finite-state transition-network determines what response to give to the student (D’Mello and Graesser, 2013).

In a series of seven experiments, VanLehn and collaborators evaluated the performance
of two other conversational intelligent tutoring systems: Why2-Autotutor and Why2-Atlas versus canned-text remediation and human instruction (VanLehn et al., 2007). Why2-Autotutor and Why2-Atlas were follow-up versions of the AutoTutor and Atlas intelligent tutoring systems, respectively. For these experiments, the canned-text was generated from textbooks or contracted logs of the tutor dialogue. The content focused on basic concepts in Newtonian physics, such as Newton’s Second Law. The seven experiments studied the relative effectiveness of the different instructional methods with varying populations of students (pre-instruction novices and intermediate students who were currently enrolled and had already seen the relevant material or had completed an introductory physics course) and different levels of physics content (the original post-instruction material, and a version with extra support for novice learners). The result was that the two conversational ITS provided significantly higher learning gains compared to the canned-text control when novice learners attempted to learn the original content, but no significant gains when intermediate students attempted the original, nor when novices attempted the novice version (VanLehn et al., 2007). Based on these findings, the authors argued that dialog-based intelligent tutoring systems lead to greater respective performance gains when there is a difference between prior student knowledge and complexity of the content.

We briefly note three other examples of full-conversation natural language tutors with varying levels of success. ITSPOKE is a spoken natural language tutor focused on qualitative physics and built upon the Why2-Atlas system (Litman and Silliman, 2004). An experimental comparison between computer training with spoken interaction and typed text resulted in no significant differences in learning gains (Litman and Rosé, 2006). Beetle II is a combined natural language and simulation ITS that trains students on basic electricity and circuits (Dzikovska et al., 2014). In addition to producing large and significant learning gains versus a no-training control, the Beetle II system successfully implemented adaptive feedback (that is, feedback that is created dynamically in response to a students input, tutorial decision rules, and simulation state, rather than being pre-coded and retrieved). Although a highly innovative example of using an ITS to provide custom-tailored feedback, there was no statistical difference between the adaptive feedback and a non-adaptive feed-
back condition in which the students were always shown the correct answer (Dzikovska et al., 2014). Finally, Rimac is a conversational physics ITS that focuses specifically on pedagogical tutoring tactics and their implementation (Jordan et al., 2013).

The use of reflection-on-action prompts and natural language discourse after quantitative problem solving has been the subject of a pair of investigations by Katz and collaborators (Katz et al., 2003, 2007). In the original 2003 paper, two studies were conducted exploring the effectiveness of reflective dialogues provided post-solution while students worked basic quantitative physics problems in Andes. The first of the two examined the transcripts from human tutor-student reflections as the students worked through problems in the ITS. In addition to identifying general patterns used by the human tutors in guiding the reflection, they found that the amount of abstraction in the reflective dialogues (coded by instances of conceptual or strategic generalization) correlated with learning gains, as measured by a quantitative-focused pre/post-test (Katz et al., 2003). The second experiment sought to compare post-solution reflection via human tutors, canned feedback, and no reflection. The result was that students in the two reflection conditions learned significantly more than the no reflection control, but no statistical difference was found between the human tutor and canned feedback reflection conditions (Katz et al., 2003). The 2007 paper then sought to study the effect of reflection after quantitative problem solving in a classroom setting. The conditions were no reflection, two different types of interactive dialogue reflection, and canned text reflection. Although there were issues of student participation (very few students completed all reflection questions), the authors found (by pairing students on pre-test score and major) that students who saw some sort of reflection learned more than those in the non-reflective control (Katz et al., 2007). Consequently, reflection-on-action prompts seem to be useful overall, but it is still unclear how the degree of interactivity, self-explanation and the type of reflection combine to influence learning.
2.2 The Self-Explanation Effect

What is the self-explanation effect? Put simply, the idea is that prompting a student to justify and elaborate as they study a new concept or work through problems aids the acquisition of knowledge, problem-solving skills and subsequent transfer to novel tasks. This effect was originally identified in a small study by Chi et al., which asked college students to voluntarily self-explain as they studied sections of an introductory physics text and worked examples (Chi et al., 1989). They found that those students that generated more self-explanations performed significantly better on follow-up tasks than their peers. This result was then confirmed and expanded upon in multiple studies in physics and other domain areas (Aleven and Koedinger, 2002; Chi et al., 1994; Chi and VanLehn, 1991; VanLehn et al., 1992).

In their original work, Chi and colleagues postulated that the self-explanation effect was primarily the result of inference generation by the learner to fill in gaps in the material to be studied (Chi et al., 1989). They argued that by explaining, the learner infers information that is missing or implied within the provided text or example. Although that hypothesis helped to explain how students interacted with several worked examples, there were inconsistencies with follow-up studies that suggested that inference generation was not the only mechanism (Chi, 2000). In particular, a careful comparison of the content of students’ explanations indicated that they did not always align with omitted information from the text (Chi, 2000). As a result, Chi and collaborators proposed that self-explanation was a dual process, involving both the generation of inferences and efforts to repair the learner’s own mental model. Chi argued that when students read a passage of text or a line in a solution, they compare their interpretation of the meaning of that text to their prior understanding. In the event of a misalignment, the student attempts to explain and repair their flawed representation through a process of mental model revision.

A study by Nokes and Vanlehn (2008) attempted to explicitly compare these two mechanisms in a study of physics problem solving. Students were asked to work through two classes of self-explanation prompts focused on the two proposed mechanisms (gap-filling and
mental model revision) with the result that the prompts focused on inference generation led to greater learning than their counterparts. The authors argued that their results should be viewed in light of the instructional-fit-hypothesis: gap-filling prompts performed better because the cognitive processing of that task better aligned with the goal of the learning task (problem solving). The corollary of such a claim is that self-explanation prompts focused on mental-model revision might perform better in contexts where students are asked to learn a new concept or explicitly address a misconception.

Other studies have corroborated the effectiveness of self-explanations. In one study by Renkl et al. (1998), an experimental group was trained on the value and method of self-explanations and compared to a control group that was trained on a more generic think-aloud protocol. The self-explanation training procedure increased the frequency of student self-explanations and subsequently improved student performance on both near and far-transfer problem solving tasks. Another study looked at how self-explanation interacts with other pedagogical techniques and factors. They found that crossed self-explanation and the fading of examples during computer-based instruction on probability did not statistically interfere with one another, with both manipulations providing significant positive effects (Atkinson et al., 2003).

Finally, there have been a number of studies in the context of computer-based tutors that have explicitly targeted the use of self-explanation. Conati and VanLehn evaluated a computer-based tutor that trained students on the use of self-explanation in support of learning worked-out examples (Conati and VanLehn, 1999, 2000) In addition to scaffolding student practice with the generation of self-explanations, the tutor gave explicit hints to encourage student explanation of key problem components. Self-explanations were selected from a drop-down menu and included both solution planning and physical rule-based explanations. Although results were mixed, there was some evidence that the addition of these selection-based explanations was beneficial for weaker students, and less so for more experienced students. The authors argued that explicit and detailed self-explanation scaffolding (as in their tutor) may be most effective for students early on in the learning state and relatively unfamiliar with the subject matter; more advanced students may find the
elaborate scaffolding counter-productive.

Work by Aleven and Koedinger took a similar approach, using a pre-built glossary to support student self-explanations in a Cognitive Tutor focused on geometry (Aleven and Koedinger, 2002). Students could either type in the name of the rule explaining a step in a geometry-focused problem or select it from the glossary. Over the course of two classroom experiments, they found that students who explained their steps were more successful on transfer problems. Moreover, the explanations were still effective when they were selected from the glossary. Another study looked at self-explanations via typed student responses (Hausmann and Chi, 2002). Students in the treatment condition were asked to explain parts of a text on the human circulatory system, to be later assessed by a human (there was no in-progress feedback on student explanations). The control condition was a read-only task. Interestingly, the experimenters found that students tended to only paraphrase the text, a result the authors argued was either a consequence of the demands of the input format or because students knew they would ultimately be assessed on their written responses. A subsequent experiment with explicit and repeated prompting did result in small increases in the amount of student explanations.

2.3 Memory & Testing Effects of Short Answer Formats

Other researchers have approached investigations into short answer formats and their effectiveness from an entirely different perspective, namely the role of format in determining the cognitive effects of testing. In particular, there has been a large amount of work studying the effect question format during testing and practice has on student retention of the relevant material.

At its core, the dissimilarity between short-answer questions and multiple choice are underlying differences in the cognitive processing required by students to answer the question. Short answer questions explicitly require the student to retrieve prior information; multiple choice questions test recognition of the correct answer amidst compelling distractors. Research suggests that - at least under some circumstances - retrieval practice us-
ing short-answer questions can benefit learning more than multiple-choice questions (Kang et al., 2007; Butler and Roediger, 2007; McDaniel et al., 2007; Clariana, 2003; Smith and Karpicke, 2014; Gay, 1980).

Feedback seems to play a central role in the relative effectiveness of the two question formats. Contrary to the above results, there have been multiple studies that have found no significant differences between short answer and multiple choice testing formats on student retention when feedback was not provided to the student (Haynie, 1989; Duchastel and Nungester, 1982; Frase, 1968). As such, some have argued that feedback is an important modifier of the success of short answer formats because students typically demonstrate lower performance - via measures of initial retrieval success - on short answer questions than on multiple choice questions (Kang et al., 2007). In essence, feedback may help account for differences in inherent difficulty necessary for retrieval practice to offer further retention benefits.

Still, results have not always been completely consistent. The study by Gay (1980) compared repeated practice with short answer and multiple choice questions in a college course and found that the short-answer questions were only significantly more effective when the final assessment was in short answer format; post-test assessments based on multiple choice formats found no differences between the initial practice formats. However, the study by Kang et al. (2007) found the exact opposite. Short answer questions significantly improved student scores on a subsequent multiple choice post-test, but not on subsequent short answer assessments (though it is important to note that the trend was in the same direction).

As a result of this network of complicated findings, some authors have argued that it is not simply retrieval difficulty that determines the relative effectiveness of the two formats, but a combination of retrieval difficulty and retrieval success (Smith and Karpicke, 2014). In short, the more difficult the retrieval practice, the greater the gains when that retrieval is successful. The implication is that the reverse is also true: when initial retrieval is too difficult, success rates will be low and subsequent retention measures will decrease. To test this, the authors studied several comparisons between short answer and multiple
choice formats during initial practice with, including a hybrid system where students first
answered a short answer version of a question, followed immediately by a multiple choice
version. Interestingly, they found that for 3 out of 4 experiments there was no significant
difference between the short answer and multiple choice formats. The only experiment
to show significant differences from the short answer format over multiple choice involved
improved retrieval success on the initial short-answer training. In addition, there was no
overall effect from the inclusion of the hybrid format over individual practice, even though
those in the hybrid conditions answered each question twice. The authors summed up their
findings, “...we found that initial retrieval practice format only mattered for learning under
very specific circumstances. Specifically, learning differences between formats only emerged
when initial retrieval practice success was as similar as possible, and when direct feedback
was provided.”

2.4 Summary

This review - though by no means exhaustive - illustrates the complex network of factors that
may potentially influence the effectiveness of short answer implementations in a computer-
based instructional setting. Feedback, grain size, student affect, and the self-explanation
effect are some of the factors influencing the potential effectiveness of student’s short answer
responses. In addition, there may be other psychological factors such as the role of recall in
memory that support short answer responses and their retention. Finally, we note previous
work has also suggested important practical recommendations, such as ensuring that the
response space for a reasonable short answer is sufficiently constrained as to be tenable for
computer-based analysis and encouraging student explanations where applicable.

This review also indicates there is further potential value to be added by studies ex-
plicitly comparing content-equivalent short answer and multiple choice formats within a
computer-based training setting, particularly in a conceptual domain - such as force and
motion in physics - where the recall and use of a concept or physical situation is both suffi-
ciently constrained, but also challenging for students. Although many studies focused on the
psychological merits of short answer responses have sought to explicitly compare the two formats, there have been fewer investigations studying their relative effectiveness specifically in the context of computer-based training. Instead, cutting-edge studies have frequently compared their natural language-based interventions to the golden standard of individual tutoring (an admirable benchmark) or to non-interactive controls like a read-only condition. Content-matched comparisons between short answer questions and multiple choice formats can help shed light on the role of feedback and student interaction, the relative efficiency, as well as effectiveness of the two formats, while both controlling for subject matter and monitoring effects due to limitations of the short answer format (computer accuracy).

The next chapter will frame the research questions that will be explored in the remaining chapters of this volume. In addition, it will describe the design and implementation of our relatively simple computer-based training system and the variations between multiple choice and short answer question formats used in the subsequent studies.
3.1 Research Questions

Six experiments were conducted with students at three different levels of introductory physics instruction: first-semester algebra-based, first-semester calculus-based, and second-semester calculus-based physics. These six experiments are illustrated schematically in Fig. 3.1. Over the course of the studies, we addressed four main research questions.

Research Question #1: Controlling for feedback and content, are short answer computer-based training formats more effective than multiple choice counterparts (in terms of overall student performance and efficiency)?

As alluded in the proceeding chapters, there are several motivations for such an investigation. First and foremost, although there is recent and extensive research effort to improve the performance of automated short assessment, there is less research focused on the instructional effectiveness of that format given its constraints. Moreover, although conversational computer tutors have established themselves as effective instructional devices (Rus et al., 2013, 2016), experimental designs have often sought to compare such interventions either to non-interactive controls (such as reading of a textbook or log of tutor output) or to human tutoring interactions. Comparison of short answer to multiple choice – controlling for the feedback provided to the students – helps address the question of whether the benefit from short answers is due to increased student interaction, the content, or the format itself.
Figure 3.1: Schematic illustration of the six short answer experiments.
Finally, such a study can assess attitudinal differences in how students view and interact with the question formats. This information could be useful to guide development of new question types for computer-based instruction systems that place a high-priority on a variety of representations and question types, such as online training via the essential skills framework (Mikula and Heckler, 2017b).

In addition to supporting the existing literature through explicit comparison of the two question formats in computer-based instruction, we seek to add unique value by studying a powerful system of physics concepts characterized in previous research: one-dimensional force and motion. This system consists of relationships between force, velocity, and acceleration in one-dimension. Previous work has suggested that this set of concepts could be compelling for comparisons of question and answer formats for several reasons.

First, there are robust and persistent student difficulties among the related concepts that may benefit from computer-based training, even though the collective relationships between force, velocity and acceleration are relatively simple (Rosenblatt and Heckler, 2011; Clement, 1982). In particular, students have a tendency to wrongfully believe that a moving object implies a net force in the direction of motion (Clement, 1982). One potential contributing reason for the persistent nature of this student misconception is that counter-cases such as oppositely aligned force and velocity vectors may simply be less available than physics contexts where an object’s net force, velocity, and acceleration are aligned (either due to everyday experiences or typical physics examples). As such, repeated retrieval, consideration, and practice with these counter-cases may help students to confront that pervasive misconception. Moreover, that practice may be more or less effective depending on the question format.

Second, previous work (Brookes and Etkina, 2009) and our own pilot studies with students have suggested that students may struggle with the nature of physics terminology in this setting. For example, in every-day parlance a “moving object” can refer to multiple different physical situations: an object accelerating and speeding up, an object moving at a constant-speed, or an object accelerating and slowing down. It is not clear that students conceptually distinguish between unique physical situations when they refer to “movement”
or “acceleration”. Short answer formats may provide benefits in helping students to distinguish different physical situations through practice, both with physics terminology (i.e., “accelerate”) and everyday parlance (i.e., “move”).

Third, previous work has developed and validated assessment tools for measuring student understanding of this set of concepts (Rosenblatt and Heckler, 2011; Thornton and Sokoloff, 1998). The FVA test is a 17 question multiple choice test designed to specifically assess student understanding of the relationships between net force, velocity, and acceleration. The test has previously been used to evaluate evidence of conceptual hierarchies amongst the physics concepts, and has indicated that students continue to struggle with these concepts even in their second year of undergraduate physics coursework (Rosenblatt and Heckler, 2011; White and Heckler, 2013).

Given the above, one-dimensional force and motion were the focus of the majority of the studies presented here (Studies #1-5). However, we conducted an additional study, Study #6, with the concepts of electric field and electric potential to corroborate results with a different set of physical concepts (for which students typically have less prior experience).

**Research Question #2:** Controlling for feedback and content, does short answer computer-based training provide retention gains compared to multiple choice?

There is a long and rich history of research into the testing effects of the two question formats on student memory and retention. However, many of the previous studies have not included feedback to the student during training. A survey of results in the field suggests that feedback is instrumental, particularly for short answer questions; without feedback, multiple-choice formats tend to result in increased student performance due to the relatively increased difficulty of the short answer counterparts. Our study complements existing research that has evaluated the two formats along with instructional feedback. Whereas many of these studies have focused on student retention of information acquired from studying a static text, our study asks the question of whether there are similar retention gains for training students on conceptual questions.
In addition, we believe that the conceptual domain of one-dimensional force and motion presents a good testing ground for the claim that differences between the two formats depend on both retrieval difficulty (the more difficult, the more beneficial an instance of successful retrieval) and retrieval success. As evidenced by clear and profound student struggles on the FVA-test, recognizing and retrieving counter-cases of anti-aligned forces and velocities is difficult for students (or anti-aligned velocities and accelerations). As such, we may be able to offer unique insights into this postulated combination of retrieval difficulty and retrieval success. Retention measurements were conducted in several of the studies presented here, namely Experiments #3 and #4.

**Research Question #3:** Does an increase in the interactivity of feedback during training affect the relative performance of short answer formats?

The success of conversational tutors suggests an important question: Do short-answer questions need to be presented in conversational contexts to be relatively successful? Or to put it a different way, how does the relative performance of short answer question formats versus multiple choice formats interact with an increase in the conversational nature of the training? Conversational tutors such as AutoTutor can take up to a hundred question turns (Graesser et al., 2005). Experiment #3 and Experiment #4 seek to compare the two formats under a more restricted and controlled comparison of stem-equivalent multiple choice and short answer questions (3-5 question turns).

**Research Question #4:** Does an increase in the variety of questions during training affect the relative performance of short answer formats?

Given the results of the previous studies, Study #5 sought to test the effect of an increased variety of question types and structures within the constraints of the two question formats: short answer and multiple choice. Along with an increase in treatment dose, the goal of this study was to examine whether further increases in question variety may lead to
relative improvements of the short answer format versus multiple choice. Given previously hypothesized interactions between retrieval difficulty and success, an increase in question variety may increase the overall difference between short answer and multiple choice formats.

In order to pursue these research goals, we designed and implemented a simple, clean computer-based training interface that could be used to administer both short answer and multiple choice questions, and automatically assess student responses and provide appropriate feedback. The design and motivation of that software is described in the next section.

### 3.2 Design of Training Software

The training software and underlying natural language implementation were developed in Python. Although the interface for the training software underwent a number of minor revisions throughout the course of the six experiments, the basic structure remained the same (see Fig. 3.2). The interface consisted of a scrolling-dialog window (left) where the question, student response, and feedback were shown; a window for question related graphics (upper right); a session progress-bar that provided the student with a rough measure of progress through the task (right); a student input line (bottom left); and an optional upvote/downvote prompt to measure student opinion on the usefulness of provided feedback (bottom right). The optional upvote/downvote prompt was included to give students in the short answer conditions a way to report incorrect assessments of their responses by the computer (and thus an increased level of student interaction in the hopes of minimizing potential frustration), but also to encourage students to monitor and reflect on the accuracy of the provided feedback. The upvote/downvote prompt appeared at the end of each question or question-chain.

In order to minimize any unintended differences from the program interface, the tutor used the same interface for both short answer and multiple choice formats. For short answer questions, students would type their response; for multiple choice questions, students would only type and submit the letter of their choice. Instructions related to the training task and how to use the interface were provided to students via a welcome screen. In addition,
for the first question students answered, a brief description appeared below the input line reminding students to type out their answer in short, concise sentences or provide the letter of their choice depending on the format of the question.

The structure of a question-turn in the tutorial depended on the experiment. In earlier studies, the majority of questions were single-turn. For those experiments, a question was
posed, the student responded either in multiple choice or natural language, and the tutorial provided corresponding feedback. In the first 2 studies, feedback depended on whether the student was correct/incorrect. If the student was marked correct, they were simply told "You are correct!" and invited to continue to the next question. If the student was marked incorrect, they were given an additional statement identifying the correct answer as part of a brief two to three sentence explanation. The feedback was identical between the two conditions. An example question, student response and feedback turn in each format are shown in Fig. 3.3.

![Table: Question Formats](image)

<table>
<thead>
<tr>
<th>Short Answer</th>
<th>Multiple Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>One of Ohio State's football players is on the field during a game. At a particular instant, his acceleration is directed towards the offensive line. What do you know about the direction of his velocity at that instant?</td>
<td>One of Ohio State's football players is on the field during a game. At a particular instant, his acceleration is directed towards the offensive line. What do you know about the direction of his velocity at that instant?</td>
</tr>
<tr>
<td>User: the velocity is towards the offensive line</td>
<td>(a) His velocity is directed towards the offensive line.</td>
</tr>
<tr>
<td>Sorry, but your answer is not correct.</td>
<td>(b) His velocity is directed away from the offensive line.</td>
</tr>
<tr>
<td>The correct answer is that the velocity could be directed towards the offensive line, away from the offensive line, or it could be zero. The acceleration does not require that the velocity have any particular direction at that instant. For instance, the player could be at rest and accelerating towards the line, or accelerating towards the line but with an instantaneous velocity directed away (that is moving away and slowing down).</td>
<td>(c) His velocity is zero.</td>
</tr>
</tbody>
</table>

Figure 3.3: Example question, student response, and provided feedback from both question formats.

Starting with Experiment #3, feedback was reformulated slightly for single turn questions, although it was kept the same across multiple choice and short answer formats. From that point on, in addition to stating whether the response was correct/incorrect, the correct answer was always provided along with a brief explanation - even if the student had been
marked correct. This was done in order to try and minimize potentially negative effects from a misidentified answer. In addition, there were two minor but additional, potential differences between the two formats. These differences were based on previous findings from Studies #1 and #2, and were implemented to ensure that the short answer format was not unfairly handicapped versus the multiple choice implementation.

First, short answer conditions included potential prompts to elicit further student explanation. These prompts were only used when students answered too briefly on questions that specifically asked for a student to justify their answer or provide their reasoning. For example, a question may have asked the student to argue which of three hypothetical students they agreed with concerning a concept related to force and motion. If a student in the free response condition simply responded ”Student B”, the program would respond, ”Ok. Please explain your reasoning.” On the other hand, the multiple choice versions of those questions included prototypical examples of that reasoning as part of the available choices.

Second, the short answer question format included one-turn clarifying prompts that asked a student to restate their answer when they used terminology or referred to directions that were not described in the problem statement. For example, a short answer question might ask a student to describe what they know about the direction of an object’s motion relative to another object at a particular instant in time (i.e., a football player moving relative to the offensive line). If a student referred to a “positive” direction when the problem had not formally defined a coordinate axis, the student was prompted to clarify their answer given the physical context of the question. This design choice was done due to avoid a common failure identified in the pilot implementation, as examined and described in Study #1.

Later, experiments also explicitly manipulated both the structure of the training questions and the presence of additional follow-up questions. For follow-up questions, student responses and subsequent feedback from the computer were added to the scrollable log so that students could go back and review earlier parts of the conversation. Further details about the form of specific feedback and follow-up questions are included in the design sections of the individual experiments.
In the event that the computer could not interpret a student response, the question turn was restated in the form of its multiple choice counterpart. This design choice is an important one, as it represents a significant departure from alternate implementations where students are repeatedly asked to rephrase their answer, either until the computer can identify their response or ultimately fail and simply provide the correct answer. Given the previous research on gaming and student affect, the choice to provide a multiple choice clarification was made to minimize student frustration and avoid gaming effects due to students simply "bottoming-out" the requested number of restatements. The hope was that the current implementation would still encourage students to try and provide a response identifiable by the computer for the sake of efficiency. If the computer was able to interpret their response to the short answer question, the implicit reward was continued progress through the training and overall less time spent on the training task.

The practical consequence is that the implementation of the short answer format here is most similar to the hybrid short answer format advocated and studied in the work by Smith and Karpicke (2014). As such, it is important to note a key result of that study, namely that a fully hybrid format - where every question was first presented in short answer format, and then immediately followed by a repeated multiple choice restatement regardless of how the student performed on the short answer format - did not differ significantly from treatments presenting only short answers. In our studies, students were always asked to consider the short answer format first, and were only prompted to answer the multiple choice counterpart when the computer was otherwise unable to interpret their initial response.

3.3 Natural Language Implementation

The purpose of the work presented here was not to advance the state-of-the-art in automated short answer assessment. Rather the evaluation of natural language in student responses was done to a) allow feedback to the student that depended both on correctness and type of error made by the student, and thus match as close as possible that of counterpart multiple

2Interestingly, a hybrid block format where students first answered the full set of short answer questions and then all multiple choice versions in the same order was also not significantly different from short answers alone (Smith and Karpicke, 2014).
choice questions and b) indicate to students that their responses were being assessed, much in the spirit of multiple choice.

As such, the natural language implementation used in the following studies is a combination of two comparatively simple, but well-established techniques: regular expressions and latent semantic analysis (LSA). Regular expressions are simply a pre-defined set of text-based matching rules that can be used to identify key words or structured phrases in a student’s response. In contrast, LSA is a high-dimensional, corpus-based method of analyzing similarity between any two pieces of text words, phrases, sentences, or entire documents (Landauer et al., 1998). Latent semantic analysis makes the assumption that a candidate text can be represented as an unordered collection of its component words, appropriately coined as a “bag of words”. Using a training corpus, LSA generates a multi-dimensional space to represent the conceptual meaning behind words in that corpus. Similarity between a candidate response and a standard text is assessed through comparison of the vector representations of the two texts within this model space. As such, similarity is measured by the projection of the candidate vector onto the standard text with a magnitude between 0 and 1 (consequently this measure is often referred to as cosine similarity). In addition to being utilized in the AutoTutor series, LSA has been implemented in multiple other intelligent tutoring systems (Graesser et al., 2004).

The advantage of using these older, but in some ways foundational, techniques is the availability of multiple robust software libraries and toolkits. In particular, we used several Natural Language Toolkit libraries (Bird et al., 2009) and the LSA implementation of GENSIM, an open-use python library specifically designed for statistical semantics and semantic similarity (Řehůřek and Sojka, 2010). The LSA model was trained on a collection of relevant sections of introductory physics textbooks as well as a database of related questions and student answers.

The general procedure for analyzing a student answer was as follows. After spell checking, the student response was compared to any question-specific rules, as represented by a pre-defined set of regular expressions. These rules included a set of specific match and avoid criteria. In order to meet all requirements a student response had to include all of
the elements specified by the relevant matching expressions, but also include none of the pre-defined elements that had been marked for avoidance. Regular expressions were particularly useful for questions where student short answers focused on important key words, such as identifying a direction or agreeable viewpoint (i.e. a student may be asked to argue in support of one of three different hypothetical students). For more open-ended questions (and when regular expression matching failed), the computer tried to match the student’s response to a catalog composed of sample answers and previous student responses using LSA. If a student response did not reach a sufficient level of similarity to an example entry, the computer opted to restate that specific question turn in multiple choice format.

3.4 Training Questions & Format

Questions in short answer and multiple choice format were designed to be stem-equivalent. In other words, short answer formats consisted of the same basic prompt in the multiple choice format, without explicit reference to any of the example choices. Occasionally, small changes in wording were necessary to make the core prompt clear, but for the most part, question formats were identical. A list of training questions used in each experiment is included in Appendix A.
Chapter 4

Study of Short Answer Format & Computer-Based Training in Physics

4.1 Experiment #1: Pilot of Short Answer vs. Multiple Choice (Force & Motion)

4.1.1 Design and Participants

The goal of this first experiment was to pilot a relatively simple comparison of multiple choice and short answer formats in the context of one-dimensional force and motion. Previous work has suggested that students struggle with the relationships between net force, velocity and acceleration even in the restricted case of one dimension (Rosenblatt and Heckler, 2011; White and Heckler, 2013). In particular, students tend to make the conceptual error of assuming that an object’s velocity and its net force must necessarily be in the same direction at all times (Clement, 1982). As such, the retrieval and consideration of alternate states of motion (such as anti-aligned force and velocity vectors) represents a unique, research-based test case for differences between computer-based question formats in training.

A total of 171 students were randomly assigned to one of four experimental conditions as shown in Fig. 3.1. The experiment conditions included computer-based training via multiple choice questions with feedback (N=45), short answer questions with feedback (N=40), short answer questions without feedback (N=48) and a no-training control (N=38).
The training consisted of 12 questions targeting the relationships between the directions of net force, velocity, and acceleration at a particular instant in time. These questions were designed to follow the structure of the FVA test developed by Rosenblatt and Heckler (2011). As such, they posed the general question: “Given the direction of $x$ at a particular instant in time, what do you know about the direction of $y$ at that moment?”, where $x$ and $y$ can represent either the net force, velocity, or acceleration of an object at a particular instant. As an example, the first question in the training asked students, “One of Ohio State’s football players is on the field during a game. At a particular instant, his acceleration is directed towards the offensive line. What do you know about the direction of his velocity at that instant?” The full set of training questions is included in Appendix A.2.

As shown in Fig. 3.1, participants completed one of the training conditions, followed by 10-15 minutes of unrelated physics tasks, and finally a multiple choice force and motion assessment. The force and motion post-test was composed of 16 (out of the original 17) questions previously studied and validated in the FVA test by Rosenblatt and Heckler (2011). One question from the original FVA test (Item 9) was not included in this particular experiment because it was originally formulated to allow students to select multiple correct answers. The full FVA test is included in Appendix A.1 (along with the modified version of Item 9 that was used for subsequent experiments).

Experiment participants were students enrolled in the second semester of a calculus-based introductory physics sequence at The Ohio State University. Students completed the tasks as part of a one-hour flexible homework assignment in their introductory physics course. Course credit was given for participation. Students completed the training and assessment in individual carrels in a quiet testing room.

4.1.2 Results

(a) Student performance on the force and motion assessment

Students’ final course grades in their physics class (second semester of the introductory calculus-based physics sequence) were collected in order to provide a general measure of
student ability. A one-way ANOVA showed no statistical differences between students' overall course grades across the four conditions, \( F(3, 167) = 0.087, p = 0.97 \).

![Mean Post-Test Score by Condition](image)

Figure 4.1: Student scores on the 16 question, multiple choice FVA post-test by condition.

The mean scores on the force and motion assessment are shown in Fig. 4.1 for each of the four training conditions. A one-way ANOVA indicated a significant difference between the four conditions \( F(3, 167) = 3.2, p = 0.03 \). In order to assess the effectiveness of each treatment relative to control, we conducted a Dunnett’s post hoc test. There was a statistically significant difference in scores between the short answer format and control \( (p = 0.037, d = 0.60) \), a marginally significant difference between multiple choice format and control \( (p = 0.078, d = 0.46) \), but no difference between short answer without feedback and control \( (p = 0.94, d = 0.09) \).

To check for an aptitude-treatment interaction, we divided students using a median split on their final course grades, as shown in Fig. 4.2. Using this split, a 2 (upper vs. lower course grade) x 4 (training condition) ANOVA showed main effects from course grade \( (p < 0.001) \), and condition \( (p = 0.02) \), but no significant interaction effect. In short,
students who performed better overall in the second semester physics course consistently performed better on the assessment, regardless of training condition. This suggests that the short answer format piloted here neither preferentially helps those who perform better in the course (who may already have a better grasp of the expected answer and language), nor those who perform worse (who may be the most to benefit) compared to stand-alone instruction or multiple choice training.

In order to more directly compare student performance with feedback across multiple choice and short answer formats, we conducted a corresponding ANCOVA on post-test scores between only the multiple choice and short answer formats with feedback, using course grade as a covariate. Controlling for course grade, there was no significant difference between the two conditions \(F(1, 99) = 3.714, p = 0.057\).
(b) Training Efficiency

The median times spent on the training were 805s for the short answer format, 683s for the short answer format without feedback, and 488s for multiple choice format. A median test showed that the training time was significantly different between conditions ($\chi^2(2) = 18, p < 0.001$). Post hoc pairwise comparisons (adjusted with a Bonferroni correction for multiple tests) showed that multiple choice was significantly different from both short answer ($z = 19.8, p < 0.001$) and short answer without feedback formats ($z = 6.746, p < 0.001$). There was no significant difference in the time spent on training between the two short answer formats ($z = 1.650, p = 0.597$).

In order to better compare potential trade-offs between learning gains and training time, we define the efficiency of training for a particular student as the ratio of assessment score over total time spent during training (in minutes). The mean efficiency ratings (% score/minutes training) for each of the training methods were 8.44 for multiple choice format, 5.16 for short answer, and 5.52 for short answer without feedback. The distribution of individual efficiency ratings for each condition is shown in Fig. 4.3. A one-way ANOVA showed that the difference between these efficiencies is significant [$F(2, 130) = 11.3, p < 0.001$]. A Tukey post hoc showed that the multiple choice format was significantly more efficient than either short answer condition, regardless of feedback ($ps < 0.001$). The difference in efficiency for short answer and short answer without feedback was not significant ($p = 0.89$).

The main finding to note is that a large part of the success of the multiple choice format – at least as defined by this efficiency metric – is the long tail of students who performed well on the assessment and proceeded quickly through the corresponding training. A similar tail for the short answer without feedback condition helps to explain the comparable efficiency performance with the short answer format that did receive feedback (even though the no feedback condition resulted in overall lower performance). In essence, although the short answer format with feedback resulted in comparable overall performance to multiple choice, these gains resulted from about a 40% time trade-off (here, approximately 5 minutes).
(c) Accuracy of automated short answer assessment

One of the goals of this pilot experiment was to analyze the accuracy and effectiveness of our simple natural language implementation and identify potential improvements. To begin, it is worth noting that out of 40 students in the short answer condition, 3 students demonstrated identifiable gaming behavior. In each case, the student rapidly (< 8 seconds) entered a short noncommittal phrase or submitted a blank response on more than one training question. Discarding the responses of these students, the natural language model produced a match for 62% of student responses.

In addition to limitations of the LSA technique and simple cases of vague wording, typical student answer patterns suggest several additional difficulties in identifying student responses. First, students would frequently invoke ambiguous coordinate systems in their free-responses to describe the direction of a physical quantity of interest. For example, a
response would refer to a “negative” direction, or a particular vector being “positive”, even though no coordinate axis was specified in the problem or explicitly stated by the student.

Second, students would occasionally state partially correct relationships or definitions, which although relevant to the question, were not necessarily what the question was asking. For example, a student would correctly define Newton’s Second Law, but then not apply it to find the direction of the acceleration. Another student, having correctly stated that acceleration was the time rate of change of velocity, incorrectly determined the resultant direction of the object’s velocity. Such cases suggest the potential for targeted clarifying dialog or follow-up questions.

For those statements which the natural language model produced a match, we tracked the number of false-positives and false-negatives – instances where the computer wrongfully declared to the student that a response was correct, or wrongfully declared a response was not correct respectively – by comparing the computer identification to hand-coded grading. We found an overall false-positive rate of 3.5%, largely driven by one question with a false-positive rate of 7.9%, and a false-negative rate of 9.3%. To test for any negative learning effects of false-positives (false-negatives were less of a concern because of the availability of explanatory feedback), we compared performance on the post-test between those students who saw at least one false-positive and those that saw no such misidentifications, as shown in Fig. 4.4. An independent samples T-test indicated that the mean score on the force and motion assessment was not significantly different between the two groups, \[ t(36) = -0.83, d = 0.36, p = 0.41 \], despite the fact that student course grades were also nominally lower – though not significantly – in the group who saw at least one false positive \[ t(36) = 0.495, d = 0.32, p = 0.486 \]. In addition, we found no correlation between the total number of misidentifications – either positive or negative – and student performance on the final assessment \[ r(38) = -0.036, p = 0.830 \].

The natural language model produced a match for 62% of student short answer responses – meaning 38% of the time, the computer was unable to successfully match the initial student answer to an answer prototype. In such cases, the student was asked a clarifying multiple choice version of that particular question. To what extent did the clarification of
the question via multiple choice influence the effectiveness of the short answer treatments? One way to provide a first-order estimate is to look at differences in student score on the final assessment as a function of the number times students were asked to clarify their answer.

We found that there was no correlation between student score on the force and motion assessment and the number of multiple choice clarifications received as part of the short answer training \[ r(38) = -0.083, p = 0.626 \]. Although the lack of statistical significance here is likely due in part to the limited number of students within the short answer treatment condition, it is noteworthy that the size of the correlation coefficient is quite close to zero. As such, at least as a rough first-order estimate, there doesn’t seem to be a large effect from asking relatively more follow-up multiple choice questions.

(d) Student Performance on Training Questions

Did student success on the training questions themselves differ based on format? To assess empirical difficulty of the training questions we compared human-graded scores of student
responses during training via short answers with feedback to student performance on the multiple choice questions with feedback. For the short answer condition, we only considered students’ initial responses (that is, if students were subsequently asked a multiple choice clarification question, we scored student responses based on their initial short answer). For short answer questions where the correct answer was that the given and requested quantities were unrelated at an instant in time (i.e. what do you know about the direction of an object’s instantaneous velocity given its instantaneous acceleration), student short answer responses were marked correct if they either listed the potential cases (as indicated by the multiple choice format), made a claim that there was not enough information, or asserted why nothing could be said of the requested quantity given their relationship. The results are shown in Fig. 4.5.

![Proportion of correct responses to each training question by condition](image)

Figure 4.5: Student performance on the training questions with the short answer and multiple choice conditions (both with feedback). Dotted lines represent the overall average. Error bars are ±1 standard errors.

The average score on the multiple choice and stem-equivalent short answer training questions was 66% and 53%, respectively. We compared total score on the training questions using a general linear model, accounting for a main effect (question format) and a
covariate (course grade). Both course grade $[F(1, 79) = 10.92, p = 0.001]$ and question format $[F(1, 79) = 9.35, p = 0.003]$ were found to significantly predict student performance on the training questions. A closer look at the performance of students on the individual training questions (see Fig. 4.5) suggests that the majority of this difference comes from four questions: Q4, Q6, Q9, Q10. Those questions all involved the (lack of) relationship between the instantaneous velocity of an object and either the net force or acceleration at that particular instant. Moreover, the questions were all framed in a concrete physical context. This decrease in performance is a reflection of two trends. First, there were a handful of insufficient student responses to the short answer questions based on irrelevant statements or unstated assumptions that were simply not choices in the multiple choice format (i.e. “there is no force besides the force of gravity down on the puck.”) However, more predominantly students would either assert the incorrect claim that the quantities must be aligned, or omit one of the other two physical possibilities in their short answer response (the zero case more often than the case of anti-alignment).

To assess the effect of feedback on the verbosity of student responses, we compared the average length of students’ responses in the short answer condition with feedback to short answer training without feedback. The mean length of a student response in the short answer with feedback condition was $7.1 \pm 1.1$ words; on the other hand, the mean length of a response in the short answer without feedback condition was $13.5 \pm 1.4$ words. The difference between the two conditions was significant $[t(83) = 3.04, p = 0.003]$. As such, this difference implies that students in the feedback condition respectively adapted their answers compared to students in the no-feedback condition, likely as a consequence of efforts to improve the computer’s interpretation.

4.1.3 Discussion

In this pilot experiment, we demonstrated that an automatically-assessed short answer question format can be at least as effective as simple multiple choice practice for learning basic force and motion concepts. Moreover, the effectiveness of the training was independent of the general physics ability of the student (as assessed by their final grade in their physics
course). However, given the finding of no significant differences between the short answer and multiple choice formats and the minimal impact of eliciting short answer responses in the absence of feedback ($p = 0.97, d = 0.09$), these findings primarily support the value of immediate and specific feedback, rather than recommend a particular question format. It is clear that feedback – rather than simple, recent exposure to the format and content of the training questions – is doing the majority of the work.

Although overall performance on the subsequent multiple choice assessment was not significantly different, there were several key differences between the respective treatments. On the one hand, the multiple choice format was significantly more time efficient than eliciting a natural language statement from the student. This was in part due to extra time spent by the students constructing their answers, but also due to the time required for students to answer clarification questions due to a failure of the computer to appropriately match an answer. Approximately 40% of the time, the student was asked to answer a clarifying multiple choice restatement of the question. However, the number of clarification requests did not correlate with student performance on the force and motion assessment. As such, the findings here are in agreement with the trends found in the work by Smith and Karpicke (2014). In 3 of their 4 reported experiments, they found no differences between short answer and multiple choice formats. More importantly, they found no significant differences compared to a fully hybrid format (where students always answered a restated multiple choice version after the initial short answer question).

However, there is some evidence that students in the short answer condition did interact with the training differently than students in the multiple choice format. In particular, there are hints that certain types of questions in force and motion may be more difficult for students in the short answer format, namely situations where the given and requested quantity need not necessarily be aligned (velocity and either net force or acceleration).

Put another way, students have a persistent and well-documented misconception that force and velocity must be aligned (Clement, 1982; White and Heckler, 2013). It appears it may be more difficult for students to recall the alternative physical possibilities than it is to recognize them from a repeated multiple choice structure. Further changes to the formats
that help emphasize that relative difference – either through more conversational emphasis discussing the physical alternatives (Experiments #3 and #4) or through an increased variety and deeper level of physics questions (Experiment #5) – may allow for subsequent gains and retention.

Experiment #2 presents an effort to replicate and extend the results reported here with a different (and potentially more relevant) population of students. Whereas this experiment involved participants from an introductory calculus-based course involving mostly topics related to electromagnetism, the next experiment was conducted with students from a first-semester introductory mechanics course.

4.2 Experiment #2: Short Answer vs. Multiple Choice

The primary motivation of this experiment was to replicate the previous findings with a more appropriate group of student participants, namely students recently introduced to the concepts of force and motion in their coursework as part of a first-semester, calculus-based introductory mechanics course. As such, the design of Experiment #2 was the same as Experiment #1 (see Fig. 3.1).

However, based on the findings of Experiment #1, a number of iterative improvements were made to the training conditions. Most importantly, Experiment #2 included several new training questions. Unlike Experiment #1 where all of the questions were in the same, predominant format used on the FVA post-test (i.e., “Given \( x \), what do you know about \( y \)?”), Experiment #2 included three multi-turn questions that asked students to think about the implications of words like “moving” and “accelerating” specifically in a physics context. In particular, these questions targeted vague, everyday descriptions of a “moving object” where it is not clear if the speaker intended an object that is moving at a constant speed (and thus, a physical situation where forces on that object must be balanced) or accelerating (an object under the influence of a non-zero net force and consequently either speeding up or slowing down). We hypothesized that repeated and vague use of descriptions like “moving” across different physics situations may contribute to the persistence of the misconception.
of a necessary force in the direction of motion. After all, if students are unable to cleanly verbalize different alignments of instantaneous net-force and velocity vectors, it wouldn’t be surprising that they repeatedly rely on only one physically accessible and intuitive case. As such, the hope was that an increase in focus on terminology and corresponding variety of question types may help students to make more precise distinctions between physical cases where an object’s instantaneous velocity was either aligned with or against an instantaneous net force/acceleration and cases where instantaneous net force/acceleration were zero. Moreover, practice with these more open-ended prompts might be more or less effective depending on their implementation as a short answer or multiple choice question. An example of one of these questions is shown in Fig. 4.6. The rest of the training questions are included in Appendix A.3.

![Consider the statements of the two students below. Student A. If there is a net force on an object, the object will move. Student B. An object will continue to move in a straight line if no net force acts upon it. Do the students each mean the same thing when they say 'move'? Explain.]()

Figure 4.6: Example of an explanation-focused training question.

A total of 212 students were randomly assigned to one of four experimental conditions: training via multiple choice questions with feedback (N=52), short answer questions with feedback (N=51), short answer questions without feedback (N=53) and a no-training control (N=56). Students completed the tasks in individual carrels in a quiet testing room. Students were awarded participation credit as part of a course homework assignment.
4.2.1 Results

(a) Student performance on the force and motion assessment

Students' final course grades in their introductory mechanics class were collected in order to provide a general measure of student ability. A one-way ANOVA showed no statistically significant differences between students' overall course grades across the four conditions, $[F(3, 208) = 1.737, p = 0.16]$.

![Mean FVA post-test score by condition](image)

Figure 4.7: Student scores on the 17 question, multiple choice FVA post-test by condition.

The mean scores on the full 17 question, multiple choice FVA post-test are shown in Fig. 4.7. A one-way ANOVA indicated a significant difference between the four conditions $[F(3, 208) = 7.007, p < 0.001]$). In order to assess effectiveness relative to control, we conducted a Dunnett’s post hoc test which indicated a statistically significant difference in scores between the short answer format and control ($p < 0.001, d = 0.90$), a marginally significant difference between multiple choice format and control ($p = 0.07, d = 0.43$), and no difference between short answer without feedback and control ($p = 0.27, d = 0.29$).
In an effort to replicate the findings of Experiment #1, we divided students using a median split on their final course grades, as shown in Fig. 4.8. Using this split, a 2 (upper vs. lower course grade) x 4 (training condition) ANOVA showed main effects from course grade \( (p < 0.001) \), and condition \( (p < 0.001) \), but no significant interaction effect \( (p = 0.956) \). This matched the trend found in Experiment #1: although higher performing students tend to score better on the FVA assessment, there appears to be no significant aptitude-treatment interaction between the different conditions.

In order to more directly compare student performance with feedback across multiple choice and short answer formats, we conducted a corresponding ANCOVA on post-test scores between only the multiple choice and short answer formats with feedback, using course grade as a covariate. Controlling for course grade, there was a marginally significant difference between the two conditions \( [F(1, 99) = 3.714, p = 0.057, d = 0.45] \).
(b) Training Efficiency

The median times spent on the training were 695s for the short answer format, 869s for the short answer format without feedback, and 556 for multiple choice format. A median test showed that the training time was significantly different between conditions ($\chi^2(2) = 23.158, p < 0.001$). Post hoc pairwise comparisons (adjusted with a Bonferroni correction for multiple tests) showed that multiple choice was significantly different from both short answer ($z = 14.8, p < 0.001$) and short answer without feedback formats ($z = 24.8, p < 0.001$). Interestingly, there was also a significant difference in the median time spent on the task between the short answer and short answer without feedback treatments ($z = 8.657, p = 0.010$).

The direction of that difference is surprising. Unlike in Experiment #1, students spent more time on the short answer training when it was presented without feedback than with feedback. What is going on here? Part of the answer seems to be that students in the without feedback condition simply wrote more in their short answer responses - in fact, considerably more. The average length of a student response in the without feedback condition was 21.1 ± 2.0 words. The average length of a student response in the with feedback condition was 9.3 ± 1.2 words. The difference between the two conditions was significant [$t(103) = 4.932, p < 0.001$].

In addition, this dramatic change in the average response length reflects a transition in the qualitative nature of the student responses. Student short answers in the feedback condition tend to shift away from proper sentences to focus on key phrases (i.e., “there is not enough information...”) and reused structures (for example, lists of physically possible directions and cases of motion). This suggests that the knowledge of the computer as an assessor is inviting those corresponding students to focus in on key cases and terminology in order to ensure that their responses can be interpreted.

Of course, time-on-task and response length are not the whole picture. To get a sense of the combination of time-on-task and subsequent student performance on the target MC assessment, we can compare efficiency ratings. The mean efficiency ratings (% score/minutes}
training) for each of the training methods were 7.14 for multiple choice format, 6.37 for short answer, and 4.11 for short answer without feedback. The distribution of individual efficiency ratings for each condition is shown in Fig. 4.9. A one-way ANOVA showed that the difference between these efficiencies is significant, \( F(2,153) = 14.721, p < 0.001 \). A Tukey post hoc showed that the multiple choice format was significantly more efficient than short answer without feedback \( (p < 0.001) \), but not significantly different from short answer with feedback \( (p = 0.393) \).

![Distributions of training efficiency by condition](image)

Figure 4.9: Distributions of training efficiency by condition. Efficiency defined as score %/minutes of time spent on training.

The efficiency metric – combined with the above results – presents a pretty stark picture for the short answer format without feedback. Although students spent more time-on-task and wrote significantly longer responses, the training was comparatively very inefficient. Although the inherent value of feedback isn’t necessarily surprising, the aforementioned
results paint a relatively more positive picture of the interplay between feedback and the verbosity of student responses. Put another way, the fact that students shorten their responses when they know they are assessed and provided feedback by the computer may not necessarily be a negative result – in this case, the resulting efficiency for the short answer with feedback treatment was not significantly different from multiple choice with feedback overall.

(c) Accuracy of automated short answer assessment

Out of 51 students who completed the tasks in the short answer with feedback condition, 1 student demonstrated clearly identifiable gaming behavior. The student submitted a blank response (or letter) on multiple short answer questions (the last response took the student approximately 1.5 seconds). Discarding the responses of this student, the natural language model produced a match for 59% of student responses. The success rate depended strongly on the type of question: the natural language model produced a match for 66% of the corresponding responses to the questions in common with Experiment #1. In contrast, the model only matched 38% of student responses to the new, more open-ended question prompts.

We once again tracked the number of false-positives and false-negatives by comparing all computer identifications to hand-coded grading. We found an overall false-positive rate of 1.1% and a false-negative rate of 1.5%. To test for any negative learning effects of the misidentifications that did occur, we compared performance on the post-test between those students that saw no misidentifications and those that saw at least one, as shown in Fig. 4.10.

For false-positives, an independent samples T-test indicated that the mean score on the force and motion assessment was significantly different between the two groups, \[ t(48) = 3.265, d = 1.5, p = 0.002 \]. Given the potentially important implications, we compared the training results for the small number of students who saw a false-positive to see if it was possible to identify any commonalities. Of the 6 students, 3 of them received the false-positive feedback early on in the training (question 3, which asked the students about the
Figure 4.10: Student performance on the FVA post-test, split by whether a student received a false-positive or a false-negative at any point during their short answer training.

relationship between acceleration and net force), one in the middle (question 5), and the other two near the end of their training (questions 8 and 10, the very last question). We did not find a consistent pattern in student interactions with the training - either in terms of time spent on the training questions, length of student responses, or overall correctness. Similarly, there were no clear shifts in student performance from before to after when the false-positive occurred. However, in the spirit of identifying common qualitative trends, we note that students who received a false-positive did have a nominally lower course-grade average compared to the class, even if not significant given the small sample size \(t(48) = 1.169, d = 0.54, p = 0.25\).

For false-negatives, an independent samples T-test was conducted between the two corresponding groups (5 students received a false-negative, with one student who saw multiple). The results indicated that the difference in mean score on the force and motion assessment between the two groups was marginally significant, \(t(48) = 1.687, d = 0.88, p = 0.098\). However, here the trend was in the other direction. Student course grades were also nominally higher in the group who saw a false-negative, though again not significantly so \(t(48) = 0.971, d = 0.46, p = 0.336\).

In addition, we analyzed the effect of varying frequencies of multiple choice clarifications within the short answer training. As in Experiment #1, we found that there was no correla-
tion between student score on the force and motion assessment and the number of multiple choice clarifications received as part of the short answer training \( r(50) = -0.031, p = 0.831 \). As before, the size of the correlation coefficient was close to zero, reinforcing the previous finding that any overall effect from asking relatively more follow-up multiple choice questions is likely quite small.

(d) Student Performance on Training Questions

Student success on the training questions was measured via comparison of human-graded scores of student responses during training via short answers with feedback to student performance on the multiple choice questions with feedback. Only a student’s initial response to a question was considered, in a manner similar to that in Experiment #1. The results are shown in Fig. 4.11.

The average scores on the multiple choice and stem-equivalent short answer training questions was 50% and 45%, respectively. Scores on the training questions were compared using a general linear model, accounting for a main effect (question format) and a covariate (course grade). While course grade \( F(1,99) = 5.822, p = 0.018 \) significantly predicted student performance on the training questions, question format did not \( F(1,99) = 2.380, p = 0.126 \).

A comparison of performance on individual questions (see Fig. 4.11) indicates that there were only likely differences between the conditions on two question items, Q6 and Q7. Q6 was a \( v \rightarrow F \) question, while Q7 asked students if it was possible for an object to not be moving and still have a nonzero net force (rather than present the question in the common form used on the post-test). Although there were not as many observable differences as in Experiment #1, the relationship between force and velocity seems to be the most likely candidate for differences between the two formats in regard to student availability and success during training.

One of the weaknesses of the current training was a tendency for students to only state a bare minimum answer on the new open-ended questions. For example, students would cite “Student B” to indicate that they agreed with a particular argument, but not provide the
requested justification for their answer. The proportion of students indicating the correct choice (both with and without a corresponding correct explanation) is included in Fig. 4.11 for reference. In order to remedy this, future implementations of short answer formats deviated slightly from their multiple choice counter-parts. In particular, if a student were to only answer with a root answer and no explanation, “Student A” or “Yes”, the program would explicitly ask the student to justify their answer.

Figure 4.11: Student performance on the training questions with the short answer and multiple choice conditions (both with feedback). Dotted lines represent the overall average. Error bars are ±1 standard errors. The proportion of students who indicated only a correct answer (i.e. “Student B”) on Q2, Q5, and Q7 (including students who did not provide a corresponding explanation) are shown separately.

4.2.2 Discussion

Perhaps the most significant - (pardon the pun) - finding of Experiment #2 is a marginally significant difference between the short answer format and the multiple choice format, with a corresponding effect size of $d = 0.49$. Though not quite statistically significant, there are several key differences between this experiment and Experiment #1 that might suggest a meaningful trend.
First and foremost, this experiment implemented several new questions that may have preferentially benefited the short answer format. These questions differed in structure from the rest of the training questions, and primarily focused on helping students to clarify the usage of terms. In addition, they were multi-turn - eliciting follow-up clarifications from the student - and avoided the structure used by previous questions. Given their novelty, the exact mechanism by which these questions may have influenced the effectiveness of the short answer format isn’t completely clear: Was it the follow-up feedback or the basic difference in question variety? Experiments #3 and #4 investigate the first possibility; Experiment #5 the latter.

Second, the participants investigated in this experiment were taken from the first semester of an introductory mechanics class. It is possible that student performance differed due to attitudinal factors separating the two populations. If the students in the introductory mechanics classroom viewed the training as more relevant to their coursework, they may have devoted relatively more effort on the short answer format questions compared to their more senior counterparts. This could help explain the overall similar success in performance on the training questions between the two conditions. For this reason, subsequent experiments included several attitudinal survey questions.

In addition to hinting at potential differences in overall effectiveness, this experiment replicated two important practical findings of Experiment #1. First, the number of multiple choice clarification questions did not correlate significantly with subsequent student performance on the multiple choice post-test. Second, knowledge that the computer will be assessing their answer seems to influence how students frame their short answers. Although the content of the feedback itself is likely having some effect on response length, the qualitative nature of student responses - focusing in on repeated directions and phrases - suggests that students were making an active effort to make their response interpretable.

Finally, it is worthwhile to consider the complicated relationship between misidentifications during the short answer training, student aptitude, and potential interactions with subsequent student performance. On the one hand, it is possible that the observed decrease in student performance is largely the result of genuine confusion on the part of the student
due to the misidentification. However, it is also possible the difference in post-test performance is reflecting another underlying difference, namely that particular students were simply more likely to receive a false-positive as a direct consequence of lower prior ability. Put another way, lower prior ability may correspond with an increased probability of mistakes during training, and thus a greater number of possible opportunities for the computer to incorrectly mark a mistake as correct. In contrast, stronger students are more likely to provide correct answers that are then marked incorrectly.

Although we cannot completely distinguish between the two cases, the takeaway is that there is evidence that a particular group of students may be at risk with the current implementation of feedback: namely students of low prior knowledge. The corresponding difference in mean grade between those who did and not see a misidentification was not significant given the small sample size. However, the trend in grade was the same as in Experiment #1. There are further structural reasons that might suggest this potential confound. Here, and in Experiment #1, if a student answer was marked correct, the student was only told that they were correct and invited to continue. It was only when a student answer was incorrect that they received knowledge of the correct answer and a brief explanation. As such, regardless of which mechanism is at play (or both), this measured difference suggests difficulties for students with low prior knowledge: they are simultaneously more likely to run into a false-positive, and they may also be more likely to lack existing resources to recognize and appropriately resolve it.

Up to this point, feedback had been structured in this manner to make a clear distinction to the student that their answer was being assessed, and that feedback was specific to their response (in order to match the functionality most natural to multiple choice formats). Given the measured effects above, one possible option would be to remove feedback focused on the status of correct/incorrect and only provided the correct answer.

However, it is important to note the dramatic changes between students with short answer training who received no feedback and those that did in terms of amount written and time-on-task. Combined with the suggestive trend of the short answer format measured here (namely a marginally significant difference between short answer and the multiple
choice format above and beyond control) it seems the role of computer-as-assessor may be particularly important. Or, to put it more generally, these results seem to support this facet of assessment as another unique and important axis to consider in the design of instructional materials (in addition to the parameter space mapped about by the permutations of what type of feedback is provided).

Based on these considerations, the following experiments implement feedback that always includes correct/incorrect status, the correct answer, and a brief explanation across all formats in order to reduce the potential confound of misidentifications, while maintaining a sense of active assessment for the student. As such, the following experiments shift focus slightly from the role of training-question difficulty, the impact of multiple choice clarifications, and assessment accuracy to measure other potentially influential factors, such as attitudinal effects, retention performance, and potential interactions with question format.

4.3 Experiments #3 and #4: Effects of Follow-up Questions

4.3.1 Design and Participants

(a) Motivation

The results of the previous experiment suggest two possible avenues for further investigation, given the suggestive nominal difference in performance between short answer and multiple choice formats – an effect size of approximately half a standard deviation, though not statistically significant. Here we consider the first hypothesis that increases in the level of interaction through multi-turn feedback may lead to potentially different gains for the respective formats.

There are a number of reasons to expect that an increase in the interactivity of the feedback provided to the student might influence the relative effectiveness of a short answer or multiple choice format. First and foremost, there are practical differences between the two formats, in additional to potential differences in mechanism (ie. recall vs recognition). Multiple choice formats typically provide more information to the student – both explicit (identifying potential distractors) and implicit (question expectations). As such, follow-up
questions that clarify the intent of a question, further elicit potential distractors, or invoke missing cases might help the short answer format to achieve a comparable grounding of question information and expectations.

There is also the possibility for attitudinal interactions with question format and the degree of interaction via follow-up questioning. On the one hand, the short answer format might benefit from increased interaction. As the training shifts away from isolated questions to more extended discussions, students may start to view the training as more conversational, and perhaps consequently, consider the training in an overall more positive light. Proponents of conversational computer tutors have typically framed the motivation for their development as an appeal to the authentic and effective nature of conversation in one-on-one tutoring settings (Graesser et al., 2005). In contrast, the additional reading demanded by nested multiple choice follow-up questions may place compounding demands on students in those conditions and corresponding negative attitudinal effects.

However, prior research suggests that there may be important constraints on the additional benefit from increases in the interactivity of feedback (and thus potential differences between question formats). This postulated upper limit in effectiveness due to more fine-grained feedback is known as the interaction plateau hypothesis (VanLehn, 2008). As a result of a meta-analysis of computer-based instruction, VanLehn (2008) argued that feedback effectiveness tends to plateau near the grain size of step-based feedback – past that point, further feedback granularity leads to diminishing or even no further returns. As a result, one of the important empirical questions here is whether or not we can break down the cognitive questions of our domain into meaningful steps for the purpose of training.

Although the relationships between the direction of force, acceleration and velocity in one dimension represent a very simple and restricted training space, one of its interesting characteristics is that prior research has measured the existence of meaningful intermediate states (Rosenblatt and Heckler, 2011). For example, students may recognize that the instantaneous velocity of an object need not necessarily be aligned with the force, but consistently only consider one other physical case (such as anti-aligned vectors). As such, students may consistently answer that the velocity “cannot-be-zero” (Rosenblatt and Heck-
ler, 2011). These intermediate states hint that it may be possible to get a student partway towards a correct answer: to help students shift from an incorrect state to a partially correct understanding, or from a partially correct state to a fully correct one. As such, we intend to assess the benefit from using follow-up questions to target the physical cases omitted by students in these intermediate states.

(b) Design

Given this, we designed a series of follow-up feedback questions to both elicit physical cases that the student did not consider (i.e. an object turning around), and to make explicit the implications of the cases they did (i.e. what do anti-aligned acceleration and velocity vectors imply about the speed of the object over time?). Follow-up feedback queries for questions of the $X \rightarrow Y$ type were implemented in collaboration with Dr. Andrew Pawl.

The precise number and nature of the follow-up questions that a student received depended on their initial answer, but the path was the same for any student with a similar response. As such, these feedback questions are much more like the feedback trees or explicit knowledge construction dialogs employed and studied elsewhere (Jordan, 2012), and differ from more open-ended intelligent tutors where the tutorial responses may vary considerably based on initial student answers and subsequent interactions. A schematic illustration of one of these trees ($A \rightarrow V$) is indicated in figure 4.12.

Although there may be a dozen unique nodes for each question, students would typically only see a maximum of 3-4 follow-up questions. For questions whose initial prompt was formulated as a $X \rightarrow Y$ question, the final tutorial step was to ask the student to retry the original question given what had previously been discussed. The student was then told whether they were correct/incorrect, the correct answer and a corresponding brief explanation. In contrast, if a student had been marked correct on their initial response, they were told they were correct via the same feedback used in the final step and invited to continue to the next question immediately.
Could be same or opposite A or zero

Velocity same as Acceleration (includes "same or zero")

Velocity Opposite Acceleration (includes "opposite or zero")

Velocity Zero

If velocity zero but nonzero acceleration, what will happen if we watch [subject] for a few seconds?

Begin to move

What direction?

Begin to move in direction of acceleration

It will go in direction of acceleration.

Any Other

What will happen to [subject’s] speed?

Stays steady

Any Other

Acceleration is defined as change in V over time. Must change. Will (increase/decrease)

Gets (faster/slower)

Will (increase/decrease)

How would velocity and acceleration have to be related to get [subject] to (slow down/speed up)

Velocity and acceleration must be (opposite/same) direction.

Any Other

The original question did not specify if the acceleration was speeding [subject] up or slowing [subject] down. With that fact in mind, what can you say about the direction of [subject’s] velocity?

Could be with or against acceleration.

Correct!

If [subject] is speeding up, we would expect V | | A (or possibly V=0 for an instant). If subject is slowing down, we expect V opposite A. Thus, we really know nothing about the distance of [subject’s] velocity.

Any Other

Incorrect...

If [subject] is speeding up, we would expect V | | A (or possibly V=0 for an instant). If subject is slowing down, we expect V opposite A. Thus, we really know nothing about the distance of [subject’s] velocity.

Any Other

Figure 4.12: Illustration of follow-up feedback structure for a question of the type $A \rightarrow V$. 
Table 4.1: Experiment #3-4: Number of students per experimental condition for both experiment administrations (algebra-based and calculus-based).

<table>
<thead>
<tr>
<th>Condition</th>
<th>Calculus-based (Fall 2015)</th>
<th>Algebra-based (Spring 2016)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-training control</td>
<td>N = 55</td>
<td>N = 50</td>
</tr>
<tr>
<td>Multiple choice (single questions)</td>
<td>N = 58</td>
<td>N = 50</td>
</tr>
<tr>
<td>Short answer (single questions)</td>
<td>N = 55</td>
<td>N = 48</td>
</tr>
<tr>
<td>Multiple choice (follow-up questions)</td>
<td>N = 59</td>
<td>N = 52</td>
</tr>
<tr>
<td>Short answer (follow-up questions)</td>
<td>N = 58</td>
<td>N = 50</td>
</tr>
</tbody>
</table>

(c) Participants

There were 5 experimental conditions: short answer with follow-up questions, multiple choice with follow-up questions, short answer (individual questions), multiple choice (individual questions), and a no-training control. These conditions are illustrated schematically in Fig. 3.1.

In order to assess performance with two different groups of students, this experimental design was administered in two different introductory physics courses: a first semester calculus-based course and a first semester algebra-based course. Whereas the calculus-based course was predominantly composed of engineering and physical science students, the algebra-based course had higher enrollments of life science, chemistry, and health-related field majors. A total of 535 students from the two introductory physics courses participated in these experiments; the number of students assigned to each condition is indicated in Table 4.1.

(d) Administration

For both experiments, students completed the research tasks as part of a flexible homework assignment in their respective physics course. In addition to attending a one-hour research session, students completed an online pre-test and a retention test. The pre-test was composed of a 10 question subset of the FVA test used here and in previous experiments. The retention test was a 5 question subset of those questions used on both the pre-test and
post-test. While the pre-test was administered as a stand alone task, the retention test was included as a single question block within a larger, online end-of-semester assignment (consisting of approximately 30 questions in total).

An outline of the experiment administration is shown in Fig. 4.13. Students first completed the online pre-test, approximately 1-week prior to attending their one-hour flex appointment. When they arrived, they were randomly assigned to one of the 5 experimental conditions. After completing their respective treatment, students completed a set of unrelated physics tasks (10-20 minutes), and then a combined force and motion post-test. The post-test consisted of the full 17 multiple choice item FVA test and 3 short answer questions. After completing the post-test, students were given a brief attitudinal survey to complete at the conclusion of the session. Approximately 6-8 weeks later students completed the retention task (the time varied based on students’ initial appointment date). Although the vast majority did (> 85% over the two experiments), not every student who completed the flex assignment completed the subsequent retention test. Those students were included in the primary analysis but excluded in the related analyses of retention scores.

Figure 4.13: Experiment administration and task sequence.
4.3.2 Results

Students’ pre-test scores and final course grades in their introductory mechanics class were collected in order to provide measures of student ability. For the calculus-based students, a pair of one-way ANOVAs showed no statistically significant differences between students overall course grades across the five conditions, \( F(4, 268) = 0.478, p = 0.752 \) or their pre-test scores \( F(4, 273) = 0.523, p = 0.719 \). For the algebra-based students, corresponding one-way ANOVAs also showed no statistically significant differences between students overall course grades across the five conditions, \( F(4, 245) = 0.671, p = 0.613 \), or pre-test scores \( F(4, 243) = 0.481, p = 0.750 \).

(a) Student performance on the multiple choice assessment

The mean scores on the 17 question FVA post-test are shown in Table 4.2.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Calculus-based</th>
<th>Algebra-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MC score</td>
<td>OT score</td>
</tr>
<tr>
<td>No-training control</td>
<td>9.47 ± 0.6</td>
<td>0.09 ± 0.05</td>
</tr>
<tr>
<td>Multiple choice (single questions)</td>
<td>10.0 ± 0.6</td>
<td>0.36 ± 0.08</td>
</tr>
<tr>
<td>Short answer (single questions)</td>
<td>10.7 ± 0.6</td>
<td>0.53 ± 0.13</td>
</tr>
<tr>
<td>Multiple choice (follow-up questions)</td>
<td>11.3 ± 0.6</td>
<td>0.49 ± 0.12</td>
</tr>
<tr>
<td>Short answer (follow-up questions)</td>
<td>11.7 ± 0.6</td>
<td>0.34 ± 0.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MC score</th>
<th>OT score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6.9 ± 0.5</td>
<td>0.16 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>8.7 ± 0.5</td>
<td>0.92 ± 0.14</td>
</tr>
<tr>
<td></td>
<td>7.9 ± 0.5</td>
<td>0.78 ± 0.15</td>
</tr>
<tr>
<td></td>
<td>8.5 ± 0.6</td>
<td>0.60 ± 0.11</td>
</tr>
<tr>
<td></td>
<td>9.1 ± 0.6</td>
<td>0.94 ± 0.18</td>
</tr>
</tbody>
</table>

Notes: MC score is the mean score on the 17 question multiple choice FVA test. OT score is the mean calculated “overtraining score” described in the results. Errors shown are ±1 standard errors.

For the calculus-based class, a one-way ANOVA indicated a significant difference between experimental conditions \( F(4, 280) = 2.556, p = 0.039 \). Subsequent Dunnett’s post hoc comparisons to control indicated a significant difference between short answer with follow-up questions and control \( p = 0.023, d = 0.48 \), a marginally significant difference between multiple choice with follow-up and control \( p = 0.087, d = 0.43 \), but no sig-
significant difference between control and the single question versions of short answer and multiple choice (respectively, \( p = 0.367, d = 0.30 \) and \( p = 0.926, d = 0.12 \)). For the algebra-based class, a similar one-way ANOVA for total score was marginally significant \( F(4, 249) = 2.233, p = 0.066 \), with only the short answer – follow-up condition being significantly different from control \( p = 0.023, d = 0.56 \). As such, only the short answer – follow-up condition was significantly better than control for both courses.

In addition to the entire FVA-test score, we considered the gain on the 10 common items represented by the pre-test; these items were all of the general form \( X \rightarrow Y \). The mean gain in this common 10 question-subscore are shown by condition in Fig. 4.14. We first compared the treatments to control and then across question formats and level of follow-up interaction.

![Figure 4.14: Mean gain in pre-test to post-test score for the 10 in-common \( X \rightarrow Y \) questions. Error bars shown are ±1 standard errors.](image)

For calculus-based students, a one-way ANOVA between conditions was significant \( F(4, 273) = 2.912, p = 0.022 \). Subsequent Dunnett’s post hoc comparisons to control indicated a significant difference between short answer – follow-up and control \( p = 0.005 \), and marginal significant differences between multiple choice – follow-up and control \( p = 0.068 \) and short answer – single and control \( p = 0.089 \). There was no significant difference
between multiple choice – single and control \((p = 0.383)\). To check for differences between the treatments, we conducted a \(2 \times 2\) (question format vs. follow-up feedback) ANCOVA on the gain, controlling for pre-test score. The effect of follow-up feedback was marginally significant \((p = 0.094)\); there was no significant effect from question format \((p = 0.174)\).

For algebra-based students, the one-way ANOVA between conditions was also significant \([F(4, 238) = 4.798, p = 0.001]\), but unlike in the calculus course, all treatments were significant (with the exception of short answer - single question, which was marginal): short answer follow-up \((p < 0.001)\), multiple choice follow-up \((p = 0.012)\), short answer single \((p = 0.058)\) and multiple choice single \((p = 0.002)\). As such, it appears part of the comparatively lower effectiveness for the calculus course may be the nominally larger testing effect demonstrated by the control condition in that experiment. For the algebra-based students, there were once again no significant main effects from either question format or follow-up feedback.

In order to monitor incorrect student over-generalization due to the training, we computed an additional metric that we refer to as an “overtraining score” (see Table 6.8). The OT score represents a rough metric by which to assess improper student over-generalization driven by the training. In essence, the OT score is simply the number of times a student incorrectly claims that there is no relationship between the interested quantities (for example, between net force and acceleration) when there should be. This choice is typically presented as the last choice for questions of the form \(X \rightarrow Y\), and represents a sort of “all of the above” answer.

Given that the training predominantly focused on getting students to recognize that force and velocity (and acceleration and velocity) need not necessarily be aligned – and the knowledge that students typically perform very well on questions asking about the relationship between the directions of the net force and acceleration (> 80%) – we tallied instances where students incorrectly applied no relationship/all-of-the-above reasoning to those questions as a measure of over-training. As such, a mean OT score of 1 would suggest on average students answered that there was no relationship between such quantities on 1 question on the post-test. Taken together, the training conditions did significantly
increase the over-generalization for both the calculus-based \((t(283) = 3.161, p = 0.002)\) and algebra-based physics classes \((t(253) = 4.263, p < 0.001)\). However, the overall increase in this incorrect generalization amounted to less than a single question, with no significant differences based on question format or the interaction level of the training. It is worthwhile to note that algebra-based students were far more likely to make this type of over-training error compared to calculus-based students with the same training regardless of treatment condition.

In summary, the training demonstrated small to medium gains (typical effect sizes in the range of \(d = 0.3 - 0.5\)) versus control, but only the short answer with follow-up treatment was significantly better than control for both experiments. Moreover, there was no overall main effect within the treatment conditions, either in terms of question format (multiple choice vs. short answer) or the amount of interaction through follow-up questions (single vs. follow-up). Finally, although there was evidence that the training did lead students to incorrectly generalize the claim of no-relationship between certain force and motion quantities to questions where it wasn’t appropriate, that effect did not differ significantly across question format or follow-up interaction. On the other hand, the algebra-based student population were more susceptible to over-training, across all conditions.

(b) Retention

There are two important questions we would like to address in terms of potential retention of student gains due to training. First, are there any gains 6-8 weeks after training? Second, if so, did question format and follow-up feedback significantly influence retention success? To answer these questions, we considered traces of 5 common items replicated on the pre-test, post-test, and retention test.

To the first question, Fig. 4.15 plots the combined mean score for all treatment conditions on the pre-test, post-test, and retention-test versus control. In order to measure whether there was any meaningful retention from the training, we first conducted a repeated measures ANOVA (pre-test to retention-test) comparing collective training to control. There was a significant effect from time \([F(1, 208) = 25.610, p < 0.001, \text{partial } \eta^2 = .11]\)
and a significant time-training interaction \(F(1, 208) = 4.229, p = 0.041, \text{partial } \eta^2 = 0.02\), suggesting that the training - taken as a whole - represented significant retention above control. The interaction is particularly important given the potential testing effect observed for the control condition earlier, and the possibility of future learning from continued general physics instruction. We can triangulate this finding with an independent samples t-test between the combined training condition and control on the final retention scores. The
difference was significant \( t(208) = 2.033, p = 0.043 \). Taken together, these results suggest that the training as a whole did lead to persistent gains above control.

For the second question, in order to ascertain whether there were effects from how students were trained (i.e. short answer vs. multiple choice) on retention after treatment, we completed a separate analysis looking at just the drop in student performance from post-test to retention within only the intervention conditions. A \( 2 \times 2 \times 2 \) mixed ANOVA was conducted with 2 between-subjects measures (question format and level of follow-up) and one within-subject repeated measure (student performance on the 5-question subset immediately after training and on the delayed retention test). There was a significant result of time \( [F(1,165) = 6.442, p = 0.012, \text{partial } \eta^2 = 0.038] \), suggesting a meaningful decay in student performance over that time period. However, there were no significant interactions with time and question format \( (p = 0.661, \text{partial } \eta^2 = 0.001) \), follow-up \( (p = 0.150, \text{partial } \eta^2 = 0.013) \), or higher-order interactions.

The above analysis was repeated for the algebra-based physics course. We first conducted a repeated measures ANOVA (pre-retention) comparing collective training to control. There was a significant effect from time \( [F(1,241) = 11.957, p < 0.001, \text{partial } \eta^2 = .05] \) but no significant time-training interaction \( [F(1,241) = 2.358, p = 0.126, \text{partial } \eta^2 = 0.01] \). Similarly, a t-test on the retention scores (control vs. training) found no overall difference between the retention scores for the combined training versus control \( (t(248) = 1.343, p = 0.18) \). To further parse this effect, we conducted individual pairwise comparisons between pre-test and retention scores. These individual comparison showed significant differences for short answer with single questions, short answer with follow-up and multiple choice with single questions conditions \( (ps < 0.05) \). Retention and pre-test scores were not significantly different for the control condition \( (t(46) = 1.218, p = 0.23) \) or multiple choice with follow-up \( (t(51) = 0.698, p = 0.488) \). The lack of retention for the multiple choice follow-up, and the relatively smaller gains explain the lack of overall retention effect. As such, it appears that there was retention for several of the treatments, but the overall level of material retained across the trainings did not differ significantly from control. To further track down the decay of the training gains, a \( 2 \times 2 \times 2 \) mixed ANOVA was
conducted with 2 between-subjects measures (question format and level of follow-up) and
one within-subject repeated measure (student performance on the 5-question subset imme-
diately after training and on the delayed retention test). There was a significant result of
time \[F(1,165) = 39.35, p < 0.001, \text{partial } \eta^2 = 0.167\], suggesting a meaningful decay in
student performance from post-test to retention test. However, there were no significant
interactions with time and question format, follow-up, or higher-order interactions.

(c) Student performance on short answer questions

In addition to the 17-question FVA test, the post-test included 3 short answer question
items, administered after the multiple choice test. The first short answer question was an
\(F \rightarrow v\) question, which stated information about the forces on an object and asked students
to describe the subsequent motion. The second question was a restatement of one of the
training questions (question 5 in Appendix A.3). The third question was intended as a
far-transfer question; it asked students to describe the motion of an object based on an
initial condition and a graph of force over time. These questions are included in Appendix
A.3. Students’ overall performance on the short answer items is shown in Table 4.3.

For calculus-based students, a one-way ANOVA indicated a significant difference be-
tween experimental conditions on overall short answer score \[F(4,276) = 2.996, p = 0.019\].
Subsequent Dunnett’s post hoc comparisons indicated that only the two short answer treat-
ments (with and without follow-up feedback respectively) were significantly different from
control (\(p = 0.007\) and \(p = 0.024\)). To verify the implied main effect, we conducted a
2 \(\times\) 2 ANCOVA with question format and level of follow-up as between-subject factors and
pre-test score as a covariate. There was a marginally significant effect of question format
\[F(1,222) = 3.627, p = 0.058, \text{partial } \eta^2 = 0.016\], but no effect of follow-up (\(p = 0.500\)) or
interaction between format and follow-up level (\(p = 0.775\)).

For the algebra-based course, a similar one-way ANOVA showed significant differences
between conditions \[F(4,276) = 2.996, p = 0.019\]. However, in contrast, subsequent Dun-
nett’s post hoc comparisons indicated that all of the treatment conditions were significant
(or close): short answer follow-up (\(p < 0.001\), multiple choice follow-up (\(p = 0.023\), short
answer \((p < 0.001)\), multiple choice \((p = 0.056)\). In part, this was due to the relatively lower floor represented by the algebra-based control condition. A \(2 \times 2\) ANCOVA with question format and level of follow-up as between-subject factors and pre-test score as a covariate showed a significant effect of question format \([F(1,191) = 7.820, p = 0.006\) partial \(\eta^2 = 0.039]\), but no effect of follow-up \((p = 0.874)\) or interaction between format and follow-up level \((p = 0.697)\).

Table 4.3: Overall short answer score by condition (algebra-based and calculus-based). Errors shown are standard errors.

<table>
<thead>
<tr>
<th>Overall short answer score</th>
<th>Calculus-based</th>
<th>Algebra-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-training control</td>
<td>41% ± 4%</td>
<td>22% ± 4%</td>
</tr>
<tr>
<td>Multiple choice (single questions)</td>
<td>50% ± 4%</td>
<td>36% ± 5%</td>
</tr>
<tr>
<td>Short answer (single questions)</td>
<td>57% ± 3%</td>
<td>49% ± 4%</td>
</tr>
<tr>
<td>Multiple choice (follow-up questions)</td>
<td>51% ± 4%</td>
<td>38% ± 4%</td>
</tr>
<tr>
<td>Short answer (follow-up questions)</td>
<td>59% ± 4%</td>
<td>49% ± 4%</td>
</tr>
</tbody>
</table>

Next, we consider the performance on the individual question items as shown in Fig. 4.16. Visual inspection indicates that the three short answer questions responded differently to training. First, there was no evidence of any differences amongst conditions for the transfer question, either for calculus-based \([F(4,280) = 0.402, p = 0.81]\) or algebra-based courses \([F(4,245) = 1.435, p = 0.223]\). The second question only demonstrates an effect from the training over control. Training conditions performed significantly higher than control for both the calculus-based and algebra-based settings: \(t(283) = 4.077, p < 0.001\) and \(t(248) = 4.608, p < 0.001\) respectively. Given that a rephrased version of that question was presented in the training, this overall difference is not particularly surprising.

Consequently, much of the structure observed in the overall short answer score is primarily coming from the first \(F \rightarrow v\) question. For the calculus-based students, a one-way ANOVA replicates part of the general structure found for the overall short answer score: the ANOVA was significant \([F(4,280) = 2.708, p = 0.031]\) and Dunnett’s post hocs revealed the
Figure 4.16: (A) Student performance on $F \rightarrow v$ short answer question. (B) Student performance on the trained question. (C) Student performance on the transfer question. Error bars shown are ±1 standard errors.

short answer – single treatment to be significantly different than control ($p = 0.017$). The short answer – follow-up condition was not significant ($p = 0.116$). The two multiple choice
treatments with and without follow-up ($p = 0.855$ and $p = 0.898$ respectively) were also not significantly different from control. A $2 \times 2$ ANCOVA with question format and level of follow-up as between-subject factors and pre-test score as a covariate showed a significant effect of question format [$F(1, 222) = 7.043, p = 0.009$, partial $\eta^2 = 0.031$] but no significant effect from follow-up ($p = 0.260$) or interaction ($p = 0.652$).

In comparison, the algebra-based course found significant differences for both short answer formats, potentially as a result of the lower floor. More specifically, the ANOVA was significant [$F(4, 245) = 4.829, p = 0.001$]. Both short answer formats were significantly different from control, with ($p = 0.002$) and without follow-up ($p = 0.002$); both multiple choice formats were not, with ($p = 0.616$) or without follow-up ($p = 0.576$). A $2 \times 2$ ANCOVA with question format and level of follow-up as between-subject factors and pre-test score as a covariate showed a significant effect of question format [$F(1, 191) = 10.838, p = 0.001$ partial $\eta^2 = 0.054$], but no effect of follow-up ($p = 0.935$) or interaction between format and follow-up level ($p = 0.975$).

In order to better qualitatively analyze student responses to this question, we looked at the proportion of students making the specific mistake of assuming the net force and velocity had to be aligned. Taken together, the training conditions significantly reduced the proportion of students making this error from 33% in the control condition to 20% in the trained conditions for calculus-based physics ($\chi^2(1) = 4.129, p = 0.042$) and from 56% (implying that yes, a student in the control condition for this population was more likely than not to make this mistake) to 31% in the algebra-based class ($\chi^2(1) = 10.851, p = 0.001$). There was no significant difference in the proportion of students making the error between short answer (16%) and multiple choice formats (24%) in the calculus-based class ($\chi^2(1) = 2.301, p = 0.129$). There was however a reduction in the proportion of students in the algebra-based class, from 38% in multiple choice conditions to 25% in short answer formats ($\chi^2(1) = 3.807, p = 0.051$). Taken together, these results suggest that the short answer formats had a small-to-medium effect on a particular type of question, namely a question of the form $F \rightarrow v$. There did not appear to be any differences for the other short answer questions.
(d) Timing data & efficiency analysis

The typical training time required by each of the interventions is indicated in Table 4.4. For the calculus-based students, an Independent-samples Median test indicated that there was a significant difference in the median training time across all treatment conditions \[ \chi^2(3) = 78.4, p < 0.001 \]. Pairwise comparisons, with a corresponding Bonferroni correction, were all significant (\( ps < 0.01 \)). The equivalent test for the algebra-based students also indicated a significant difference in the median training time \[ \chi^2(3) = 73.5, p < 0.001 \]. Taking into account a Bonferroni correction for multiple comparisons, all pairwise comparisons were statistically significant (\( p < 0.01 \)), except for the comparison between short answer – single and multiple choice – with follow-up (\( p = 0.66 \)). As such, it is clear that there are very significant differences in the total time required by the training, representing a significant addition in time from shifting from multiple choice to short answer, and from single questions to questions with follow-up interactions.

Table 4.4: Mean and median time spent (in minutes) on training and corresponding efficiency for each training condition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Calculus-based</th>
<th>Algebra-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple choice (single questions)</td>
<td>8.0 (2.7)</td>
<td>8.1 (2.8)</td>
</tr>
<tr>
<td>Short answer (single questions)</td>
<td>13.6 (4.5)</td>
<td>11.7 (3.7)</td>
</tr>
<tr>
<td>Multiple choice (follow-up questions)</td>
<td>10.8 (3.1)</td>
<td>13.0 (3.7)</td>
</tr>
<tr>
<td>Short answer (follow-up questions)</td>
<td>17.9 (5.9)</td>
<td>20.5 (6.3)</td>
</tr>
</tbody>
</table>

Note: Standard deviation indicated by (values). Efficiency represents the mean of individual efficiency scores, calculated as \( \epsilon_i = (Post_i - Pre_i) / (\sigma_{control} \cdot time_i) \). Training time is in units of 10 minutes. Errors shown with ±1 standard errors.
The table also presents a mean efficiency value. Here, we define efficiency as \( \epsilon_i = (Post_i - Pre_i) / (\sigma_{control} \cdot time_i) \), where the standard deviation is with respect to student scores in the control condition on the post-test (specifically for the common 10 questions repeated on the pre-test). The standard deviation of student scores for the control condition did not differ significantly between pre and post-testing. In addition, the time unit has been scaled to 10 minutes in order to make the values of \( \epsilon \) more easily readable. Although there are other metrics that we could use to represent a measure of performance gains as a rate of training time, this one is useful for its conceptual clarity. In essence, this efficiency rating can be interpreted as “gain pre-to-post in units of control standard deviation per 10 minutes spent training.” Other efficiency metrics - using total multiple choice score or combined short answer and multiple choice scores - suggest similar qualitative results.

The major take-away is that single question multiple choice is nominally the most time efficient treatment condition for both courses. For the calculus-based course, an ANOVA on efficiency was marginally significant across conditions \( [F(3, 221) = 2.242, p = 0.084] \). For the algebra-based course however, the ANOVA was significant \( [F(3, 192) = 3.480, p = 0.017] \). A Tukey HSD post hoc indicated that multiple choice – single was significantly more efficient than multiple choice with follow-up \( (p = 0.042) \) and short answer with follow-up formats \( (p = 0.027) \), and marginally significant compared to short answer single questions \( (p = 0.089) \).

In addition to overall training time, we considered the time students spent on the feedback provided as part of the training. The feedback time for the follow-up conditions was taken to only include the time students spent on the end-of-question feedback (that is, the final statement of correct/incorrect, along with the correct answer and justification). In addition to providing a comparison of the in-common feedback presented to the students across all conditions, this was also done for practical reasons - namely that there was no way post-administration to separate out how much time the student spent reading the previous feedback (provided along with the next follow-up question) and the time the student spent constructing their next response. For the end-of-question feedback, the amount of time students spent on the feedback was measured from the time they submitted their answer.
until they clicked to continue to the next question.

The time spent on feedback for each question, for each training condition, as well as averages over question format and follow-up level is shown in Fig. 4.17. There are several valuable trends and overall effects. First, we note that time spent on feedback generally decreases throughout the course of the training - a finding common to many forms of computer-based instruction.

Second, students focus on particular feedback in similar ways across the two physics courses. For example, question 3 represents a dip in the time students typically spend on the feedback. For all trainings, this corresponded to a question of the form $F \rightarrow a$. 

Figure 4.17: Time spent on feedback for each training question by condition. (Left) Time spent by condition. (Center) Time spent by question format. (Right) Time spent by follow-up level. Error bars are ±1 standard errors.
As our earlier experiments have shown, questions focused on the relationship of Newton's Second Law tended to have the highest ratio of student success. Question 5 represents a significant bump, especially for the short answer students. This question asked students to think about and clarify the word “move” - making a distinction between cases of constant velocity and acceleration. As this question was not of the form $X \to Y$, students in the short answer conditions spent more time making sense of the feedback, compared to the multiple choice students (who likely benefited from the provided choices and inherent scaffolding for the question’s expectations). Finally, question 9 represents a special case of the $F \to v$ relationship, namely the special case: $F_{Net} = 0 \to v$. This echoes the trend of question 5, suggesting that whenever the underlying structure of the question varied, there was a noticeable spike in the time students in the short answer conditions spent on that feedback.

In contrast, the multiple choice students typically did not demonstrate as large a jump in time, possibly because that processing had been rolled into time during reading the question.

Third, there seem to be overall trends in the total duration of time spent on feedback based on the question format and type. For the calculus-based course, a $2 \times 2$ ANOVA on total feedback time with question format and level of follow-up as between-subject factors showed significant effects of question format [$F(1, 224) = 55.285, p < 0.001$, partial $\eta^2 = 0.198$] and follow-up [$F(1, 224) = 15.722, p < 0.001$, partial $\eta^2 = 0.066$] but no significant interaction of the two factors ($p = 0.720$). For the algebra-based course, the $2 \times 2$ ANOVA also showed significant effects of question format [$F(1, 196) = 20.385, p < 0.001$, partial $\eta^2 = 0.094$] and follow-up [$F(1, 196) = 5.268, p = 0.023$, partial $\eta^2 = 0.026$] but again no interaction of the two factors ($p = 0.551$). These results suggest that the two main effects are consistent across the two classes and relatively large in size. Students in the short answer conditions spend significantly longer on the feedback than students in the multiple choice conditions, while students in the follow-up conditions spend significantly less time on the feedback than students in the single question formats (likely because they have viewed that relevant feedback through the scaffolding of the follow-up questions). Moreover, Fig. 4.17 suggests that the effect for question format is even larger for the first few training questions, as students try and understand the expectations of the task. This may suggest
to front-load challenging questions with important feedback at the start of short answer focused trainings.

(e) Attitudinal data

Students were asked a set of brief attitudinal questions at the completion of their one-hour session, as shown in Table 4.5. The first two questions asked students to predict typical performance on the force and motion assessment. Student predictions of their own performance on the post-test were significantly correlated with their actual scores for both courses and across all conditions ($r = 0.42$ for the calculus-based control, $r = 0.49$ for calculus-based training, $r = 0.37$ for the algebra-based control, and $r = 0.50$ for algebra-based training). Although students appeared to have a good sense of their relative ability, there was an overall tendency to over-estimate proficiency (as evidenced by the difference between average scores on the post-test and the average predicted score). In part, this effect seems to be linked to the amount of feedback and interaction provided during the training - students predicted higher scores in the conditions where they received the more interactive follow-up questions. Separate $2 \times 2$ ANOVAs over format and follow-up levels on predicted score showed a significant effect from follow-up questions [$F(1, 225) = 13.530, p < 0.001$, partial $\eta^2 = 0.057$] and algebra-based courses [$F(1, 195) = 9.974, p = 0.002$ partial $\eta^2 = 0.049$]. There were no significant main effects of question format in either course. Given that previous results found no significant main effect from follow-up feedback on performance on the multiple choice test, it seems that students may be over-estimating their understanding (though the trend is mostly in the right direction).

The role of follow-up feedback also seems to significantly affect the degree to which the students consider the training questions useful. For the calculus-based experiment, a $2 \times 2$ ANOVA over format and follow-up levels on student rating of question usefulness found a significant effect of follow-up questions [$F(1, 225) = 9.211, p = 0.003$], but no format main effect. A complementary $2 \times 2$ ANOVA for the algebra students also found a significant effect of follow-up questions on student rating of the question usefulness [$F(1, 194) = 8.627, p = 0.004$].
Table 4.5: Student responses to attitudinal exit survey by condition. The top-line for each question is the mean score for the calculus-based course, the bottom-line is the mean score for the algebra-based course. Errors shown are ±1 standard errors.

<table>
<thead>
<tr>
<th>Condition (Top = calculus, bottom = algebra-based)</th>
<th>Multiple Choice</th>
<th>Short Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean FVA post-score</td>
<td>Control</td>
<td>Single</td>
</tr>
<tr>
<td>56%</td>
<td>59%</td>
<td>66%</td>
</tr>
<tr>
<td>41%</td>
<td>51%</td>
<td>50%</td>
</tr>
</tbody>
</table>

Survey question

Using a standard 100% scale, how well do you think you did on the force, velocity, and acceleration post-test?

- 80% ± 2% 72% ± 2% 79% ± 2% 73% ± 2% 82% ± 2%
- 77% ± 2% 66% ± 3% 71% ± 2% 67% ± 3% 77% ± 2%

Using a standard 100% scale, how well do you think the typical student in this course does?

- 77% ± 1% 71% ± 2% 74% ± 2% 74% ± 1% 80% ± 1%
- 71% ± 2% 69% ± 3% 70% ± 2% 67% ± 3% 75% ± 1%

* In the first force-velocity-acceleration task, were the [free-response/multiple choice] format questions useful?

- - - 3.59 ± 0.12 4.03 ± 0.11 3.60 ± 0.14 3.89 ± 0.12
- - - 3.78 ± 0.16 3.96 ± 0.14 3.25 ± 0.16 3.94 ± 0.13

* In the first force-velocity-acceleration task, was the feedback on the [free-response/multiple choice] questions useful?

- - - 4.34 ± 0.13 4.41 ± 0.12 4.18 ± 0.14 4.16 ± 0.13
- - - 4.50 ± 0.10 4.43 ± 0.12 4.02 ± 0.16 4.20 ± 0.12

* In the first force-velocity-acceleration task, were the follow-up multiple choice questions useful (if applicable)?

- - - - - 4.00 ± 0.11 4.09 ± 0.14
- - - - - 4.00 ± 0.16 3.98 ± 0.15

† In the first force-velocity-acceleration task, did the program perform well in analyzing your responses?

- - - - - 3.82 ± 0.14 3.94 ± 0.15
- - - - - 3.79 ± 0.16 4.04 ± 0.15

* Scale ranged from 1 (No, not at all) to 5 (Yes, very helpful)
† Scale ranged from 1 (No, not at all) to 5 (Yes, very well)
In addition, there seems to be hints of an interesting interaction for the algebra-based students. For the algebra-based students, both the main effect of question format \( F(1, 194) = 3.476, p = 0.064 \) and the interaction between format and follow-up levels were marginally significant \( F(1, 194) = 2.944, p = 0.088 \). A closer look at student ratings of the question usefulness suggests the source of this effect. When the algebra-based students rated the usefulness of the short answer format with single questions, the mean rating fell below the comparable multiple choice condition. However, when follow-up interactions were added to the mix, the mean rating for the short answer format jumped up to the same rating as the multiple choice version. Although not a significant claim at the 0.05 level, this finding is suggestive. It appears that increasing the level of interaction might have increased the relative rating of usefulness of the short answer format, at least for the algebra-based students.

The only significant effect of question format was in terms of student ratings of the usefulness of the feedback. There was a significant main effect of question format for the algebra-based population \( F(1, 194) = 7.779, p = 0.006, \text{ partial } \eta^2 = 0.039 \), with ratings of feedback usefulness being higher for multiple choice rather than short answer. The effect of format on usefulness of feedback was not significant for the calculus-based experiment \( F(1, 225) = 2.492, p = .116, \text{ partial } \eta^2 = 0.011 \). Given that the feedback was identical, this is interesting. One possible explanation for the difference in scores is that students in the algebra-based course are placing extra value on the feedback for multiple choice because it matches the expectations and context represented by the provided answer choices. In a sense, the multiple choice helps the student to know what to expect from the feedback. On the other hand, the student in the short answer conditions has to make sense of the feedback in light of their interpretation of the question and their own assessed response. Finally, we note that in light of the above, students reported a more positive-than-not outlook on the performance of the natural language algorithm in matching their responses (ranging from approximately 3.8 - 4.0 out of 5 depending on the level of follow-up and the course).
4.3.3 Discussion

There are several main takeaways from this combined pair of experiments. First and foremost, although there were significant differences compared to no-training control (in particular for short answer format with follow-up questions, which was nominally the highest performing intervention in both experiments), there were no overall effects of either question format (multiple choice vs. short answer) or level of follow-up (single question vs. follow-up question dialog) on student performance on the multiple choice post-test. Moreover, this finding of no significant difference extended to retention of learning gains; there was no significant effect from either factor on subsequent retention, measured approximately 2 months after the initial training. However, there was evidence - particularly for the calculus-based students - that there were still significant learning gains (as compared to pre-test scores and relative performance of the control condition) despite measurable decays (as compared to immediate post-testing). As such, this suggests that recurrent training at opportune spacings may offset such learning decay and potentially allow for further gains.

The second main finding was that there was a significant effect of question format on subsequent student performance with a short answer assessment. In particular, there were gains on one specific type of question, namely a short answer $F \rightarrow v$ question, where students were asked to describe the motion of an object at an instant, given the net force acting upon it at that time. This difference was driven in part by a significant difference in the proportion of students making the conceptual error that the force and velocity need always be aligned. However, there were no differences on the other two short answer questions by format (a transfer question involving a graph, and a question based on one of the earlier trained questions).

The third finding, though not as surprising, is that single question multiple choice was by far the most time efficient training method. To this, the addition of follow-up questions and a switch to short answer formats both represented a significant increase in training time. In fact, the combination of short answer with follow-up questions represented an average training time double that of the single multiple choice version for both populations.
In addition to addressing these main research questions, these experiments have identified potential factors between the treatments that may have further pedagogical implications. First, there was a significant difference in the amount of time students spent on the feedback depending on the format of the training question. Students were likely to spend more time on the feedback in the short answer conditions compared to multiple choice versions, despite the fact that the feedback was identical. Moreover, this effect was amplified at the start of the training, when the task was novel, and later when the student encountered a question that broke from the standard $X \rightarrow Y$ format. As such this suggests that an increased variety of questions may help the short answer condition to shift farther from multiple choice, as students in the short answer question are less able to rely on previously recalled question structure. Second, there was evidence of a small but statistically significant finding of over-training for both the calculus-based and algebra-based populations. Although there were no differences between the treatments based on either question format or level of follow-up, the fact that algebra-based students were more likely to demonstrate over-training behavior suggests that any further potential gains in one-relationship may be hindered by such competing effects, particularly for lesser-prepared students. Finally, there seem to be hints of an interesting interaction between follow-up questions and students’ view of the usefulness of the short answer format. In particular, there is marginally significant evidence that students viewed the short answer question format as more useful when it was coupled with follow-up questions, as compared to single question settings. Perhaps further supporting this effect, it is important to note that this increase in rating of the usefulness of the training questions corresponded to situations with a significant increase in the time-on-task required of the student in order to complete the training. As such, this hints that there may be attitudinal benefits from the follow-up structure for short answer formats - either due to how students related to the training (i.e. as more conversational) or in how students viewed the follow-up questions themselves.

Given the implications of over-training and the differences in how students interacted with feedback, the next experiment sought to investigate whether an increase in the variety of question types during training had a significant effect on the relative effectiveness of the
two question formats.

4.4 Experiment #5: Effects of Question Variety

4.4.1 Design and Participants

Based on the results of the previous pair of experiments – and in particular, the finding that students spent different amounts of time on feedback based on the question format – we sought to test the hypothesis that an increased variety of question types might lead to appreciable gains for the short answer format. In the previous experiment, student time on feedback in the short answer conditions tended to spike when the question differed from its counterparts, whereas the time on feedback for the multiple choice conditions tended to more closely follow a general, decreasing trend. One possible explanation is that students in the short answer condition had internalized the basic multiple choice structure for a typical $X \rightarrow Y$ question. As such, although they were forced to recall the answer, that recall may not have been significantly more effortful than a multiple choice counterpart. Moreover, in that case, their answer was likely to match the structure of the anticipated feedback, minimizing any cognitive effort for reconciling the feedback with their own response. As such, an increased variety of questions may lead to further cognitive processing for the short answer format - both upfront when they formulate their answers, and after as they consider the provided feedback - compared to multiple choice.

To test this hypothesis we designed two different versions of training questions, crossed with the two versions of format (multiple choice vs. short answer). The standard training consisted of questions in the $X \rightarrow Y$ format. Given the prior (marginal) finding that follow-up question formats increased the relative student rating of the short answer format, we used the follow-up feedback versions for the $X \rightarrow Y$ format questions - each with the same structure for any given force and motion relationship. That is, if there was a $F \rightarrow v$ question involving a football player, and later another involving an abstract context, the feedback pattern and set of follow-up questions would be the same given similar student responses.

On the other hand, the varied versions of the training consisted of blocks of 2-4 questions,
Figure 4.18: Schematic illustration of training structure and question versions. The varied-question training included question variants that specifically deviated from the typical $X \rightarrow Y$ type.
each of a slightly different question structure. As much as possible, these questions were restricted to the relationship of the directions of net force, velocity, and acceleration in one-dimension. However, that space was too limiting to ensure a large enough student answer space, and so the varied versions of the questions also included several references to constant velocity, constant acceleration, and examples of changing magnitude. In addition, the varied question format included a block of questions that asked students to make a distinction between the relationship between two quantities at a single point in time, and over an extended time interval. Sets of $X \rightarrow Y$ questions were included in the varied training, but always separated by other question types. An illustration of the two training conditions, with examples of different question versions are shown in Fig. 4.18.

Finally, efforts were made to keep the content and surface features of the two versions as similar as possible. If a question in the varied-format involved a rocket or an elevator, the corresponding $X \rightarrow Y$ question did as well. In addition, we tried to control for overall exposure to the inter-relationships between net force, velocity, and acceleration. For example, if a question in the varied format asked a student to consider different applications of the word “move” in everyday language and how those corresponded to specific physical scenarios, the matching $X \rightarrow Y$ question was an $V \rightarrow A$ question.

This experiment made another overall change to address a potential competing explanation for the results in the previous experiments, namely that the mean training time spent on the treatments in the previous experiment was limited to a relatively short duration. On average students spent between 10-20 minutes on the force and motion training depending on condition. Given the repeated difficulty demonstrated by students and the pervasiveness of student misconceptions related to the topic, the training dose may simply have been insufficient to observe overall differences between conditions. As such, we made structural changes to the administration of the task to increase the dosage.

Here, students once again completed the research tasks as part of a flexible homework assignment in their introductory algebra-based physics course. Unlike prior experiments, we brought students in twice to the laboratory setting, with an enforced time delay of 2-3 weeks (depending on student schedule and availability) between the two sessions. Students
received full participation credit for completing both sessions. An outline of the experiment administration is shown in Fig. 4.19. During the first flex-session students completed the pre-test and the first part of their randomly assigned training. The first training portion consisted of 14 questions. During the second session, students completed the second half of their previously assigned training (11 questions), followed by a 10-15 minute unrelated task, and then finally the FVA post-test.

As shown in figure 3.1, a total of 288 students were randomly assigned to one of five experimental conditions: no-training control (N = 57), multiple choice with $X \rightarrow Y$ questions, short answer with $X \rightarrow Y$ questions (N = 58), multiple choice with varied questions (N = 61), and short answer with varied questions (N = 55).

### 4.4.2 Results

(a) **Student performance on the multiple choice assessment**

The mean scores on the 17 question FVA post-test and corresponding mean over-training scores for each condition are shown in Table 4.6. The over-training score was calculated in the manner previously used in Experiment 3 and 4 (see section 4.3.2). A one-way ANOVA showed no statistical differences between students pre-test scores across the five experimental conditions, $[F(4, 283) = 0.672, p = 0.612]$.

A one-way ANOVA indicated a significant difference between experimental conditions on total score on the full FVA test $[F(4, 283) = 5.142, p = 0.001]$. Subsequent Dunnett’s post
Table 4.6: Mean multiple choice post-test score and over-training score by condition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>FVA Post-test score</th>
<th>Over-training score</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-training control</td>
<td>6.19 ± 0.43</td>
<td>0.19 ± 0.08</td>
</tr>
<tr>
<td>Multiple choice (X → Y questions)</td>
<td>8.43 ± 0.51</td>
<td>0.50 ± 0.11</td>
</tr>
<tr>
<td>Short answer (X → Y questions)</td>
<td>7.93 ± 0.50</td>
<td>0.30 ± 0.08</td>
</tr>
<tr>
<td>Multiple choice (varied questions)</td>
<td>9.10 ± 0.55</td>
<td>0.54 ± 0.12</td>
</tr>
<tr>
<td>Short answer (varied questions)</td>
<td>8.91 ± 0.56</td>
<td>0.49 ± 0.10</td>
</tr>
</tbody>
</table>

Notes: MC score is the mean score on the 17 question multiple choice FVA test. OT score represents the mean calculated “overtraining score”. Errors shown are ±1 standard errors.

Post hoc comparisons to control indicated a significant difference between control and three of the treatment conditions: multiple choice with X → Y questions (p = 0.008), multiple choice with varied questions (p < 0.001) and short answer with varied questions (p = 0.001). Short answer with X → Y questions was marginally significant (p = 0.060). A 2 × 2 ANOVA on format and question variety found no significant main effects of either factor ([F(1, 227) = 2.401, p = 0.123] and [F(1, 227) = 0.422, p = 0.517] respectively.

In addition to the entire FVA-test score, we considered the gain on the 10 common items represented by the pre-test. The mean gain in this common 10 question-subscore are shown by condition in Fig. 4.20. A one-way ANOVA between conditions was significant [F(4, 283) = 7.434, p < 0.001]. Post hoc comparisons were conducted via a Dunnett’s test to control. These comparisons indicated a significant difference between control and all 4 of the treatment conditions (p < 0.001). There were no significant differences between the different interventions. Given that these items were all of the general form X → Y, it is somewhat surprising that more varied practice – which deviated from this structure wherever possible – was at least as effective at improving performance on these specific sub-items. After all, this particular gain measure represents a question format identical to the type that students were trained on in the X → Y conditions.

In order to monitor incorrect student over-generalization due to the training, we compared student over-training scores across conditions. A one-way ANOVA was marginally significant [(F, 4, 283) = 2.327, p = 0.06]. Only the multiple choice treatment (varied questions) was significantly different from control (p = 0.043). Multiple choice with X → Y
questions was marginally significant ($p = 0.10$). There was no significant difference with respect to control for either short answer with varied questions ($p = 0.12$) or short answer with $X \rightarrow Y$ questions ($p = 0.87$). A comparison of treatment conditions indicated no significant effects from either format or question variety.

Interestingly, despite the increase in dose from previous experiments, the highest over-training score here was only around 0.5, suggesting that less than one question on average per student demonstrated incorrect application of the “no relationship” case. As such, it may be tempting to attribute the strength (or lack) of this effect to the proportion of the training focused on that relationship; however, the training questions here target the $F \leftrightarrow a$ relationships in approximately 20% of the opening question prompts. The previous experiments had a similar proportion of questions (out of 10 training questions one explicitly used a $F \rightarrow a$ question, with two other questions focused on students correct descriptions of acceleration given a net force).
(b) Student time spent on feedback during training

The mean and median time spent by students on the corresponding trainings are shown in Table 4.7. In addition, the table breaks the total time into time spent during both sessions. Although the first session represented an increase in time over the second, it also corresponded to a higher number of questions (14 training questions in the first session vs. 11 in the second). The second session had a smaller number of questions to allow for both a spacer task and the subsequent FVA post-test.

An Independent Samples Median test indicated a significant difference in the means of the different treatment groups $[\chi^2(3) = 60.99, p < 0.001]$. Bonferroni-adjusted pairwise comparisons showed that while the two versions of multiple choice training were not significantly different from one another ($p = 1.00$), they were both significantly lower in time than the two short answer conditions ($ps < 0.001$). The short answer conditions were also not significantly different in time from one another ($p = 0.708$). Therefore, it seems that format (multiple choice vs. short answer) is the largest determining factor in the training time, with an approximately 50% increase in self-determined time-on-task.

Table 4.7: Mean and median time spent (in minutes) on training in total and for mean each session. $\langle Time_{S1}\rangle$ and $\langle Time_{S2}\rangle$ are the mean times spent on the first session training task and the second session task respectively.

<table>
<thead>
<tr>
<th>Condition</th>
<th>$\langle Time_{Total}\rangle$</th>
<th>Median Time</th>
<th>$\langle Time_{S1}\rangle$</th>
<th>$\langle Time_{S2}\rangle$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple choice ($X \rightarrow Y$ questions)</td>
<td>20.2 (6.3)</td>
<td>20.4</td>
<td>12.8 (4.3)</td>
<td>7.4 (2.6)</td>
</tr>
<tr>
<td>Short answer ($X \rightarrow Y$ questions)</td>
<td>30.4 (4.5)</td>
<td>29.1</td>
<td>19.0 (5.5)</td>
<td>11.4 (4.0)</td>
</tr>
<tr>
<td>Multiple choice (varied questions)</td>
<td>22.4 (8.0)</td>
<td>19.8</td>
<td>13.2 (5.2)</td>
<td>9.2 (3.8)</td>
</tr>
<tr>
<td>Short answer (varied questions)</td>
<td>34.4 (10.3)</td>
<td>31.9</td>
<td>21.0 (6.6)</td>
<td>13.4 (8.4)</td>
</tr>
</tbody>
</table>

Note: Standard deviation indicated by (values).

In addition to looking at total time, we tracked the amount of time students specifically spent on the feedback. As before in Experiment 3 and 4, the time reported as time on feedback was based only on the summative feedback provided at the end of a
question/follow-up interaction. This time was recorded based on the difference between when the student submitted their final answer and when they clicked to continue onto the next question. A $2 \times 2$ ANOVA with format and question variety as factors indicated significant main effects of format ($F(1, 198) = 27.27, p < 0.001, \eta^2 = 0.12$) and variety ($F(1, 198) = 7.934, p = 0.005, \eta^2 = 0.039$) as well as a significant interaction effect ($F(1, 198) = 5.949, p = 0.016, \eta^2 = 0.029$).

Figure 4.21: Time spent on feedback for each training question by condition. (Left) First session. (Right) Second session. (Top) Comparison plot of multiple choice with $X \rightarrow Y$ questions and short answer with $X \rightarrow Y$ questions. (Bottom) Comparison plot of multiple choice with varied questions and short answer with varied questions. Error bars are ±1 standard errors.
To better elucidate these effects, traces of the mean feedback time for each individual training question are shown in Fig. 4.21. The plots separate the traces by question variety ($X \rightarrow Y$ vs. varied) and session (first or second) for clarity and to emphasize the transitions in the type of question within the varied format. The dotted lines indicate the as-designed switches in question format (that is, the line represents the last question in a similar set, with the next question representing an intended different style of question).

The figures suggest several main findings. First, students consistently spent more time on the feedback in the short answer condition regardless of the variety of question. Second, students in the $X \rightarrow Y$ trainings demonstrated a relatively clean decay in time spent on feedback. Although such decays are common for training a related skill or set of concepts, it is noteworthy that both short answer and multiple choice $X \rightarrow Y$ trainings tended to follow the same trajectory. On the other hand, there were distinct spikes and a generally slower decay in the time students spent on feedback in the varied question format. The short answer condition with varied questions experienced larger spikes in time on feedback compared to multiple choice with varied questions due to switches in the type of question during training. This is the likely source of the interaction effect identified in the above $2 \times 2$ ANOVA. Moreover, the jumps in time students allocated to the feedback seemed to line up relatively well with the predicted switches in question type designed by the training. In turn, this suggests that the changes to the questions did have the predicted effect of representing a different type of training content. Finally, the time students spend on the feedback at the start of the second session has not reset, suggesting that students retained some level of general familiarity with the task over the 2-3 week delay.

4.4.3 Discussion

There are 2 main takeaways from the results reported here - one, it appears that changing the training to include more question variety did significantly affect how students interacted with the training, particularly in terms of the time spent on feedback. Rather than follow a smooth exponential decay, student time on feedback spiked throughout the short answer (varied question) training, with those changes generally aligning with the instructionally-
designed differences in type of training question. However, despite this fact, the overall gains across the treatment interventions were practically identical. This second finding could suggest two main possibilities.

First, there may be relatively equal and offsetting effects contributing to the effectiveness of the treatments. For training with the $X \rightarrow Y$ style of questions only, the similar decay on feedback for both multiple choice and short answer may hint that the short answer students are in fact internalizing the structure of the anticipated multiple choice questions. In essence, they are practicing a multiple choice question where the choices were read a minute or so prior to the question. On the other hand, the significant time students devote to feedback in the short answer condition, particularly when switching to a novel question set, suggests that students may be spending effort to interpret that feedback in light of their written responses. As such, it is slightly surprising that multiple choice performed equally well with varied questions. One possibility is that students in the multiple choice variant performed better on the question itself given the scaffolding provided by the questions - further work is needed to assess this potential effect.

The second possible explanation is that the training is starting to hit a sort of pseudo-ceiling in terms of further gains. It is interesting that the gains demonstrated pre-post here (20% in the common 10 questions) are similar to the highest gains observed in the earlier experiments (short answer with follow-up feedback), despite changes in the type of questions asked within this experiment. As a result, this may suggest that this kind of availability-based, restricted training can only get students so far in terms of additional gains. Moreover, given the persistence of student difficulties, it may suggest that students’ prior experience with these general force and motion concepts is providing unique resistance against learning gains and subsequently hindering potential observations of differences in training format. As such, in the next section, we describe an experiment investigating comparisons of multiple choice and short answer formats with a different system of physics concepts.

Finally, an additional twist is that despite the increase in training dose, we did not observe a further increase in over-training compared to control. This is particularly inter-
testing given that the algebra-based students were more likely (nearing an over-training rate of 1 problem per student) to commit such errors in the previous experiment. As such, one possible explanation may be that given the strength of the Newton’s Second Law relation and students’ prior instruction and practice, it may be the number of recent repetitions with the concept, rather than the relative proportion, that determines how likely students are to over-generalize the consideration of no-relationship.

4.5 Experiment #6: Short Answer vs. Multiple Choice (Electric Field & Potential)

4.5.1 Design and Participants

Up until this point, the investigations into different question formats have focused on training the set of simple relationships between force, velocity and acceleration in one-dimension. As motivated previously, that conceptual system was chosen based on prior research that identified specific student difficulties with this set of relationships. In particular, students tend to assume that force and velocity need to be aligned at all points in time. In addition, previous studies with a multiple choice test had verified the existence of intermediate states (Rosenblatt and Heckler, 2011) – i.e. the velocity can be aligned or anti-aligned with the force, but it cannot be zero – as well as potential training hierarchies (White and Heckler, 2013). As such, we postulated that short answer and multiple choice training that targeted the availability of those different physical contexts (and clarified those contexts with specific practice with the related language and terms) might result in respectively different gains depending on the format.

Although the previous experiments did demonstrate significant gains versus no-training controls, there was no significant difference in training with short answer or multiple choice formats, at least as measured by metrics like the aforementioned multiple choice test. As such, one remaining concern is that the previous conceptual domain may have been too difficult for any observations of relative effectiveness between the two question formats. Although the training always occured after in-course instruction on the topic, it is possible
student mastery of the related concepts may not have met a threshold necessary for these relatively simple, availability-based training questions to be effective. In essence, there may not have been any observable differences in the question formats because training was not the appropriate intervention; further instruction was necessary first.

Given that consideration, we selected the relationship between electric potential and electric field for a complementary experiment. In one dimension, electric field and electric potential are related in a mathematically similar way to acceleration and velocity (rate of change). As such, many of the same reasoning mistakes that occur in force and motion are possible in this context as well, but with less prior experiences to fortify students’ preconceptions (Allain, 2001). For example, although students have relatively less everyday experience, students still have a tendency to assume that potential and electric field are directly proportional - in essence, an assumption of “more is more.”

To this, we designed a set of stem-equivalent short answer and multiple choice training questions focused on the relationship between electric field and potential in one dimension: \( E_x = -\frac{dV}{dx} \). The training and test questions addressed both the magnitude and direction of the electric field given a known potential, the distinction between electric field as a vector and electric potential as a scalar, as well as the (lack of) relationship between the magnitude of potential and the magnitude of the electric field at only a single point. In addition, several of the training questions included multi-turn follow-up questions, in the spirit of Experiments #3 and #4. For the most part, these follow-up questions asked the student to restate the relationship between the two quantities, or consider an important physical counter-example when they provided an incorrect answer. The assessment consisted of 12 multiple choice questions followed by 3 short answer questions. The training and assessment questions are included in Appendix A.5.

Experiment participants consisted of students from a second semester introductory physics course focused on topics in electricity and magnetism. A total of 118 students were randomly assigned to one of three experimental conditions: training via multiple choice questions (N=39), training via short answer questions (N=39), and a no-training control (N=40). Students completed the tasks in individual carrels in a quiet testing room as part
of an one-hour long flexible homework assignment for their course. Student grades were based solely on participation. At the end of the training session, students were asked to predict their score on the multiple choice post-test, as well as a series of Likert scale attitudinal questions. Due to an administration oversight, students in the no-training control condition did not receive the attitudinal survey.

4.5.2 Results

(a) Student performance on the multiple choice assessment

Students' final course grades in their introductory electricity and magnetism class were used to provide a general measure of student ability. A one-way ANOVA showed no statistically significant differences between students overall course grades across the three conditions, \( F(2, 115) = 0.475, p = 0.623 \).

![Figure 4.22: Student scores on the 12 question multiple choice post-test by condition.](image)

The mean scores on the multiple choice post-test are shown in Fig. 4.22. A one-
way ANOVA indicated a significant difference between the three conditions \( F(2, 115) = 33.005, p < 0.001 \). Follow-up Tukey post hocs indicated a statistically significant difference in scores between the short answer format and control \( (p < 0.001, d = 1.70) \), and between multiple choice format and control \( (p < 0.001, d = 1.40) \), but not between the short answer and multiple choice formats \( (p = 0.61, d = 0.22) \). Although they were not significantly different from one another, both multiple choice and short answer represent a large performance gain compared to control, approximately 1.5 standard deviations in both cases.

(b) Student performance on the short answer assessment

Student performance on the short answer post-test questions was highly correlated with their performance on the multiple choice assessment \( (r(40) = 0.649, p < 0.001 \) for control; \( r(39) = 0.684, p < 0.001 \) for multiple choice; and \( r(39) = 0.563, p < 0.001 \) for short answer).
way ANOVA indicated a significant difference between the three conditions \( [F(2, 115) = 18.431, p < 0.001])\). Follow-up Tukey post hocs indicated a statistically significant difference in scores between the short answer format and control \( (p < 0.001, d = 1.12)\), and between multiple choice format and control \( (p < 0.001, d = 1.19)\), but not between the short answer and multiple choice formats \( (p = 0.808, d = 0.15)\). Once again, both multiple choice and short answer represent a sizable performance gain over control (over a standard deviation for both cases).

Moreover, there were significant gains on each individual short answer question as shown in Fig. 4.24. Question 13, the first short answer question, simply asked students to explain what they knew about the electric field in a region where the potential increases linearly in the positive x-direction. The question was graded on a 3-point scale, with a point for magnitude (field is constant), direction (negative x-direction), and reasoning (description of the equation in their own words). Chi-squared tests showed that there were significant differences between the conditions in the proportion of students providing the correct direction \( (\chi^2(2) = 6.737, p = 0.034)\) and reasoning \( (\chi^2(2) = 19.508, p < 0.001)\), and a marginally significant difference in the proportion of students who identified the field was constant \( (\chi^2(2) = 5.497, p = 0.064)\). Subsequent pairwise comparisons with corresponding Bonferroni adjustments showed no significant differences between the multiple choice and short answer trainings at the 0.05 level.

Question 14 and Question 15 were scored on a 2-point scale (1 point each for the correct answer and justification) and dealt with two other cases, namely: the field is zero in a region (what do you know about the potential?) and the potential is zero at a point (what do you know about the field?). In addition to finding the same trend - students performed significantly better after training, but with no difference between treatments - an analysis of student responses to Q15 indicated clear evidence of incorrect proportional reasoning between potential and field. In particular, approximately 20% of students across the experimental conditions were able to state the correct mathematical relationship between electric field and potential, but then misinterpreted it to conclude that the electric field was zero. The common reason given was captured by the following example student response, “It [the
Figure 4.24: Student performance for each short answer question by condition.

electric field] will also be zero. E is the negative derivative of V, and if V is already 0 there is nothing to take the derivative of, therefore making E zero.”

Unfortunately, training did not significantly reduce the number of students making this mistake versus control ($\chi^2(2) = 1.460, p = 0.482$). However, it is interesting to note the robustness and prevalence of this misconception given its structural similarity to similar reasoning errors in one-dimensional force and motion. In the future, it may be beneficial to scaffold training questions that have students specifically focus on explaining the mathematical relationship itself and how that connects with the physical interpretation.

(c) Attitudinal data

Students in the two training conditions were asked a series of quick attitudinal questions at the completion of their one-hour session, as shown in Table 4.8. The first two questions asked students to predict typical performance on the electric field and potential assessment. For
both conditions, student predictions of their own performance on the post-test were highly correlated with their actual scores, suggesting that students had an accurate representation of their ability ($r(39) = 0.659, p < 0.001$) for multiple choice and $r(39) = 0.507, p = 0.001$ for short answer format).

Students were also asked to rank the usefulness of the training questions and the training feedback using a pair of 5-point Likert scales. Student ratings for both short answer and multiple choice treatments were extremely high, with the lowest mean rating being 4.3/5. In addition, there were no significant differences between student ratings between conditions.

Finally, students had a more positive-than-not outlook overall on the performance of the natural language algorithm in matching their response. Students’ rating of the natural language matching did not correlate with their performance on the task ($r(38) = 0.049, p = 0.769$) or their general ability, as represented by their course grade ($r(38) = 0.069, p = 0.681$).

4.5.3 Discussion

The goal of this experiment was to duplicate the basic structure of the earlier experiments in a new conceptual domain in order to investigate whether the previous findings of no significant difference between multiple choice and short answer formats was in part due to conceptual resistance to training on force and motion concepts. In essence, one possible reason for no difference between the treatments is that there was not a large enough overall effect from training to observe it.

Here, training on the concepts of electric potential and electric field resulted in large gains in subsequent student performance on similar conceptual questions compared to a no-training control (approximately 1.5 standard deviations in the context of the multiple choice assessment, and 1 standard deviation for the short answer assessment questions). There were no significant effects from question format.

Given the success of the training as measured by student performance on the multiple choice assessment (post-test scores were near 80%), it is possible that student performance is also starting to hit a ceiling here. However, performance on the short answer questions was
Table 4.8: Student responses to attitudinal exit survey by condition. Errors shown are ±1 standard errors.

<table>
<thead>
<tr>
<th>Question</th>
<th>Condition</th>
<th>Multiple Choice</th>
<th>Short Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using a standard 100% scale, how well do you think you did on the electric potential and electric field post-test?</td>
<td></td>
<td>78 ± 3</td>
<td>80 ± 2</td>
</tr>
<tr>
<td>Using a standard 100% scale, how well do you think the typical student in this course does?</td>
<td></td>
<td>73 ± 2</td>
<td>76 ± 2</td>
</tr>
<tr>
<td>*In the first potential and electric field task, were the questions useful?</td>
<td></td>
<td>4.43 ± 0.09</td>
<td>4.33 ± 0.12</td>
</tr>
<tr>
<td>*In the first potential and electric field task, was the feedback on the questions useful?</td>
<td></td>
<td>4.60 ± 0.12</td>
<td>4.60 ± 0.08</td>
</tr>
<tr>
<td>*In the first potential and electric field task, were the clarifying multiple choice questions useful (if applicable)?</td>
<td></td>
<td>-</td>
<td>4.71 ± 0.08</td>
</tr>
<tr>
<td>†In the first potential and electric field task, did the program perform well in analyzing your responses?</td>
<td></td>
<td>-</td>
<td>3.85 ± 0.14</td>
</tr>
</tbody>
</table>

*Scale ranged from 1 (No, not at all) to 5 (Yes, very helpful)
†Scale ranged from 1 (No, not at all) to 5 (Yes, very well)

noticeably lower, with the correspondingly lower floor represented by the control condition suggesting that student performance gains were meaningful. Consequently, it seems unlikely that those questions were capped by ceiling effects. There were once again no significant differences between the two question formats.

4.6 Summary

Taken together, this combined study found that training via short answer and multiple choice formats were both effective at improving student performance on subsequent multiple choice assessments – as measured relative to course instruction alone – and similar in overall performance. Experiment 2 found a marginally significant difference ($p = 0.056, d = 0.45$) in
student performance between the short answer format and multiple choice formats. There was no significant main effect from question format in any of the other experiments as measured by subsequent multiple choice assessment. However, there was replicated evidence across two different courses (an algebra-based and a calculus-based introductory mechanics course) that the short answer format did make a difference on later performance with a specific type of short answer question, namely a $F \rightarrow v$ question. In contrast, there were no significant differences on either a graph-based transfer question or a question replicated from the training. As such, this may suggest that short answer training on simple conceptual relationships can be at least as useful as multiple choice overall; it resulted in comparable benefits on multiple choice assessments, and potentially increased performance on certain types of short answer questions.

One important trade-off, however, is that multiple choice was repeatedly more efficient from the simple perspective of gain in pre-post scores as a function of training time. In particular, when short answers were combined with other changes to the training questions (either increased interactive follow-up or increased question variety) there was a tendency for the gap in time-on-task between short answer and multiple choice formats to widen. However, there is some evidence that part of this additional time-on-task has key instructional relevance; students in the short answer conditions spend significantly more time on training feedback (even when it is identical) compared to multiple choice. To this, it is also important to note that the time spent on training in all of the experiments here was self-determined by the student. As such, it appears that short answer formats may be a useful and authentic way to increase the amount of time students spend on training material and in particular, the amount of time they spend on question feedback. In fact, Experiments 3 and 4 found that students rated trainings using follow-up questions - which were significantly more time intensive than their single question counterparts - as more useful.

Experiment 4 hinted at a particularly interesting attitudinal interaction: students were not only statistically more likely to prefer follow-up questions to single questions as rated by overall question usefulness, but there was a marginal interaction with the format of the question. In short, students tended to rate single multiple choice questions as more
useful than their short answer counterparts. When follow-up interactions were added to both, students rated the formats as equally useful. This finding suggests that the increased interaction of the training helped students’ attitudinal perceptions for the short answer format – either by making the training more conversational, or by scaffolding the intent of the main short answer questions. In Experiment 6, where the training was extremely successful (with effect sizes of $d \approx 1.5$ relative to control) students tended to rate both formats with equally high marks (> 4 out of 5).

Experiments 3 and 4 included explicit measurements of student retention approximately 1.5-2 months after initial training. In both cases, there was evidence that the gains from the training as measured between pre-test and retention test were significantly different from control. As such, it appears that not only is the training retained on a time-scale that is potentially valuable (periodic re-training could be used to lengthen retention further), but that it was above any gains from further course instruction or testing effects. On the other hand, there was no significant effect of question format on retention.

Experiments 3 and 4 also found no significant effect of follow-up question on the effectiveness of the two formats, as measured by either multiple choice or short answer assessment questions. Similarly, Experiment 5 found no significant effect of increased question variety on relative performance of the two question formats. However, there was evidence that the increase in question variety altered how students interacted with the training. Students with varied short answer questions are much more likely to spend longer considering the feedback when the structure of the question changes compared to their multiple choice counterparts.

Finally, this collected set of experiments illustrated several important findings related to effective training via short answers for similar conceptual topics. First, Experiments 1 and 2 suggested that there was no significant correlation between student performance after training and the number of multiple choice clarifications invoked for students during training. Second, care needs to be taken to ensure that the feedback is designed in such a way to minimize any potential interference, particularly for false positives. Third, the act of assessment seems to dramatically alter the way students respond to short answer questions: students significantly reduce the length of their answers and tend to focus on key terminology.
and clear lists of potential physical cases when they knew they were being provided feedback on the correctness of their response. This finding may suggest that the act of assessment is another important dimension by which to design and assess instructional interventions, in addition to more traditional comparisons of feedback information and interaction levels.

Overall, these experiments suggest that controlling for content, feedback, and level of interaction, short answer and multiple choice formats are equivalently effective for the purpose of training students on conceptual physics relationships – at least to first order and for a comparable level of training question as studied here.
Volume II

What Works with Worked Examples: Extending Analogical Comparison and Self-Explanation to Synthesis Problems
Chapter 5
THEORETICAL BACKGROUND

5.1 Synthesis Problems

Problem solving is a complex and multi-faceted process. Accordingly, there has been a significant investment in problem solving research in physics exploring problem solving frameworks, novice vs. expert problem solving strategies, and related procedural skills (for reviews, see (Docktor and Mestre, 2014; Hsu et al., 2004; Maloney, 2011)). However, the vast majority of these studies have typically focused on problems requiring the application of one single, isolated physics concept (e.g., Larkin (1979); Reif and Heller (1982); Van Heuvelen (1999); Hardiman et al. (1989); Kohl and Finkelstein (2008); Meltzer (2005)).

The following series of experiments seeks to investigate a specific subclass of physics problem, which we will refer to as a synthesis problem: namely, a question requiring the application of more than one major physics concept, often from disparate parts of the teaching timeline (Ding et al., 2011). Synthesis problems are of importance for both theoretical and practical reasons. Practically, synthesis problems are often closer to real world situations in their complexity. As a result, improving student success on synthesis problems is consistent with the goal of better preparing future engineers and scientists. In the context of physics education research, synthesis problems are similar to context-rich problems in this pursuit – a topic of ongoing research in both general problem solving (Heller et al., 1992; Heller, 1992; Ogilvie, 2009) and computer-aided tutoring (Antonenko et al., 2011; Ryan et al., 2016).

Synthesis problems are also of theoretical interest as they provide unique difficulties for students in the recognition and joint application of multiple concepts (Ding et al., 2010, 106).
2009; White et al., 2014). Our previous studies showed that these distinct challenges extend beyond just the sum of difficulties represented by the individual component concepts. In particular, the recognition of multiple concepts becomes a significant bottleneck in the context of these more complicated problems (White et al., 2014). This difficulty with synthesis problems is likely exacerbated by end-of-chapter textbook exercises and homework activities focusing on practicing only the most recently learned material in the context of single-concept problems. Students often approach these end-of-chapter exercises with documented plug-and-chug algorithms that do not necessarily scale successfully to situations with multiple interconnected physics concepts (Allen et al., 1996; Duch, 2001; Sabella and Redish, 2007). As such, the experiments here represent a novel focus on extending instructional methods based on worked examples specifically to synthesis problems and the unique challenges therein, namely multiple concept recognition.

5.2 Worked Examples

Worked examples consist of a problem statement and a corresponding set of solution steps, often with the implicit goal of modeling an expert-like approach to the solution of the problem. Previous research has shown that worked examples can be extremely effective in aiding novice learners as they attempt to master domain-specific knowledge and problem solving skills, especially in highly structured domains such as physics (Sweller and Cooper, 1985; Zhu and Simon, 1987; Atkinson et al., 2000). Moreover, seminal work by Sweller et. al. demonstrated that with careful, principle-based instructional design, studying worked examples can be significantly more effective than individual practice solving problems (Sweller and Cooper, 1985). This “worked example effect” has traditionally been framed in terms of cognitive load theory – worked examples are effective because they reduce extraneous load associated with inefficient problem solving strategies (Sweller, 1988; Renkl and Atkinson, 2007). Rather than devote limited cognitive capacity to plug-and-chug algorithms and equation matching heuristics, a fully worked example allows the novice to instead focus on extracting the relevant solution structure and construct a conceptual schema for subsequent
5.3 Analogical Comparison

Analogical reasoning is a mechanism of applying what has been previously learned from a base situation to a new, analogous target situation. Successful analogical reasoning requires that a person recognize base-target similarity, perform structural mapping, and subsequently apply the base solution to the target (Catrambone and Holyoak, 1989; Gentner, 1983, 1989; Gentner and Markman, 1997; Gentner et al., 2003; Ferguson and Forbus, 1998). In physics, researchers have used analogical reasoning to facilitate student conceptual learning (Brown and Clement, 1989; Clement, 1993; Ding et al., 2011; Podolefsky and Finkelstein, 2006). Although the methods and implementation have differed, the primary goal has often been to help novices acquire understanding of a novel concept via analogies to a situation that the student already comprehends (such as invoking the idea of water flow to understand current in a circuit, or scaffolding a series of analogies to aid conceptual understanding of normal force).

Here, we focus on a specific type of analogical reasoning known as analogical comparison. Analogical comparison invokes student comparison between two worked examples with the intent that students extract the necessary structure to tackle a related target problem. The technique was explored in a study by Gentner, Loewenstein and Thompson that tested the use of analogical comparison with undergraduates and business-negotiation techniques (Gentner et al., 2003). In their studies, participants were asked to explicitly compare and contrast two isomorphic base examples before solving a related target problem. It was found that this analogical comparison between base examples facilitated learners to recognize, map, and apply key principles significantly better than did the traditional technique of using only one single base example. Given previously documented student difficulties recognizing component concepts when solving synthesis problems (White et al., 2014; Ding et al., 2011), we posit that this technique of analogical comparison may be particularly suited to helping students solve physics synthesis problems. Since analogical comparison emphasizes the
identification of conceptual structure, it may assist students to overcome the characteristic multiple concept recognition and joint application bottlenecks that were identified in our prior studies on synthesis problems (Ding et al., 2010, 2011; White et al., 2014).

This proposal is further supported by previous studies in physics that have tested the effectiveness of isomorphic worked examples and analogical reasoning as a method to improve student problem solving. In particular, Lin and Singh have previously shown that invoking student discussion and comparison of a single isomorphic worked example to a target multi-concept problem can improve student use of the relevant physics concepts (Lin and Singh, 2011). Interestingly, they found that students who were first asked to think about how to solve the target problem before comparing it to the provided worked example performed significantly better on that target problem compared to participants who were provided the worked example, scaffolding prompts and explicitly told that the target problem and provided worked example shared the same physical concepts (energy conservation and centripetal acceleration).

In comparison to the work of Lin and Singh, where a single isomorphic worked example was provided to the students for study and use on the target problems, the series of studies conducted here specifically employ analogical comparison across pairs of worked examples. By providing pairs of worked examples with similar solution structure, we test the hypothesis that analogical comparison can assist students to extract the overall solution structure of a target synthesis problem while minimizing the impact of surface features from the provided worked examples. In short, analogical comparison – through an appeal to similarities and differences across the worked examples – may serve as an effective way to help students create a generalizable, and thus readily transferable, solution schema.

5.4 The Return of Self-Explanation

Effective interventions based on worked examples are often coupled with prompted or spontaneous self-explanations, whereby novices seek to explain the rationale and structure of the worked examples either to themselves or an interested third-party. The self-explanation
effect is described more extensively in Section 2.2. However, it is worth pointing out that the importance of self-explanation was identified in a study by Chi et. al, which asked college students to voluntarily self-explain to an experimenter as they studied examples of introductory mechanics problems (Chi et al., 1989). Students that generated more high-quality self-explanations performed significantly better on follow-up problem solving tasks than their peers. That result was then confirmed and expanded upon in multiple studies in physics and other domain areas, such as biology, algebra, and computer programming (Chi et al., 1994; Chi and VanLehn, 1991; Aleven and Koedinger, 2002; VanLehn et al., 1992).

However, as with many of the aforementioned studies on worked examples, the problems and applications used in this previous work have predominately focused on mastering isolated concepts and their application to single concept problems, such as Newton’s second law in the context of an equilibrium problem (in the case of the original Chi experiments). As such, our goal is two-fold: first, to extend the application of self-explanation specifically to the domain of synthesis problems in physics; and second, to compare the effectiveness of self-explanation within individual worked examples to analogical comparison across a pair of worked examples.
In light of previous results, we sought to explore how students utilize worked examples specifically in the context of synthesis problems. In particular, we sought to test whether or not short answer prompts invoking analogical comparison across examples facilitated student recognition of relevant concepts and improved their performance when solving a novel synthesis problem. Along with this overall research goal, we considered the following related research questions. First, given the increased complexity of synthesis problems, is it more effective to invoke comparisons between worked examples that break down the target synthesis problem into its single-concept parts, or worked examples that include the concepts in combination? Second, how does the focus of the prompts influence analogical comparison (i.e. prompts involving holistic structure and overall concept recognition vs. prompts for fine-grain applications of the individual concepts)? Third, how does analogical comparison across a pair of worked examples compare with self-explanation of each worked example independently? These questions have been addressed by a series of four experiments, illustrated schematically in Fig. 6.1.
Figure 6.1: Schematic illustration of the four experimental designs included in the synthesis study.
6.1 Experiment #1: Analogical Comparison - Effects of Type of Worked Examples

6.1.1 Design and Participants

The goal of this first experiment was to compare several methods of analogical comparison to baseline performance from course instruction alone (control) and to recent practice solving single-concept problems (priming). In order to test the effectiveness of analogical comparison in training students to solve synthesis problems, we designed a target synthesis task that would require application of two physics concepts: energy conservation and circular motion. The target synthesis problem used for this study is shown in Fig. 6.2B. In addition to being relevant to the students’ course – it represents a canonical situation presented in various problems within introductory physics courses – the problem was chosen based on previous work which documented significant student difficulties with a similar problem (White et al., 2014).

Three different interventions using variations of methods for analogical comparison were designed. Examples of problems used for the worked example are included in Fig. 6.2A. Full versions of all worked examples and corresponding comparison prompts are included in Appendix B.1.

The interventions varied in both the type of worked examples (single-concept vs. synthesis problems) and the type of prompts invoking comparison (single-concept mastery vs. concept recognition). The three analogical comparison interventions were (1) single-concept examples with mastery prompts (subsequently referred to as Single-concept – Mastery), (2) synthesis examples with mastery prompts, and (3) synthesis examples with recognition prompts.

A no-training control was included to establish a baseline of student performance solely from course instruction. The final condition was a priming intervention. Rather than having students explicitly compare the worked examples, the priming condition asked students to solve two of the single-concept problems – one of each concept – used as worked examples in the analogical comparison conditions. The priming intervention was included to provide a
Figure 6.2: An example single-concept problem and synthesis problem provided as worked examples during the analogical comparison interventions (A) and the target synthesis problem (B).

comparison for effects from recent single-concept practice with the relevant physics concepts, namely increased concept availability. The full experiment design is shown in Fig. 6.3.

The combination of Single-concept - Mastery and Synthesis - Mastery conditions were designed to measure the effect of the type of worked example utilized for analogical comparison. These conditions used the same prompts for comparison with only minor changes to account for different line numbers in the solutions. In addition, the physical contexts of the solutions, diagrams, and problem statements were kept as similar as possible between the synthesis and single-concept worked examples. In order to keep time-on-task as similar as possible across interventions, the analogical comparison conditions with synthesis problems included a comparison of only a single pair of worked examples. Students in the single-concept analogical comparison condition compared two pairs of worked examples: a pair of
worked examples involving circular motion, and a pair of energy conservation examples.

In principle, there are compelling reasons to expect both methods to be successful. On the one hand, the single-concept problems are the embodiment of a reductionist approach: break the overall problem solution structure into its component parts and minimize cognitive load at each individual stage of analogical comparison. As a result, the reduced complexity combined with comparing examples immediately after one another may assist students to recognize how to apply both individually to a following novel problem. On the other hand, the synthesis worked examples are structurally more similar to the target synthesis problem, and include the structural step of joining the two concepts.

The effect of different comparison prompts was assessed by the combination of Synthesis – Mastery and Synthesis – Recognition conditions. These two sets of prompts were designed to focus student attention on different elements of the worked examples. The prompts involving single-concept mastery explicitly targeted the application of the individual physics concepts within the worked examples (e.g. “Is the direction of the $a_c$ term the same or different in the two solutions?” and “Why is the term $mgh$ on the left side of Solution 2, but on the right side in Solution 1?”). The recognition prompts focused on concept recognition and the combination of the concepts within the worked examples (e.g., “What
are the main physical concept(s) used in both of the students’ solutions?” and “identify the elements of the problem statement or diagram which indicate the need to use each of the physical concept(... Are the elements you identified similar between the two problems?).

Tasks were administered during the Fall of 2015. Participants were students in a calculus-based introductory mechanics course at The Ohio State University who participated as part of a 1-hour long flexible homework assignment for course credit. Students earned full credit for the assignment based on participation. A total of 196 students were randomly assigned into one of the five study conditions.

Students completed the training tasks and target synthesis problem in individual carrels in a quiet room. An equation sheet similar to those used in the course was provided to all students. Tasks were administered and collected by the proctor one at a time, and students were allowed to work at their own pace. Students first completed their selected training, followed by 10-15 minutes of unrelated physics tasks, and then the target synthesis problem. Students in the control condition completed a set of unrelated physics tasks and the target synthesis problem, as shown in Fig. 6.3.

Table 6.1: Scoring rubric for the target synthesis problem. 1 point was awarded for each item, for a total of 9 points.

<table>
<thead>
<tr>
<th>Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1 Energy Conservation</td>
</tr>
<tr>
<td>+1 Centripetal Acceleration</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1 $a_c$ applied correctly with Newton’s Second Law</td>
</tr>
<tr>
<td>+1 Identify normal force = 0 (minimum criteria)</td>
</tr>
<tr>
<td>+1 Correct initial potential energy</td>
</tr>
<tr>
<td>+1 Correct final potential energy</td>
</tr>
<tr>
<td>+1 Included final kinetic energy (velocity top of loop $\neq 0$)</td>
</tr>
<tr>
<td>+1 Substitute correct final velocity from centripetal motion constraint</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1 Correct final answer \ No mathematical mistakes</td>
</tr>
</tbody>
</table>
A rubric for assessing student solutions of the target synthesis problem was determined by two of the authors. In an effort to provide an authentic measure of student performance, the rubric was designed to mirror a grading scheme that could be applied in the students introductory course. The rubric is shown in Table 6.1. After discussing the rubric, the two authors coded all of the target student responses independently with an inter-coder agreement of 80%. Concept recognition was coded generously (for example, a student earned credit for recognizing energy conservation if they tried to apply a $\frac{1}{2}mv^2$ term), but required the student to commit to using the concept as part of their solution. If a concept-relevant equation was jotted down separately but not used as part of the final solution by the student, the student was not given credit for recognizing that concept. Assessment of concept recognition was in complete agreement between the two coders. The vast majority of scoring differences occurred with low scoring solutions and typically represented a 1-point shift in the rubric (74% of scoring differences). All disagreements were discussed leading to the agreed upon scores presented here.

6.1.2 Results

Student final course grades in their introductory mechanics class were collected and compared across experimental conditions. To eliminate outliers, two cuts were uniformly conducted across all conditions: students must have completed the course (removing 1 student), and have scored no lower than 2-standard deviations below the course mean (removing a total of 5 students, ranging from 0-3 students per condition). Given that the synthesis problem represents the combination of single-concepts covered as part of the students’ introductory mechanics course, these cuts were conducted to minimize uninformative student difficulties in synthesizing those concepts that may have been due to simple unfamiliarity with the related physics course material. A one-way ANOVA of course grade showed no significant differences across conditions \[F(4, 185) = 1.228, p = 0.301\].

The mean score on the target synthesis problem (out of a maximum of 9) and the number of students per condition are shown in Table 6.2, with corresponding score distributions included in Fig. 6.4. The distributions are distinctly non-normal and roughly clustered into
Table 6.2: Mean score on target synthesis problem out of a maximum score of 9 points. Errors shown are standard errors. Analogical comparison conditions are labeled AC.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean Score ± SE</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>4.15 ± 0.37</td>
<td>40</td>
</tr>
<tr>
<td>Priming</td>
<td>5.12 ± 0.39</td>
<td>43</td>
</tr>
<tr>
<td>AC: Single-concept Examples with Mastery Prompts</td>
<td>5.69 ± 0.40</td>
<td>32</td>
</tr>
<tr>
<td>AC: Synthesis Examples with Mastery Prompts</td>
<td>7.23 ± 0.36</td>
<td>35</td>
</tr>
<tr>
<td>AC: Synthesis Examples with Recognition Prompts</td>
<td>6.83 ± 0.40</td>
<td>40</td>
</tr>
</tbody>
</table>

two distinct groups: one group centered near a score of 3-4 and the other at a total score of 8-9.

A Kruskal-Wallis H test was conducted to determine the effectiveness of the 4 interventions versus control (course instruction only). The mean rank of scores on the target synthesis problem was significantly different between conditions \([\chi^2(4) = 36.2, p < 0.001]\). Pairwise comparisons of each intervention to control were conducted using Dunn’s (1964) procedure with a Bonferroni correction for multiple comparisons \((n = 4)\). The adjusted p-values are presented. Results indicated that while the priming condition was not significantly different from control \((z = 1.97, p = 0.196)\), all three analogical comparison conditions were significantly higher than control: Single-concept – Mastery \((z = 2.632, p = 0.034)\), Synthesis – Mastery \((z = 5.436, p < 0.001)\) and Synthesis – Recognition \((z = 4.379, p < 0.001)\).

To test for hypothesized differences (H-1) between analogical comparison and priming, (H-2) between single-concept and synthesis worked examples, and (H-3) between mastery and recognition prompts, we conducted a Kruskal-Wallis H test across only the intervention conditions. There was a significant difference in mean rank of scores on the target synthesis problem between the interventions \([\chi^2(3) = 16.72, p = 0.001]\). Five pairwise comparisons were conducted using Dunn’s (1964) procedure with a Bonferroni correction for multiple comparisons. The adjusted p-values are presented. To test H-1, comparisons showed there were significant differences between priming and both the Synthesis – Mastery \((z = 3.718, p < 0.001)\) and Synthesis – Recognition conditions \((z = 2.602, p = 0.045)\), but not with Single-concept – Mastery \((z = 0.721, p = 1.000)\). To test H-2, pairwise comparison
showed there were significant differences between Single-concept – Mastery and Synthesis – Mastery ($z = 2.772, p = 0.03$). To test H-3, there were no significant differences between Synthesis – Mastery and Synthesis – Recognition ($z = 1.187, p = 1.000$). Taken together, these results demonstrate that the effectiveness of analogical comparison extends beyond just single-concept practice and concept activation (as evidenced by comparison to priming via single-concept problem-solving exercises). Moreover, while varying the type of prompts had no significant effect on student performance with the target synthesis problem ($d = 0.17$), students who compared synthesis worked examples performed significantly better on the target problem than those who compared examples highlighting the component concepts in isolation ($d = 0.70$).

In addition to considering total scores on the target synthesis problem, we analyzed the proportion of students recognizing each of the two component concepts. The proportion of students recognizing a concept is shown in Fig. 6.5. Almost all students ($\geq 95\%$) recognized
and utilized energy conservation as part of their solution.

The proportion of students recognizing centripetal acceleration on the target synthesis problem varied considerably. A chi-squared test across treatment conditions showed there was a significant difference in proportion of students recognizing centripetal acceleration when solving the target synthesis problem \( \chi^2(3) = 29.899, p < 0.001 \). To test for hypothesized differences between the interventions, post-hoc comparisons between treatments were conducted using pairwise chi-squared tests with a Bonferroni correction for multiple comparisons \( (n = 5) \). The proportion of students recognizing centripetal acceleration
in both the Synthesis – Mastery and Synthesis – Recognition conditions was significantly different from priming, $\chi^2(1) = 18.059, p = 0.001$ and $\chi^2(1) = 19.267, p < 0.001$ respectively. The proportion of students in the Single-concept – Mastery condition was not significantly different than priming $\chi^2(1) = 1.099, p = 1.000$. Comparison between the Single-concept – Mastery and Synthesis – Mastery conditions also showed a significant difference $\chi^2(1) = 9.600, p = 0.01$. There was no difference between the Synthesis – Mastery and Synthesis Recognition conditions. Taken together these results support the trend suggested by the analysis of students’ total score on the target synthesis problem: analogical comparison significantly increased recognition of centripetal acceleration, but only when students compared synthesis examples.

6.1.3 Discussion

There are four important findings from this experiment. First and foremost, training via analogical comparison of worked examples was effective in improving student scores on a target synthesis problem relative to control. Second, training via analogical comparison behaved markedly better than priming (via problem-solving exercises). Third, the effectiveness of analogical comparison depended significantly on the type of worked examples to be compared, but not the specific nature of the comparison prompts.

The fourth finding is that student success on the problem was bottlenecked primarily by their ability to recognize the presence of the centripetal motion constraint. Given the analysis of student concept recognition across control and the four treatment conditions, the non-normal distributions of student total scores are telling: without intervention students primarily solved the target synthesis problem as if it were a single-concept problem. Once they were able to conceptually recognize both the need for conservation of energy and centripetal acceleration, they almost universally shifted from only applying one concept to successfully applying both. As a result, one of the strongest potential gains from training via analogical comparison – at least, specifically in the context of synthesis problems – may be the improvement of student conceptual recognition.

Moreover, the finding that comparisons of synthesis worked examples was significantly
more effective than comparisons of single-concept examples (using the same prompts) supports the importance of this full structural transfer (Gentner et al., 2003). Combined with the fact that priming students via explicit problem-solving practice was not significantly better than course instruction alone, these results are also prescriptive. Namely, these results weaken the often-held assumption that students can repeatedly practice physics concepts in isolation, and simultaneously expect success on problems that combine them. Instead, these results suggest that integration does not happen spontaneously, at least not for solving complex physics problems. As such, in contrast to the vast majority of introductory homework problems and end-of-chapter exercises, success with synthesis problems may best be facilitated by explicit practice with synthesis problems.

We consider two potential explanations for the finding of no significant difference between the comparison prompts focused on individual concept application and those focused on overall concept recognition and structure. First, noting that both conditions were quite successful relative to additional practice solving single-concept problems, it is possible that ceiling effects are limiting the possibility for any difference in overall effectiveness. Second, students tend to switch from applying a one-concept solution on the target problem in the control and priming conditions, to providing fully correct solutions after completing the analogical comparison tasks. As such, it is possible that the greatest benefit of the analogical comparison prompts is that they force the student to sufficiently encode the two synthesis examples; because the students are so successful with the recognition of the individual concepts once they have encoded the combined concept structure, the specific comparisons themselves are less important.

Overall, the results of this experiment suggest that analogical comparison can be an effective technique for training students to solve synthesis problems – at least for problems that demonstrate the level of conceptual and mathematical complexity represented by this target synthesis problem. However, even considering the vast array of potential physics concepts and combinations, the success here is promising. After all, this target synthesis problem and the corresponding base worked examples are already more involved than other previous, successful examples of analogical comparison (Gentner et al., 2003;
Ferguson and Forbus, 1998) Still, the results here suggest natural follow-up questions: Is analogical comparison as effective for a problem where recognition is not a major bottleneck? Is analogical comparison effective for a synthesis problem with increased complexity? Is analogical comparison as effective as another known problem solving intervention method, namely self-explanation? These questions will be considered in the following studies.

6.2 Experiment #2: Analogical Comparison when Recognition isn’t a Bottleneck

6.2.1 Design and Participants

This experiment was designed to replicate the findings of the previous experiment in a new physics context, namely the findings that 1) analogical comparison increased student performance on a target synthesis problem relative to course instruction 2) analogical comparison increased performance relative to practice with single-concept problems (priming) and 3) analogical comparison was effective only when comparisons were invoked between synthesis worked examples, and not between single-concept examples. This experiment also included additional conditions to test the hypothesis that the effectiveness of the analogical comparison intervention is modulated by the structural similarity of the base worked examples to the target synthesis problem. To test these hypotheses, we designed a synthesis task that would require application of two physics concepts: Ohm’s law (in the context of simple circuit analysis) and magnetic fields due to a current carrying wire.

The target synthesis problem is shown in Fig. 6.6C. Beyond testing a different set of physical concepts, there is a nuanced and important difference between this problem and the previous target synthesis problem of Experiment #1. A full solution of this problem requires sequential application of both component physics concepts: Ohm’s Law (and simple circuit reduction) to solve for the currents in the circuit, followed by subsequent calculation of the magnetic field from the component currents. In Experiment #1, a student could arrive at an answer without consideration of the constraint from centripetal acceleration. The solution of this second target synthesis problem is blocked in a way that the first was
not: application of the second physics concept requires the physical quantity solved for by
the first. Consequently, this problem represents a different kind of synthesis problem, and
as such, may result in different bottlenecks in student performance.

Four analogical reasoning interventions and a priming treatment were designed. As in
Experiment #1, the analogical reasoning interventions consisted of pairs of worked examples
(an example of individual questions are shown in Fig. 6.6A and Fig. 6.6B) and several short
answer questions that sought to focus student attention on key similarities and differences
between them (i.e., a question would highlight lines in the compared solutions that focused
on the differences in circuit structure, or differences in the order that the concepts were
applied). The final question for each analogical comparison was intended as a summative
prompt; the question asked students to create a short guide explaining how to solve similar
problems to a friend (a.k.a. create-a-guide question). Full versions of the worked examples
and examples of the comparison prompts are included in Appendix B.2.

The four analogical reasoning treatments were created in a 2x2 design, crossing similarity
of worked examples to the target synthesis task (near vs. far) with the type of problems
used as worked examples (single-concept vs. synthesis). Similarity to the target synthesis
problem was varied by changing what variables were provided and requested. The two
near analogical reasoning conditions matched the structure of the target synthesis question: features of the battery or the currents were provided and the total magnetic field was requested. The provided information was switched for the two far conditions (given a total \( B \)-field, a voltage or current was requested).

As before in Experiment #1, in order to keep time-on-task as similar as possible across conditions, the analogical reasoning conditions based on synthesis problems included a comparison of only one pair of worked examples. Students in the single-concept conditions compared two pairs of worked examples: a pair of circuit problems and a pair of magnetic field problems. The priming treatment consisted of two single-concept questions selected from the worked examples in the single-concept analogical reasoning conditions – one involving circuits and one involving magnetic fields. Students solved both independently. The final condition included in the design was a no-training control. The full experiment design is shown in Fig. 6.7.

![Figure 6.7: Illustration of the design and administration of Experiment #2.](image)

A total of 278 participants were randomly assigned to one of the six conditions, as
shown in Table 1. The participants were students enrolled in the second semester of a calculus-based introductory physics sequence at The Ohio State University. The tasks were administered to students in individual carrels in a quiet testing room, as part of a flexible homework assignment for course credit. An equation sheet used for exams in the introductory electricity and magnetism course was provided.

Students first completed a conceptual pre-test in order to control for student mastery of the individual physics concepts, consisting of 3 magnetic field questions, 3 circuit questions, and 2 non-relevant questions involving electric field and potential. Afterwards, students completed a short unrelated physics task, then their selected training, followed immediately by the target synthesis problem. Students completed all tasks at their own pace. The analogical reasoning and priming tasks required most students 15-20 minutes to complete. The target synthesis problem required approximately 10 minutes.

<table>
<thead>
<tr>
<th>Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1 Ohm’s Law</td>
</tr>
<tr>
<td>+1 Magnetic field from top wire</td>
</tr>
<tr>
<td>+1 Magnetic field from middle half-loop</td>
</tr>
<tr>
<td>+1 Magnetic field from bottom wire</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1 Correct circuit reduction \ resistor combination</td>
</tr>
<tr>
<td>+3 Correct top current</td>
</tr>
<tr>
<td>Correct middle current</td>
</tr>
<tr>
<td>Correct bottom current</td>
</tr>
<tr>
<td>+3 Correct top magnetic field (using top current)</td>
</tr>
<tr>
<td>Correct middle magnetic field (using middle current)</td>
</tr>
<tr>
<td>Correct bottom magnetic field (using bottom current)</td>
</tr>
<tr>
<td>+1 Final magnetic field direction</td>
</tr>
<tr>
<td>+1 No irrelevant contributions (e.g. side wires)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>+2 Correct final answer \ No mathematical mistakes</td>
</tr>
</tbody>
</table>
After review of a subset of student solutions, a 15 point rubric for the target synthesis problem was agreed upon by two researchers. The full rubric is shown in Table 6.3. All solutions were independently graded and the scoring rubric for each student was subsequently compared. Initial agreement of scores was 68%, with the vast majority of disagreements consisting of a one point difference in score. Each disagreement was discussed and resolved, leading to the agreed upon scores presented in Table 6.4.

Table 6.4: Mean score on target synthesis problem out of a maximum score of 15 points. Errors shown are standard errors. AR (Combined) represents the mean of the four analogical reasoning conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean Score ± SE</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>9.6 ± 0.6</td>
<td>44</td>
</tr>
<tr>
<td>Priming</td>
<td>10.4 ± 0.6</td>
<td>47</td>
</tr>
<tr>
<td>AR (Combined)</td>
<td>10.7 ± 0.2</td>
<td>187</td>
</tr>
<tr>
<td>AR Single-concept Near</td>
<td>11.0 ± 0.5</td>
<td>44</td>
</tr>
<tr>
<td>AR Single-concept Far</td>
<td>10.8 ± 0.5</td>
<td>46</td>
</tr>
<tr>
<td>AR Synthesis Near</td>
<td>10.6 ± 0.5</td>
<td>48</td>
</tr>
<tr>
<td>AR Synthesis Far</td>
<td>10.5 ± 0.5</td>
<td>49</td>
</tr>
</tbody>
</table>

6.2.2 Results

(a) Student performance on the target synthesis problem

Student pre-test scores and course grades were compared between conditions. There were no significant differences between pre-test scores \([F(5, 272) = 0.097, p = 0.993]\) or course grade \([F(5, 272) = 0.433, p = 0.826]\) between the experiment groups.

In order to test the hypothesis that training with analogical reasoning is effective for synthesis problems (H-1), an AvaNCOVA was conducted between the combined analogical reasoning treatments (AR) and control, using pre-test score and final course grade as covariates in order to account for differences in student ability. Together, the analogical reasoning treatments represent a significant effect on student performance with
the target synthesis question after controlling for pre-test score and final course grade, \[ F(1, 227) = 4.593, p = 0.033, d = 0.31 \].

To test the effectiveness of AR training vs. priming (H-2), an ANCOVA was conducted between the combined AR treatments and the priming treatment. There was no significant difference between AR and priming after controlling for pre-test score and final course grade, \[ F(1, 230) = 0.652, p = 0.420 \].

Finally, a 2x2 ANCOVA with final course grade and pre-test score as covariates was conducted between the four AR conditions to test whether the similarity to the target synthesis problem (near vs. far) or the question type (single vs. synthesis) used during training influenced treatment effectiveness (H-3). Controlling for pre-test score and final course grade, there was no significant main effect of either similarity \[ F(1, 181) = 0.336, p = 0.563 \] or question type \[ F(1, 181) = 1.645, p = 0.201 \]. Overall, the analogical reasoning interventions had a significant positive effect on student performance with the target synthesis problem, but there was no difference between the combined AR conditions and priming, or between the four different AR conditions.

**(b) Solution Bottlenecks & Common Errors**

An analysis of solution bottlenecks and difficulties with the target synthesis task suggest several potential reasons for the similar effectiveness of AR and priming. First and foremost, most students were successfully able to identify the correct physics concepts required by the target synthesis problem. This is a far different situation from the previous study where concept recognition was perhaps the definitive driving factor of student performance with the corresponding synthesis problem-solving task. As such, it appears that one major potential advantage of analogical reasoning, namely facilitation of students recognizing pertinent concepts, is not manifested in this case as many students are already capable of doing so without intervention.

Although concept recognition was not a significant bottleneck, Table 6.5 suggests several common failure points that did hinder many students from correctly solving the problem. First, a significant portion of students failed to include all 3 relevant B-fields (approximately
30% across conditions). Only slightly over a third of the students included all relevant fields, matched them to the correct physical current, and correctly solved for the component currents. As shown in Table 6.5, neither analogical comparison nor priming improved student progress through these key problem-solving steps.

Table 6.5: Proportion of students successfully accomplishing key synthesis problem-solving steps. Errors shown are ±1 standard errors of a proportion.

<table>
<thead>
<tr>
<th></th>
<th>Control N=44</th>
<th>Priming N=47</th>
<th>AR N=187</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognize both correct concepts</td>
<td>89% ± 5%</td>
<td>94% ± 4%</td>
<td>97% ± 1%</td>
</tr>
<tr>
<td>+ Include all 3 relevant B-fields</td>
<td>61% ± 7%</td>
<td>66% ± 7%</td>
<td>67% ± 3%</td>
</tr>
<tr>
<td>+ Match each field to correct current</td>
<td>43% ± 7%</td>
<td>45% ± 7%</td>
<td>51% ± 4%</td>
</tr>
<tr>
<td>+ Correctly calculate all currents</td>
<td>34% ± 7%</td>
<td>40% ± 7%</td>
<td>41% ± 4%</td>
</tr>
</tbody>
</table>

The mismatch of individual currents to their corresponding magnetic field is a particularly interesting “synthesis” failure point. Approximately 10% of the students across all conditions made the serious conceptual error of finding the total current through the battery, and then using that current for calculating all of their identified B-fields. In this case, the main synthesis failure is students inability to identify that unique-and-initially-unknown currents within the circuit each contribute to the total magnetic field. As such, this failure point provides some evidence that the failure modes of synthesis problems can extend beyond a product of difficulties found in simpler single-concept problems. In this case, the complexity of the physical situation itself may represent an obstacle for students attempting to combine the two physics concepts. In other words, it is not conceptual recognition or individual concept mastery that is at issue – after all, these students often correctly performed the overall circuit reduction to find the total current through the battery and also used correct forms for the magnetic field. Instead, this is an example of joint application of physics concepts; and this failure in joint application seems to be intimately connected with the particular physical context of this problem, namely that it is not just the total current that ultimately determines the strength of the component magnetic fields, but rather the
distribution of that current through the different circuit branches.

As shown in Table 6.6, AR and priming treatments both proved ineffective in helping students address this difficulty compared to control, suggesting that this may be a useful idea to target explicitly in future implementations. In addition, AR and priming performed the same bringing attention to common minor mistakes (for example, missing the factor of $\frac{1}{2}$ in formulating the field due to a half-circular arc or completely ignoring the magnetic field direction). Other errors, such as incorrectly assigning a B-field direction through a misapplication of the right-hand-rule, a skill addressed only briefly in the analogical training conditions, also persist across conditions. Overall, neither analogical comparison nor priming reduced single-concept related errors for this particular target problem.

(c) Analogical Comparison & Grain Size

In addition to comparing the relative effectiveness of the different AR conditions on the target synthesis problem, it is worthwhile to compare how the students responded to the comparison prompts. Overall, student answers to the AR prompts were appropriate and of satisfactory quality. The most interesting result involves the create-a-guide question included at the end of all four of the different analogical reasoning tasks. The students in the single-concept conditions were far more likely to include specific information and solution steps compared to those in the synthesis conditions. To illustrate, Fig. 6.8 shows the percentage of students who included a subset of the potentially relevant components related to Ohms law and circuits. Identifying parallel/series structure and applicable rules for such cases,

Table 6.6: Proportion of students demonstrating several specific errors on the target synthesis problem. Errors shown are ±1 standard errors of proportions.

<table>
<thead>
<tr>
<th>Error Description</th>
<th>Control N=44</th>
<th>Priming N=47</th>
<th>AR N=187</th>
</tr>
</thead>
<tbody>
<tr>
<td>Used total current in all included B-fields</td>
<td>14% ± 5%</td>
<td>9% ± 4%</td>
<td>10% ± 2%</td>
</tr>
<tr>
<td>Missed the 1/2 for the half-loop</td>
<td>39% ± 7%</td>
<td>28% ± 7%</td>
<td>28% ± 3%</td>
</tr>
<tr>
<td>Ignored B-field direction</td>
<td>39% ± 7%</td>
<td>26% ± 6%</td>
<td>26% ± 3%</td>
</tr>
<tr>
<td>Included incorrect B-field direction</td>
<td>27% ± 7%</td>
<td>30% ± 7%</td>
<td>33% ± 3%</td>
</tr>
</tbody>
</table>
reducing the circuit to equivalent resistances, and explicit citation of Ohm’s law are more fine grained descriptions, whereas implicitly using Ohm’s law (i.e. "use the given voltage to solve for the currents") and procedural steps (e.g. identifying unknowns) are more course-grained. Where students in the single-concept conditions were more likely to specifically cite Ohm’s law $[\chi^2(1) = 20.834, p < 0.001]$, circuit reduction $[\chi^2(1) = 28.190, p < 0.001]$, and identify series/parallel circuit structure $[\chi^2(1) = 40.096, p < 0.001]$, students in the synthesis condition discussed the concepts more broadly, and were significantly more likely to implicitly refer to Ohm’s Law $[\chi^2(1) = 14.046, p < 0.001]$, and more general problem solving steps $[\chi^2(1) = 34.525, p < 0.001]$.

The difference between the conditions is likely due in large part to students’ expectations of the task - regardless of condition, students would respond to the prompt with a guide of approximately 3-4 unique steps. Since the prompt was included twice in the single-concept conditions (once for each concept-pair), it is reasonable that those students would include more detailed steps overall.
However, students in the synthesis conditions have at least two *a priori* options: respond with a subset of the important steps at the same grain level (due to availability, cue strength, etc.) or adjust grain level and describe the synthesis problem more generically. Almost invariably students in the synthesis conditions provided guides based on the latter. Although this difference did not seem to introduce a significant effect in performance on the target synthesis question, it may be that analogical comparison using synthesis problems provides an authentic way for students to practice analyzing a physics problem from a more course-grained level, where the focus is on recognition of the relevant concepts and a plan for their application. On the other hand, it seems that comparisons of single-concept examples might be more suited to highlighting the nuances necessary for concept mastery, since such comparisons can focus on those issues without introducing additional load.

### 6.2.3 Discussion

The major takeaways from this study are three-fold: 1) analogical comparison resulted in a significant but small improvement ($d = 0.31$) on the target synthesis problem over course instruction alone 2) analogical comparison was not significantly different from priming and 3) there were no significant main effects from the similarity of the base worked examples either in terms of structure (single-concept vs. synthesis) or provided information (matching the target problem vs. switched).

These findings are inherently connected by the absence of any statistical difference in student recognition of the component physics concepts on the target synthesis problem. As noted in Table 6.5, almost all students recognized the need to calculate both magnetic fields and circuit reduction via Ohm’s law. As such, concurrent concept recognition was not a significant problem for students. This is likely a consequence of the particular combination of concepts and the structure of the target synthesis problem. A schematic of the circuit and given information about particular circuit elements are very strong cues for Ohm’s Law, and the problem statement specifically tipped its hat at the need of at least one (if not more) magnetic fields. This lack of a recognition bottleneck is perhaps the primary reason that this experiment found a smaller effect in overall score, and unlike Experiment #1, no
significant difference between analogical comparison and priming.

As such, the fact that Experiment #2 found no difference between analogical comparison and priming is of particular interest given the earlier results of Experiment #1. For one, it suggests that analogical comparison is likely to be most effective in situations were students are expected to struggle with identification of at least one of the physics concepts. In Experiment #1, energy conservation overshadowed the application of the circular motion constraint: since students could arrive at an answer using only energy-conservation (and the reasonable, but incorrect, assumption that the “minimum” criteria corresponded to the case where the cart had zero velocity at the top of the loop), they ignored the application of a second concept completely. That was simply not the case here.

On the other hand, there are hints that students still struggled with the joint application of both physics concepts. In addition to concept specific difficulties, a small but consistent proportion of students (about 10%) incorrectly combined Ohm’s Law and circuit analysis with a calculation of the total magnetic field. In particular, these students (often correctly) calculated the total current in the circuit, and then proceeded to substitute that current into every identified contribution to the magnetic field. As such, this problem represents evidence of one of the key synthesis problem solving bottlenecks, namely joint application, hypothesized in our earlier work (White et al., 2014). Unfortunately, neither priming nor analogical comparison were effective in eliminating this issue. However, now that this student difficulty has been identified, it may be possible to scaffold further analogical comparison prompts that explicitly confront it.

It is important to note that one of the initial motivations for this study was to try to identify effects from the structural similarity of the base worked examples to the target synthesis problem. Likely in part due to the aforementioned student success with concept recognition, we found no evidence of an effect either due to the structure of the worked examples (single-concept vs. synthesis) or the provided information (matching the target problem vs. switched) on the target synthesis problem. However, there was some evidence that students did react differently to the training via single-concept and synthesis based worked examples. In particular, student responses to the final comparison prompt (create
a short, 3-4 step guide identifying the important elements of the worked examples) suggests that students adjusted the grain size of their responses according to the type of worked example. Rather than focus on a subset of the elements included by students in the single-concept conditions, students who compared synthesis problems were far more likely to take a “big picture” view. This might hint that synthesis problems can be particularly useful for encouraging similar reflection on general problem solving approaches.

Overall, this study suggests that analogical comparison is best suited for problems that represent a challenge for students, particularly in regards to recognition of the pertinent concepts. If a student is able to rely on easily identifiable cues and given-unknown heuristics, analogical comparison is unlikely to invoke the significant gains previously identified by Experiment #1. On the other hand, physics problems that require students to identify concepts through careful consideration of physical context and competing conceptual cues may benefit greatly from such an intervention. As such, the next two experiments seek to test the benefits of analogical comparison with a relatively more complicated set of physics concepts.

6.3 Experiment #3: Analogical Comparison vs. Self-Explanation

6.3.1 Design and Participants

The goals of Experiment 3 were to explore the effectiveness of analogical comparison in the context of more complicated introductory-level synthesis problems (see the target problem in Fig. 6.9B) and to compare the effectiveness of analogical comparison to the method of self-explanation. A full solution of the target problem requires three main conceptual components: simple circuit analysis (Ohm’s law), induced EMF (Faraday’s Law), and magnetic force. In particular, the concepts of magnetic force and Faraday’s Law represent distinct and documented challenges for students (Maloney et al., 2001; Zuza et al., 2014). In the case of magnetic force, students must successfully interpret cross-products. Faraday’s Law requires an implicit understanding of magnetic flux and the consideration of direction via
Lenz’s Law. Correspondingly, a compact solution of the target synthesis problem utilizes not only three basic physics equations, but considerably more algebraic manipulations than that required to solve Experiment 1. Along with the target synthesis problem, we designed a corresponding set of single-concept and synthesis worked examples (see Fig. 6.9A).

Figure 6.9: Example problems given as worked examples (A) and target synthesis problem (B).

Experiment 3 had six conditions (see Fig. 6.10). Four of the conditions (control and three analogical comparison conditions) were similar in structure to Exp. 1. The fifth condition was aimed at avoiding a potential shortcoming of the other analogical comparison conditions. This condition represented a “best-effort” attempt to scaffold compar-
isons through a mix of recognition and single-concept focused prompts. In particular, the
prompts explicitly addressed the concept of induced emf and its application within the two
synthesis worked examples. For example, one prompt in this condition states, “One of the
important concepts is that of an induced emf due to a changing magnetic flux. Compare
the physical reason for a changing magnetic flux in each of the problems. Explain your
answer highlighting any differences between the two problems.”. In contrast, the other
conditions did not explicitly invoke the concept of Faraday’s Law. Further, the combined
“best-attempt” condition tried to scaffold comparison that followed the worked examples:
starting with recognition of the relevant concepts, through consideration of the physical
context for Faraday’s Law, and then subsequent application of the component concepts in
the worked examples. The other analogical comparison conditions were restricted to focus
on either recognition of the concepts or their application as in Experiments 1 and 2. Full
versions of all worked examples and corresponding comparison prompts are included in
Appendix B.3.

Finally, the sixth condition was the self-explanation condition. This condition simply
prompted students to explain (write) both of the synthesis worked examples to a friend,
but did not invoke explicit comparison between the two. Ample space was provided for the
explanation.

Tasks were administered during the Spring of 2016. Participants were students in the
second semester of a calculus-based introductory electromagnetism course at The Ohio State
University. Students participated as part of a 1-hour long flexible homework assignment for
course credit. Students completed the flexible assignment over a three week window, after
course instruction on Faraday’s Law, and in close proximity to a course midterm covering
the relevant material. A total of 254 students were randomly assigned into one of the six
study conditions.

Students completed the training tasks and target synthesis problem in individual carrels
in a quiet room. An equation sheet similar to those used in the course was provided to all
students. Tasks were administered and collected by the proctor one at a time, and students
were allowed to work at their own pace. Whereas students in the analogical comparison
conditions were given all relevant worked examples and prompts together, students in the summary condition were given one synthesis worked example to summarize at a time. Students first completed their selected training, followed by 10-15 minutes of unrelated physics tasks, and then the target synthesis problem. Students in the control condition completed a set of unrelated physics tasks and the target synthesis problem.

A rubric for assessing student solutions of the target synthesis problem was determined by two of the authors. The rubric is shown in Table 6.7. As in Experiment 1, recognition of component physics concepts was coded generously, but required the student to commit to using the concept as part of their solution. A random sample of 25 student solutions was coded by two researchers with an inter-coder agreement of 84%. Disagreements were discussed and resolved.
Table 6.7: Scoring rubric for the target synthesis problem. 1 point was awarded for each item, for a total of 10 points.

<table>
<thead>
<tr>
<th>Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1 Faraday’s Law/induced emf</td>
</tr>
<tr>
<td>+1 Ohm’s Law</td>
</tr>
<tr>
<td>+1 Magnetic Force</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1 Correct induced emf</td>
</tr>
<tr>
<td>+1 Correct induced emf direction/Lenz’s Law</td>
</tr>
<tr>
<td>+1 Ohm’s Law: series circuit/combined resistors in series</td>
</tr>
<tr>
<td>+1 Ohm’s Law: combined voltage sources</td>
</tr>
<tr>
<td>+1 Correct magnetic force equation</td>
</tr>
<tr>
<td>+1 Correct applied force direction</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1 Correct final answer \ No mathematical mistakes</td>
</tr>
</tbody>
</table>

6.3.2 Results

(a) Student performance on the target synthesis problem

Student final course grades in their introductory electromagnetism class were collected and compared across conditions. The same cuts conducted in Experiment 1 were applied to eliminate outliers. Students must have completed the course (removing no students), and have scored no lower than 2-standard deviations below the mean (removing a total of 9 students, ranging from 1-4 students per condition). A one-way ANOVA of course grade showed no significant differences across conditions \( F(5, 239) = 0.554, p = 0.735 \).

Mean scores on the target synthesis problem are shown in Table 6.8. We first compared the interventions to control. A one-way ANOVA showed significant differences between conditions \( F(5, 239) = 5.961, p < 0.001 \). A Tukey HSD post-hoc showed significant differences between Summarization and Control \( (d = 1.14, p < 0.001) \) and Synthesis Combined Prompts and Control \( (d = 0.67, p = 0.018) \). The other analogical comparison conditions were not significantly different from control: Single-concept – Mastery \( (d = 0.18, p = 0.959) \), Synthesis – Mastery \( (d = 0.44, p = 0.363) \), and Synthesis – Recogni-
Table 6.8: Mean score on target synthesis problem out of a maximum score of 10 points. Errors shown are standard errors. Analogical comparison conditions are labeled AC.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean Score ± SE</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>5.19 ± 0.35</td>
<td>42</td>
</tr>
<tr>
<td>AC: Single Concept Examples with Mastery Prompts</td>
<td>5.60 ± 0.35</td>
<td>42</td>
</tr>
<tr>
<td>AC: Synthesis Examples with Mastery Prompts</td>
<td>6.16 ± 0.35</td>
<td>38</td>
</tr>
<tr>
<td>AC: Synthesis Examples with Recognition Prompts</td>
<td>6.19 ± 0.35</td>
<td>42</td>
</tr>
<tr>
<td>AC: Synthesis Examples with Combined Prompts</td>
<td>6.75 ± 0.37</td>
<td>40</td>
</tr>
<tr>
<td>Summarization</td>
<td>7.54 ± 0.28</td>
<td>41</td>
</tr>
</tbody>
</table>

A one-way ANOVA was conducted between only the intervention conditions to test for hypothesized differences between single-concept worked examples and synthesis worked examples, between mastery and recognition prompts, and between summarization and the combined, best-effort analogical comparison condition. The one-way ANOVA showed significant differences between interventions \( F(4, 198) = 4.697, p = 0.001 \). A Tukey HSD post-hoc showed no significant difference in total score on the target synthesis problem between Single-concept – Mastery and Synthesis – Mastery \( (d = 0.25, p = 0.776) \). There was no significant difference between Synthesis – Mastery and Synthesis – Recognition \( (p = 1.00) \). Finally, although there were no significant differences between Summarization and Synthesis – Combined \( (d = 0.38, p = 0.48) \), the Summarization treatment significantly outperformed all of the other analogical comparison conditions: Single-concept – Mastery \( (d = 0.95, p = 0.001) \), Synthesis – Mastery \( (d = 0.70, p = 0.042) \), and Synthesis – Recognition \( (d = 0.66, p = 0.042) \).

To sum up, only the best-effort attempt at analogical comparison (synthesis worked examples using a combination of scaffolded comparison prompts) and the summarization intervention were significantly better than course-instruction alone (control) in terms of overall student performance on the target synthesis problem. In addition, there were no statistical differences in the effectiveness of analogical comparison based on either the type of worked example or the type of comparison prompts. Finally, summarization via self-explanation was the most effective intervention, with significant differences in student performance ver-
sus all analogical comparison conditions except for the highly-scaffolded “best-attempt” condition.

In addition to overall performance on the target synthesis problem, we compared student conceptual recognition across conditions. The proportion of students recognizing each component physics concept is shown in Fig. 6.11. The vast majority of students recognized and utilized Ohm’s Law (≥ 93%) and magnetic force due to a current carrying wire (≥ 83%) across all conditions.

A chi-squared test across treatment conditions showed there was a significant difference in the proportion of students recognizing and utilizing Faraday’s Law on the target synthesis problem [χ²(4) = 25.543, p < 0.001]. To test for hypothesized differences between the interventions, post-hoc comparisons between treatments were conducted using pairwise chi-squared tests with a Bonferroni correction for multiple comparisons (n = 3). The adjusted p-values are reported. To test for the hypothesized difference due to type of worked example,
a comparison of Single-concept – Mastery and Synthesis – Mastery showed no significant difference in the proportion of students recognizing Faraday’s Law [$\chi^2(1) = 0.029, p = 1.00$]. To test for the hypothesized difference due to type of comparison prompt, a comparison of Synthesis – Mastery and Synthesis – Recognition also showed no significant difference [$\chi^2(1) = 1.749, p = 0.558$]. Finally, a comparison between Synthesis – Combined and Summarization showed a significant difference in the proportion of students recognizing Faraday’s Law on the target synthesis problem [$\chi^2(1) = 6.316, p = 0.036$].

Overall, there was a statistically significant difference in the proportion of students identifying Faraday’s Law between the Summarization condition and the best-attempt, combined analogical comparison treatment with the summarization group outperforming the combined analogical comparison group. However, there was no significant effect from either the type of worked example or the type of comparison prompts on student recognition of Faraday’s Law for the target synthesis problem. Both results are in agreement with similar findings for total score on the target problem.

In order to further examine the effect of training on how students approached the target synthesis problem, we compared the proportion of students across conditions who recognized and applied Faraday’s Law, Ohm’s Law, and the magnetic force on a current carrying wire, and also explicitly calculated a total current in the circuit that combined the voltage due to the battery with the induced emf (but not necessarily with correct directions or magnitudes). This combination was used to represent the minimum structure necessary for a correct approach to the target synthesis problem. The proportion of students successfully meeting this threshold is shown in Fig. 6.12.

A chi-squared test across treatment conditions showed there was a significant difference in the proportion of students meeting this structural threshold [$\chi^2(4) = 32.482, p < 0.001$]. To test for hypothesized differences between the interventions, post-hoc comparisons between treatments were conducted using pairwise chi-squared tests with a Bonferroni correction for multiple comparisons ($n = 3$). Tests for hypothesized differences due to type of worked example and type of prompt showed no significant differences (i.e. comparison between Single-concept – Mastery and Synthesis – Mastery and between Synthesis –
Figure 6.12: Proportion of students who had the correct solution structure on target synthesis problem. Students with the correct structure included all 3 relevant physics concepts, and both contributions to the total emf: the battery voltage and the induced emf. Errors shown are ±1 standard errors for proportions.

Mastery and Synthesis – Recognition respectively). However, comparison between Synthesis – Combined and Summarization did show a significant difference in the proportion of students who applied all component concepts and calculated the combined total current $\chi^2(1) = 8.320, p = 0.012$. Simply put, the difference does not come from the first four treatment groups in Fig.7; instead it comes from the last group.

(b) Student performance on the summary training task

In order to identify which elements of the worked examples were attended to by students, student summaries collected from the training intervention were coded for attended concepts and any included details related to the application of those concepts within the synthesis worked examples. Although there were pairs of student summaries that either provided only general problem solving steps (1 student) or paraphrased the steps of the example solutions without additional explanation (5 students), most student summaries included additional, relevant elaborations of at least one of the included concepts (88%). No students included references or comparisons to the first worked example during their discussion of the second worked example.
Most students included Faraday’s Law (97%), magnetic force on a current carrying wire (92%), and Ohm’s Law (78%) in both summaries. Students invoked specific details relating to the application of those concepts less frequently. For example, for the first worked example involving a circuit falling out of a uniform magnetic field, students explicitly identified the changing area in the magnetic field as the reason for the changing magnetic flux 46% of the time; while 39% discussed the resulting direction of the induced emf. For the second worked example, which placed a circuit in a time-dependent magnetic field, the students explicitly identified the changing magnitude of the magnetic field as the reason for the changing magnetic flux 68% of the time. In addition, more student summaries included a discussion of the resulting EMF direction (71%) than with the first worked example. In the first problem, student’s primarily discussed the magnetic force in the context of Newton’s second law and corresponding force balance. Few students explicitly verified that the direction of the magnetic force was consistent with the direction of the total current in the circuit and the direction of the magnetic field (12%). Slightly more students explicitly discussed the cross-product in the context of the second problem (39%). Overall, more students correctly identified the cause of the emf in the second problem than in the first.

The order in which students discussed the physics concepts was tracked across the student summaries (summaries often proceeded single-column and top-down, or were labeled to indicate order). The majority of student summaries discussed the concepts in the order that they were presented in the worked solution (71% for the first worked example, and 83% for the second worked example). In addition, the majority of students alluded to the identification of unknowns as part of their summaries. In particular, students frequently (63% of the time) identified the current in the worked examples as an important intermediate quantity. For example, one student noted, “The next step is to solve for the current since L, B, g are all known...however, to properly solve this we need to solve for the \( \mathcal{E} \) created by the B-field.”
6.3.3 Discussion

There are two main takeaways from Experiment 3. First, of the analogical comparison interventions, only the scaffolded “best-attempt” analogical comparison condition performed significantly better than control in terms of overall performance on the target synthesis problem. Moreover, there were no significant differences in the effectiveness of the analogical comparison interventions based on the type of worked example or the type of comparison prompt. With other factors controlled, single-concept and synthesis worked examples produced similar results. The case was the same for prompts focused on mastery or recognition. Second, summarization of synthesis worked examples alone was not only the most effective intervention in improving students’ overall performance on the target problem, but it also significantly improved the recognition and use of Faraday’s Law compared to every other intervention, including the best-attempt analogical comparison condition.

The increased complexity of the target problem in Experiment 3 (compared to Experiments 1 and 2) manifested itself in the details of student performance on the problem. In Experiment 1, student conceptual recognition was the dominant bottleneck to correctly solving the target problem; once that difficulty was successfully overcome by the analogical comparison interventions, students correctly applied the component physics concepts. In Experiment 3, recognition alone was not enough to guarantee a completely correct solution even for the best performing intervention, summarization. The summarization condition resulted in 90% of students recognizing and applying Faraday’s Law (and even more recognizing the other two component concepts), but the mean score was still only 7.54/10, in part due to remaining difficulties applying Lenz’s Law and cross-product-based reasoning.

Along with the increased complexity of the target synthesis problem, the synthesis worked examples were of a correspondingly higher level of complexity. We suggest that this manipulation of the synthesis worked example complexity is a likely reason for the subsequent lack of statistically significant gains for those analogical comparison conditions. The importance of the increased complexity seems even more likely considering that the only statistically different analogical comparison condition versus control was the combined
analogical comparison treatment. The combined prompts provided more information to
the student (explicitly identifying Faraday’s Law) and scaffolded the comparison of the two
worked examples more extensively than any of the other analogical comparison interven-
tions. In order for the analogical comparison to be effective with the more complicated base
examples, additional scaffolding was necessary.

The necessity for additional scaffolding may not be too surprising in and of itself given
the cognitive load and increased demands from the more complicated worked examples.
However, it is important to note that once-again comparisons of considerably simpler single-
concept worked examples did not significantly improve student performance on the target
problem. It appears that the structure of a synthesis problem cannot simply be broken
into parts, even when students are asked to compare the full set of parts immediately after
one another, and despite the fact that the single-concept worked examples maintained the
surface features and component steps of the combined synthesis worked examples. Unfor-
tunately, this also suggests that analogical comparison with single-concept examples alone
is unlikely to be an effective way to help students reduce the cognitive load inherent to
complicated synthesis problems.

On the other hand, despite the lack of instructional scaffolding or information beyond
the worked example, students in the summary condition performed significantly better than
the control group on the target synthesis problem. Further, students in the summarization
condition performed significantly better in recognizing the need for Faraday’s Law compared
to the best analogical comparison condition – despite the fact that the best-effort analogical
comparison condition explicitly pointed out the concept as one of the comparison prompts.
This result was further supported by the proportion of students who generated the correct
solution structure on the target problem – using all three identified physics concepts and
explicitly combining both the induced emf and the battery voltage (c.f. Fig. 6.12).

Given these results, there are several possible explanations for the relative success of the
summarization condition. First, summarization might have been more successful because
the task of self-explaining each problem, one-at-a-time, allowed students to better encode
the entire structure of each independent worked example. In contrast, students in the
analogical comparison condition may have made sense of smaller-grain component concepts across the base examples, but without encoding the full structure of either example. This lack of encoding the full structure in the analogical comparison conditions could be due to a simple failure of students to satisfactorily read through the base worked examples – that is, beyond what was necessary to answer each specific invoked comparison – or more nuanced differences in how the students extracted the structure of the provided solutions. Experiment 4 was designed to help exclude the first possibility. Two mechanisms proposed by Chi (2000) to explain the success of self-explanation, namely inference generation and mental-model revision, may account for other underlying differences in student encoding.

The mechanism of inference generation suggests that the summarization condition may have led to higher performance via self-explanation by prompting students to fill-in necessary information and reasoning steps missing in the worked examples. There is some evidence for such an effect: most students included not only the relevant physics concepts in their summaries, but also additional justifications. For example, Faraday’s Law and induced emf were addressed in almost all student summaries and more than half of the summaries over the two worked examples described the physical reason for the changing magnetic flux. Moreover, student summaries seemed to follow the reasoning of the provided solutions. The majority of student summaries discussed the concepts in the order that they were presented in the worked solutions, while also identifying important intermediate quantities. Although that recognition was probably driven in part by students relying on a given-unknown problem solving heuristic, it may have allowed students to identify the electric current as a structural connection between the physics concepts in the problem. In contrast, the connections between the analogical comparisons prompts may not have been as strongly internalized by students, even though the combined analogical comparison condition invoked comparisons that explicitly targeted those exact elements.

The second mechanism, mental-model revision, could also account for the relative success of the summarization condition. The target synthesis problem – and its potentially novel combination of an induced emf and a battery – represented a significant challenge for students. It is possible that students had an incomplete or disjointed prior understanding
of electromotive force. If that was the case, summarization via self-explanation may have helped students to reconcile their mental model with the worked examples. Future work is needed to differentiate between these possible mechanisms.

### 6.4 Experiment #4: Analogical Comparison using Self-Explanation

#### 6.4.1 Design and Participants

This experiment was built upon the previous study to both replicate the relative success of the summarization intervention and test whether analogical comparison and self-explanation of the individual worked examples can be combined for further benefit. As such, this study used the same target problem employed in Experiment 3. In addition to a no-training control, four treatments were included in the experimental design. To replicate the results of Experiment 3, we once again included a “best-effort” analogical comparison condition and summarization condition, using the same worked examples as Experiment 3. The prompts for the analogical comparison condition were similar to those used previously and explicitly invoked the concept of Faraday’s Law. Given student’s prior difficulty with applying the individual concepts – in particular, difficulties with direction-based considerations due to Lenz’s Law and cross-products – we made adjustments to the prompts to encourage further comparison of the important directions identified in the worked examples. The full set of prompts is included in Appendix B.4.

The other two conditions were designed to test whether an increased emphasis on encoding the individual worked examples before analogical comparison would increase student performance on the target synthesis problem. First, we included an “annotation” condition that asked students to very briefly comment on the two individual worked examples. This was done in order to test whether ensuring that students first thoroughly read through the worked examples prior to completing the comparison prompts would increase student performance on the target synthesis problem. The prompts were intended to be brief checks on students reading and understanding of the example solutions and simply asked the students
to identify both the concepts used and the goal of each section in the worked examples (i.e. “Ohm’s Law” and “Find the total current”). The presentation of these reading annotations is also shown in Appendix B.4.

The last condition sought to test the hypothesized idea that summarization and analogical comparison could be combined for additional benefit. Given the increased complexity of the base worked examples, we posited that inviting students to first summarize the individual worked examples independently would facilitate subsequent analogical comparison. Consequently, students may have a better holistic understanding of the base examples and how individual comparisons fit within the two overall solution structures, rather than viewing them as a set of unconnected, piecewise comparisons. As such, the combined summarization + analogical comparison condition asked students to first briefly summarize each worked example before prompting them to compare across the two worked examples. The full experiment design is shown in Fig. 6.13.

![Figure 6.13: Illustration of the design and administration of Experiment #4.](image)

Tasks were administered during the Fall of 2016. Participants were students in an off-sequence calculus-based introductory electromagnetism course at The Ohio State University,
who participated as part of a 1-hour long flexible homework assignment for course credit. The flexible homework assignment was administered over a three week period near the end of the semester, approximately one month on average after course instruction on Faraday’s Law. A total of 232 students were randomly assigned into one of the five study conditions.

Students completed the training tasks and target synthesis problem in individual carrels in a quiet room. Tasks were administered and collected by the proctor one at a time and an equation sheet was provided to all students. Students first completed their selected training, followed by 10-15 minutes of unrelated physics tasks, and then the target problem, as shown in Fig. 6.13. Whereas students in the analogical comparison and annotation conditions were given all relevant worked examples and prompts together, students in both the summary and combined summary + analogical comparison conditions were given only a single synthesis worked example to summarize at a time. After completion, the worked example and the student’s summary were collected and the student was given the second worked example to summarize. Students in the control condition once again completed a set of unrelated physics tasks and the target synthesis problem.

Student time-on-task was digitally recorded by the proctor. Due to time constraints from the testing format, students in the combined summary + analogical comparison condition were given a time limit of approximately 7 minutes per individual summary. Time-on-task during the training interventions was collected for the vast majority of students (95%).

Student solutions to the target synthesis problem were graded with the same rubric used in Experiment 3 (shown in Table 6.7). A random sample of 25 student solutions was coded independently by two researchers. Any differences in coding were discussed and resolved leading to an inter-coding agreement of 88%.

6.4.2 Results

(a) Student performance on the target synthesis problem

Student final course grades in their introductory electromagnetism class were collected and compared across conditions. The same cuts conducted previously were applied to eliminate
Table 6.9: Mean score on target synthesis problem out of a maximum score of 10 points. Errors shown are standard errors.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean Score ± SE</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>4.04 ± 0.30</td>
<td>45</td>
</tr>
<tr>
<td>Analogical Comparison</td>
<td>5.74 ± 0.35</td>
<td>46</td>
</tr>
<tr>
<td>Analogical Comparison + Reading Annotations</td>
<td>5.93 ± 0.32</td>
<td>45</td>
</tr>
<tr>
<td>Summarization</td>
<td>6.29 ± 0.34</td>
<td>45</td>
</tr>
<tr>
<td>Analogical Comparison + Summarization</td>
<td>7.07 ± 0.34</td>
<td>44</td>
</tr>
</tbody>
</table>

outliers. Students must have completed the course (removing no students), and have scored no lower than 2-standard deviations below the mean (removing a total of six students, ranging from 0-2 students per condition). Almost all students satisfactorily completed the training tasks. One student, who did not complete the training task, was removed from the combined analogical comparison + summarization condition. A one-way ANOVA of course grade showed no significant differences across conditions \(F(4, 220) = 0.522, p = 0.719\).

Mean scores on the target synthesis problem are shown in Table 6.9. First, interventions were compared to control. A one-way ANOVA showed significant differences in total score on the target synthesis problem between conditions \(F(4, 220) = 11.351, p < 0.001\). A Tukey HSD post-hoc showed significant differences for all treatment conditions compared to control: Comparison-only \((d = 0.77, p = 0.003)\), Comparison + Annotations \((d = 0.92, p = 0.001)\), Summarization \((d = 1.05, p < 0.001)\), and Comparison + Summarization \((d = 1.44, p < 0.001)\).

A one-way ANOVA was conducted between only the intervention conditions to test for hypothesized differences in student performance on the target synthesis problem. The one-way ANOVA showed significant differences between the treatments \(F(3, 176) = 3.002, p = 0.032\). In order to test for hypothesized differences, a Tukey HSD post-hoc was conducted. The post-hoc tests showed no significant difference in total score on the target synthesis problem between Comparison-only and Comparison + Annotations \((d = 0.08, p = 0.977)\), nor between Comparison-only and Summarization \((d = 0.23, p = 0.653)\), but there was a significant difference between Comparison-only and Comparison + Summarization \((d = \)
In summary, the combination of summarization and analogical comparison was statistically better than analogical comparison alone. However, there was no significant difference between students who completed only the analogical comparison prompts and the students who were first explicitly asked to read through each worked example and provide brief annotations before making comparisons. There was also no significant difference between summarization alone and analogical comparison alone ($d = 0.23$), though the trend was in the same direction as in Experiment 2 ($d = 0.38$), with summarization performing nominally better than analogical comparison. Taken together, this replication may suggest a small, but potentially significant effect ($d \approx 0.3$).

The proportion of students recognizing each component physics concept and employing it as part of their solution on the target synthesis problem is shown in Fig. 6.14. Across all conditions, the majority of students successfully recognized and utilized Ohm’s Law ($\geq 80\%$) and magnetic force due to a current carrying wire ($\geq 93\%$). A chi-squared test across all conditions showed there was a significant difference in the proportion of students recognizing and utilizing Faraday’s Law on the target synthesis problem [$\chi^2(4) = 39.140, p < 0.001$]. However, such a difference went away if only the treatment groups were compared [$\chi^2(3) = 5.658, p = 0.129$]. This suggests that while the four treatment conditions all significantly improved concept recognition versus control, they did not differ significantly from one another.

In addition, we compared the proportion of students across the conditions who met the minimum threshold for a correct approach to the target problem, namely recognize and apply all three concepts and calculate a total current using both the induced emf and battery voltage. The results are shown in Fig. 6.15. Along with clear differences between the interventions and control, a chi-squared test was used to compare students who did not summarize the individual worked examples with those who did as part of their training. The chi-squared test showed a significant difference in the proportion of students who met the proposed threshold on the target synthesis problem [$\chi^2(1) = 24.360, p < 0.001$]. Students who summarized the individual worked examples as a part of their training were significantly
more likely to recognize all three concepts and combine the two sources of emf.

(b) Time on task

Time-on-task during training was recorded for the vast majority of students (95%) and compared across the four intervention conditions. Students without timing data were removed from the subsequent analysis, resulting in the corresponding boxplots presented in Fig. 6.16. A median test showed significant differences in time spent during training across the conditions. In particular, Comparison + Annotation and Comparison + Summarization represent an increase of approximately 20% and 35% (five and eight minutes respectively) in training time over the Comparison-only condition.

To assess the effect of time-on-task and student aptitude in determining student performance on the target synthesis problem, we compared total score on the target synthesis problem between the Comparison + Summarization and Comparison-only conditions using
Figure 6.15: Proportion of students who had the correct solution structure on target synthesis problem. Students with the correct structure included all three relevant physics concepts, and both contributions to the total emf: the battery voltage and the induced emf. Errors shown are ±1 standard errors for proportions.

Figure 6.16: Time spent on each respective training task by condition. The number of included students in each condition is noted. Time-on-task during training was recorded for 95% of students.
a general linear model, accounting for a main effect (condition) and two covariates (course grade and time-on-task). Outliers in time-on-task were removed based on inspection of the boxplot presented in Fig. 6.16. Both course grade (partial $\eta^2 = 0.14, p = 0.001$) and time-on-task (partial $\eta^2 = 0.05, p = 0.048$) were found to significantly predict student performance on the target synthesis problem. Condition was marginally significant (partial $\eta^2 = 0.04, p = 0.093$). These results suggest that student aptitude, as measured by course grade, is the strongest predictor of subsequent performance on the target synthesis problem. Moreover, though the model lacked the statistical power to observe the effect at the 0.05 level, it suggests some evidence that combining analogical comparison and summarization may provide a small, positive effect beyond that accounted only by the additional time-on-task.

(c) Student performance on the training tasks

Student annotations of the two worked examples from the Analogical Comparison + Annotation condition were examined to verify that students satisfactorily read through the full solutions. With the exception of several mislabeled concepts (for example, two students cited Ampere’s Law instead of Faraday’s Law, and others Kirchoff’s Law instead of Ohm’s Law), almost all student annotations clearly identified both the use of component concepts (induced emf) and intermediate steps (solve for currents).

In addition, student summaries collected from the training interventions were coded for attended concepts and details of application of those concepts relevant to the synthesis worked examples. We compared students in the Summarization-only condition to students who were asked to briefly summarize each worked example (with an approximate 7 minute limit per summary) in the combined Comparison + Summarization condition. The majority of students in both conditions included conceptual elaborations explaining at least one of the concepts in both of the worked examples (89% for summarization alone and 80% for the combined condition); the remaining students provided only paraphrases of solution steps, without additional explanation (five and nine students respectively).

However, students in the Summarization-only condition frequently presented more over-
all detailed descriptions than students in the combined Comparison + Summarization condition. To this, we first consider student summaries from the Summarization condition. Students discussed Faraday’s Law (89%), Ohm’s Law (82%), and magnetic force (93%) in both summaries. For the first worked example, which involved a circuit falling out of a uniform B-field, students explicitly identified the changing area in the magnetic field as the reason for the changing magnetic flux 40% of the time. The direction of the induced emf was considered in 36% of student summaries. For the second worked example, which placed a circuit in a time-dependent magnetic field, students explicitly identified the changing magnitude of the magnetic field as the reason for the changing magnetic flux 62% of the time, while 60% of student summaries included a discussion of induced emf direction.

In comparison, student summaries from the combined Comparison + Summarization condition referred to the individual component physics concepts less frequently, with corresponding decreases in the frequency of elaborations of the physical situation, physical quantities, and directions. Faraday’s Law (75%), Ohm’s Law (50%), and magnetic force (80%) were explicitly discussed in both summaries by the majority of students. However, there was a significant difference between the proportions of students who included explicit discussions of Ohm’s Law \( \chi^2(1) = 10.337, p = 0.001 \) and marginal differences between the proportion of students who discussed magnetic force \( \chi^2(1) = 3.626, p = 0.057 \) and Faraday’s Law \( \chi^2(1) = 2.910, p = 0.088 \) when compared to Summarization-only.

The decrease in students explicitly discussing Ohm’s Law reflects a shift in student summaries towards less-detailed and larger grain descriptions in the time-restricted Comparison + Summarization condition. For example, rather than write out explicit descriptions of Ohm’s law, voltage sources, and the simple series circuit, students in the combined Summarization and Analogical comparison condition would often only mention the need to find a current. In a similar fashion, fewer students included explicit discussions of the physical situation and the corresponding cause of the induced emf (16% and 27% for the first and second worked examples respectively), and induced emf direction (9% and 43% respectively). Overall, these results suggest that while students in the time-restricted Comparison + Summarization condition considered Faraday’s Law and its role in both worked
examples, students provided fewer details and in depth discussions of the application of that concept.

### 6.4.3 Discussion

Taken together, the results from Experiment 4 support three broad conclusions. First, both analogical comparison and summarization were effective in improving student performance on the target synthesis problem versus control, though there were differences between analogical comparison and summarization in the proportion of students employing the correct solution structure on the target problem. Second, the combination of summarization and analogical comparison was significantly more effective than analogical comparison alone. Third, student annotations of the worked examples were accurate, but did not significantly improve student performance on the subsequent target problem relative to analogical comparison alone.

While the mean score on the target synthesis problem for the control condition in this study was 4.04/10, with only 16% of students recognizing and applying Faraday’s Law, the mean score on the same target problem in Experiment 3 was 5.19/10, with 38% of students recognizing and applying Faraday’s Law. In order to provide context for these results in light of previous findings, we note that the population sample used in this study differed in two potentially important ways from the sample in Experiment 3. First, though the two samples were drawn from different semesters of the same introductory electromagnetics course, students in this study completed the course off-sequence. Second, students in this study completed the training task and target problem over a three week period near the end of the semester, approximately one month on average after course instruction on Faraday’s Law. In contrast, students in Experiment 3 completed the training over a similar period that began closely following in-course instruction, and in proximity to an in-course exam on the relevant topics. Although we cannot exclude differences in on/off-course sequence, previous research tracking student understanding over the duration of an introductory course suggests timing differences between in-course instruction and administration of the training and target synthesis problem are a potential explanation for the observed differences in baseline
student performance between the two studies (Heckler and Sayre, 2010; Ding et al., 2008).

This difference in baseline performance suggests an important distinction when discussing the relative effectiveness of the analogical comparison and summarization conditions. This study found no significant differences in either total score on the target synthesis problem or the proportion of students recognizing and applying Faraday’s Law between the summarization and analogical comparison conditions. In contrast, Experiment 3 found that the summarization condition resulted in significantly more students recognizing and applying Faraday’s Law than in the analogical comparison condition. Moreover, the overall proportion of students in the summarization condition, who were able to construct the correct solution structure for the target synthesis problem, was considerably lower than in Experiment 3.

At the same time, these findings support the hypothesis that there is a meaningful difference in how successfully students encoded the base worked examples between the summarization and analogical comparison conditions: in particular, the replicated finding that significantly more students in the summarization condition constructed the correct solution structure for the target synthesis problem than in the analogical comparison condition. Moreover, the finding of no significant difference between analogical comparison alone and analogical comparison with annotations suggests that these differences are likely not due to students simply failing to sufficiently read through the worked examples in the comparison conditions. In other words, students in the annotation condition satisfactorily labeled physical concepts and key steps within both worked examples with no difference in student performance on the target synthesis problem compared with just analogical comparison alone.

As a result, it is more likely that the success of analogical comparison of the two worked examples was limited by cognitive load and not a failure of students to appropriately attend to the task – here, and potentially in Experiment 3. There are several additional pieces of evidence in support of such a claim. First, student performance on the target problem once again indicated significant and persistent student difficulty with applications of the single-concepts – in particular, determining physical directions associated with Lenz’s Law.
and cross products. Unlike Experiment 1 where students demonstrated a high degree of mastery of the two component concepts after recognizing the need for their simultaneous application, students continued to struggle with these single-concept difficulties regardless of intervention. Second, student summaries indicate that there might have been differences in difficulty between the worked examples themselves that might have limited students’ ability to create a general solution schema via analogical comparison. As such, students might have relied on only the more accessible worked example when solving the target synthesis problem, rather than a more generalized schema from comparison of the two. Third, there was a significant difference between analogical comparison alone and analogical comparison after summarization, both in terms of total score and the proportion of students demonstrating the correct solution structure on the target synthesis problem, suggesting that the initial self-explanation from the summaries aided the comparison.

One limitation of this study is that it was unable to make a definitive distinction between the value added by the combination of analogical comparison and summarization and the corresponding additional time on task. This limitation was a consequence of constraints involving the administration of the task, available contact time, and number of participants. However, the overall significant difference between analogical comparison alone and the combination of summarization and analogical comparison is still of particular value, as it suggests at least one pedagogically relevant way to help students analyze similar, complicated synthesis problems. In other words, the additional approximate 8 minutes to time on task from the combined intervention was well spent, engaging, and resulted in significant gains on the target synthesis problem; in contrast, the additional time spent required by the reading annotations was not inherently productive.

6.5 Summary

Taken as a whole, these four experiments demonstrate that the instructional methods of analogical comparison and self-explanation of worked examples can successfully be extended to improve student performance on target synthesis problems. As such, this work represents
a novel contribution to the study of these techniques beyond previous work predominantly studying their application with single-concept examples. Moreover, the results of these experiments suggest several principles regarding the conditions under which the two instructional methods are likely to be effective.

First, analogical comparison only resulted in significant increases in student performance over baseline when the comparisons were invoked between two worked synthesis examples. There were no significant improvements when students were asked to compare corresponding sets of single-concept problems, despite the same prompts and surface features in the worked examples. This is particularly important given the increase in the target synthesis problem complexity from Experiment 1 to Experiment 3 both in terms of the conceptual difficulty represented by the involved concepts and the requisite number of equations and algebraic manipulations necessary for a complete solution. The finding that breaking the base worked examples into component parts was not significantly different than either unguided problem solving practice (Experiment 1) or baseline performance (Experiments 1&3) emphasizes the importance of the combined structure and joint application of concepts within synthesis problems. Even with explicit and sequential comparisons of the component parts, students still cannot be expected to successfully transfer those parts to a novel synthesis problem without extensive efforts to explicitly scaffold the missing structure of the synthesis problem. Unfortunately, this also suggests that the reduction of synthesis worked examples into only individual single-concept parts is unlikely to be a successful way to reduce cognitive load for more complicated problems.

Second, these results show that while analogical comparison and self-explanation of worked examples can be effective in improving student conceptual recognition and the use of the correct solution structure on a target synthesis problem, pervasive difficulties associated with single-concept mastery - such as how to apply Lenz’s Law - may not be as successfully remedied. In part, this suggests that synthesis problems and the instructional methods used here can best be employed to help students explicitly practice concept recognition, as opposed to the practice with the plug-and-chug search heuristics often associated with single-concept problems. Meanwhile, issues of single-concept mastery may benefit
from complementary and targeted practice in single-concept problems. As such, synthesis problems represent another type of tool for physics instructors, similar to other classes of physics problems like context-rich problems and jeopardy problems.

Third, the structure of the synthesis problem in question likely plays an important role in determining both the necessity and potential success of interventions using analogical comparison and self-explanation. In Experiment 1, students can forgo (and frequently did) the circular motion constraint and subsequent application of centripetal acceleration if they naively assumed that the velocity of the cart was zero at the top of the loop. In Experiments 3 and 4, students could arrive at an answer by only considering the battery voltage. As such, students were not blocked from the final answer only because of a missing unknown, but rather conceptual consideration of the physical situation. It is possible that the large gains in concept recognition reflect this inability of students to successfully rely only on a given-unknown search heuristic. On the other hand, in Experiment 2 where student recognition was not a significant bottleneck, analogical comparison was not significantly different than single-concept practice. As a result, future multi-concept problems where recognition is strongly cued by obvious constraints, given information, or problem statements may not benefit as strongly from these types of interventions. Similarly, mathematically sequential synthesis problems that require students to first use one concept to solve for an unknown that is then necessary for the application of the second concept may not represent as significant a challenge for students.

Fourth, there is evidence that summarization and analogical comparison can be combined for additional gains. Although it is not completely clear whether those gains are solely due to increased time on task or represent an additional inherent difference between the treatments, this finding is of pedagogical value. The addition of < 10 minutes of total time on task necessary for students to briefly self-explain the individual worked examples before comparison resulted in significant improvements in total score on the target synthesis problem.
This thesis addressed two distinct areas of research. In Volume I, we discussed six experiments that investigated the relative effectiveness of short answer questions versus multiple choice in the context of computer-based training. Volume II examined a series of four experiments with worked examples that compared the effectiveness of analogical comparison and self-explanation in order to improve student performance on multi-concept physics problems. The following sections discuss the main research questions, implications for instruction, limitations and future work for each of these areas of research.

7.1 Case Study #1: Short Answers in Computer-Based Instruction

7.1.1 Research Questions

Chapter 1 presented a series of main research questions. Here, we briefly consider each of the research questions related to the study of short answer formats in the context of computer-based instruction and the key experimental results that addressed them.

1. Controlling for feedback and content, are short answer computer-based training formats more effective than multiple choice counterparts?

Overall, short answer and multiple choice formats were equally effective across the six experiments when subsequent student performance was assessed by multiple choice assess-
ments. Experiments on training students on the relationships of net force, acceleration and velocity in one dimension demonstrated similar and significant gains (approximately 20% pre-test to post-test). This finding was replicated with another set of physics concepts in Experiment #6, namely electric potential and electric field. Here, the training was highly successful (with an effect size of $d \approx 1.5$), but there were again no significant differences between question formats. Only one experiment out of the six found a marginally significant difference ($p = 0.056$) between short answer and multiple choice: Experiment #2, with a medium-sized effect ($d = 0.045$). As such, this hints that any underlying difference between formats on subsequent performance with a multiple choice assessment is likely to be at most a relatively small effect.

There is some evidence that the short answer format performed better than multiple choice in terms of student performance on subsequent assessment via short answer questions. In particular, Experiments #3 and #4 both found that students performed better on a short answer question in a similar format to those during training. However, there were no differences on other short answer assessment questions, including a transfer question that asked students to interpret motion from a graph.

2. Controlling for feedback and content, does short answer computer-based training provide retention gains compared to multiple choice?

Experiments #3 and #4 measured student retention of training on force and motion with a delay of approximately 1.5-2 months. For calculus-based students, the difference in pre-test to retention score was significant, suggesting meaningful retention of the training even at this timescale. Moreover, student retention scores were significantly higher than no-training control, suggesting that the retention performance was not primarily the result of additional course instruction or repeated testing effects. However, there was no significant difference in retention based on the format of the training question.

3. Does an increase in the interactivity of feedback during training improve the relative
performance of short answer formats?

Experiments #3 and #4 included two training conditions with more interactive feedback. These training conditions invoked follow-up questions and elicited specific counterexamples based on a students’ initial response. We found no significant effect of this increased interactivity on overall effectiveness of the training, nor an interaction with question format.

However, the addition of follow-up questions did hint at an important interaction with students’ attitudinal ratings of the training for students’ in the algebra-based course. Although student ratings of the usefulness of the training questions were already generally positive, there was a marginally significant interaction between the level of follow-up feedback, the question format, and students’ view of the training. In general, students tended to view single multiple choice questions as more useful than their short answer counterparts (though still positively). When follow-up questions were included in the training, short answer and multiple choice formats were rated as equally useful. This suggests that the increased interaction of the training helped students’ attitudinal perceptions for the short answer format, perhaps as a result of making the training seem more conversational.

4. *Does an increase in the variety of questions during training improve the relative performance of short answer formats?*

Experiment #5 compared student performance after training on the one-dimensional force and motion relationships via two sets of the same type of questions versus training sets composed of questions with varying structure. The increased question variety did not significantly improve student performance on the subsequent multiple choice assessment. Given that student gains on force and motion tended to plateau at around 20% from pre to post-assessment across all conditions (and all experiments) this may suggest an upper limit for this type of training in improving student conceptual performance on this topic, rather than conclusively argue that question variety is likely to have no impact on the relative
effectiveness of the two formats.

Still, despite this contextual constraint, there were noticeable interactions with how students interacted with the feedback in the training conditions. While student time on feedback in the same-question-type trainings tended to follow a clear exponential decay, there were noticeable spikes in student time with feedback when the questions varied. Moreover, this effect depended on question format, with students in the short answer question condition spending a significantly larger amount of time on the feedback than students in the multiple choice questions.

7.1.2 Instructional Implications

The main instructional takeaway of this study is that when content and feedback are controlled, multiple choice and short answer formats are similar in overall effectiveness for training simple concepts in physics. In essence then, at least for this level of training question and conceptual domain, the content of the question is the key. However, there are hints that if an instructional goal is to specifically improve student performance on subsequent student short answers – either from a desire to make sure that students can recall the concept on their own, or perhaps demonstrate correct usage of terminology – there may be situations where training via short answer formats such as those studied here can be beneficial.

Perhaps the most interesting finding in terms of instructional implication is that students tend to interact very differently with feedback, depending on the format of the question. Students are not only more likely to spend additional time on feedback with short answer questions in general, but particularly so at the start of the training and when the questions change in style. For those interested in designing instructional materials with a variety of questions and feedback, this may suggest placing short answer questions at the start of a training or homework set.
7.1.3 Limitations and Future Work

The most important restriction on our results is that all six experiments were conducted in a research setting. Future work implementing and investigating these types of questions in an authentic course setting is necessary. A second key limitation of this study is that it was limited to sets of highly restricted physics concepts - either one-dimensional force and motion or one-dimensional electric potential and field. It is not a given that training other physical concepts would necessarily arrive at the same result, given that this set of topics was specifically chosen based on potential issues of concept availability.

In addition, we only investigated one method for implementing a short answer format during training. We chose to use a hybrid approach where all initial questions were in short answer format and any subsequent clarifications were via multiple choice. In a similar vein, we only investigated one form of feedback - namely presentation of correct/incorrect along with knowledge of correct answer and explanation. This was done to minimize student frustration, opportunities for time-on-task to diverge unproductively (when students were unable to restate their answer in a satisfactory manner), and the potential for student gaming of the training task. Although our initial experiments found no correlation between the frequency of multiple choice clarifications and subsequent student performance when students were trained with single questions, it is not completely clear how our particular implementation may differ from other possible methods. Comparisons to other short answer implementations and levels of feedback might yield interesting results.

Our first two experiments suggest an alternative, but potentially valuable thread for future research. In particular, we identified clear differences in how students answered the short answer questions based on whether feedback (and assessment of correct/incorrect status) was provided to the student. With feedback, student answers became significantly shorter, with more precise terminology and repeated structure across similar questions (i.e. lists of physical cases). We hypothesized that this difference was not just a consequence of the content of the feedback, but the fact that the student was being assessed and adjusted their answers to maximize the chance that the computer would correctly interpret their
response. As such, this dimension of assessment may represent another axis to previous and ongoing research efforts on the optimum level of feedback for training. In order to separate out these effects, further research with different combinations of assessment, its indication to the student, and level of feedback would be needed.

In addition to the above, future work that attempts to parse interactions between training question difficulty (and therefore student success rates), time on feedback, and interplay with inherent complications of automatic assessment of short answer formats (misdetections, need for clarification, etc.) would be valuable. In particular, future work to look at the relationship between these factors within the context of more complicated training conditions (for example, with follow-up questions) could be particularly interesting. The results of such investigation may not only have implications for training, but potential extensions to natural language use in contexts like conversational computer tutors.

7.2 Case Study #2: Worked Examples & Synthesis Problem Solving

7.2.1 Research Questions

As above, we briefly consider the proposed research questions related to synthesis problem solving and the main experimental results addressing each of these questions.

1. Given the increased complexity of synthesis problems, is it more effective to invoke comparisons between worked examples that break down the target synthesis problem into its single-concept parts, or worked examples that include the concepts in combination?

This research question was addressed in Experiments 1 and 3. Students were given analogical comparison materials that either asked students to compare worked synthesis examples or single-concept examples, using the same prompts for comparison (with minor changes for line numbering). The context and surface features of the single-concept examples and synthesis examples were kept as similar as possible, with the single-concept solutions
representing portions of the full synthesis solution.

Neither experiment found a significant difference between control and treatments where analogical comparisons were invoked across single-concept examples. On the other hand, Experiment 1 found that comparison of worked synthesis examples was not only significantly better than control, it was also significantly better than the single-concept treatment, both in terms of overall score and the proportion of students recognizing and applying the correct concepts on the target synthesis problem. For the more complicated synthesis problem in Experiment 3, the comparison of worked examples using the same prompts in the single-concept condition was not significantly different than comparisons of single-concept examples - but a set of more scaffolded comparison prompts with synthesis worked examples did result in significant gains.

Taken together, these results support the importance of structural transfer. In particular, the difference in concept recognition in Experiment 1 is particularly telling. Whereas less than half (47%) of the students in the single-concept condition applied centripetal acceleration to the target problem, over 80% of the students did so in the synthesis worked example treatment. The fact that the proportion of students in the single-concept condition recognizing centripetal acceleration was not significantly different from control also speaks to how strongly biased students are to only apply a single physics concept to a given problem. Here, students were willing to invoke an entirely unstated assumption (that the velocity at the top of the loop had to be zero) in order to satisfy the constraints of their energy conservation equation.

It is also interesting that the comparison of single-concept worked examples did not improve its relative standing on the more complicated target problem. One potential reason for this is that although analogical comparison of synthesis worked examples seems to be incredibly powerful in terms of improving student recognition and application of component concepts, the technique seems less useful in improving student difficulties with the application of the individual concepts. In Experiment 2, where concept recognition was not a significant bottleneck, the analogical comparisons were relatively less effective at improving student performance.
2. How does the focus of the prompts influence analogical comparison (i.e. prompts involving holistic structure vs. prompts for fine-grain applications of the individual concepts)?

Different types of prompts - focused either on single-concept mastery or concept recognition - were investigated in Experiments 1 and 3. In Experiment 1, there was no significant difference in the effectiveness of prompts focused on single-concept mastery and prompts focused on overall concept recognition, either in terms of overall score on the target problem or the proportion of students recognizing the correct concepts. In Experiment 3, only a highly scaffolded set of comparison prompts - that included specific guidance on the inclusion of one of the component concepts - resulted in significant gains.

As a result, it seems that one of the important functions of the comparisons prompts is that they push the student to encode the structure of the base worked examples; provided that students are able to internalize the structure of the worked example, the specific nature of the prompts may be relatively less important. In Experiment 3, with the more complicated synthesis problem, the comparison prompts required additional scaffolding to meet this threshold.

3. How does analogical comparison across a pair of worked examples compare with self-explanation of each worked example independently?

One of the truly surprising – and ultimately interesting – results of this study is that unguided self-explanation dramatically outperformed all other treatment conditions in Experiment 3. In particular, whereas only 17% of students in the control condition applied the correct structure on the target synthesis problem, a shocking 80% applied the correct solution structure after self-explaining the provided worked examples, one at a time. For comparison, the best analogical comparison condition helped 50% of students to apply the correct structure. Although the difference wasn’t quite as large in Experiment 4, where students were relatively more removed from when the related material was presented in their
course, summarization via self-explanation once again outperformed analogical comparison alone.

Moreover, Experiment 4 provided evidence that this difference was not due to a simple failure of students to read through the worked examples when conducting analogical comparisons (in essence, only comparing the individual items highlighted by the questions one at a time). Prompting students to first read through and briefly annotate the worked solutions prior to their analogical comparison did not change subsequent student performance (despite adding a significant amount of time-on-task).

As such, it seems that the success of analogical comparison of the two worked examples was limited by cognitive load - invoking comparisons across two relatively complicated synthesis problems at once is a difficult and demanding task. On the other hand, an analysis of student summaries suggests that self-explanation may help students to understand the structure of each individual base example via a combination of filling-in-the gaps between important problem steps and potential mental model revision (i.e. recognizing that the induced emf and the battery emf can be combined). Naturally, this suggests that the two techniques might be complementary: inviting students to self-explain the individual worked examples first may help support further analogical comparison and extraction of important problem solving structure. Experiment 4 hinted that this was in fact the case, though the result is confounded by a corresponding increase in time-on-task.

7.2.2 Instructional Implications

The work presented here represents a novel approach to improving student performance on multi-concept physics problems (synthesis problems) using worked examples. Therefore, perhaps the most important instructional finding was that self-explanation and analogical comparison of worked examples can, in fact, be effective techniques for improving student problem solving performance even with these more complicated problems. Although the effectiveness of analogical comparison, in particular, is modulated by the complexity of the base and target problems, we have shown that analogical comparison can be further combined with self-explanation for additional improvements. As such, these activities could
potentially be used in a course recitation setting or in homework sets as a means to improve student practice with conceptual recognition, emphasize physical understanding, and de-emphasize equation hunting heuristics.

In addition to providing a proof of concept, there is another important instructional implication from our results. For two different sets of physics concepts, synthesis problems proved significantly more effective than comparisons of component single-concept questions and additional practice solving related single-concept problems. This result is particularly illuminating, given the implicit assumption that students can repeatedly practice physics concepts in isolation and expect success on problems that combine them. Our results suggest that this integration does not happen spontaneously, even when the component single-concept problems are presented sequentially for comparison. As such, in contrast to the vast majority of introductory homework problems and end-of-chapter exercises, our results imply that success with synthesis problems may best be facilitated by explicit practice with synthesis problems.

### 7.2.3 Limitations and Future Work

There are several important limitations to the work presented here. First, this series of experiments investigated synthesis problems involving only three different combinations of physics concepts. Further study with an increased variety of concepts and combinations is necessary. As a corollary, it is worthwhile to note that the target synthesis problems that responded effectively to analogical comparison and summarization all shared a subtle but important commonality: they were structured so that students can arrive at an answer, albeit incorrect, without considering at least one of the component concepts. In other words, students were not blocked from arriving at a solution merely from the mathematical structure of the involved physics equations. The intervention techniques may not prove as effective when students can successfully fall back on plug-and-chug problem solving heuristics.

A second key limitation of this series of studies is that it was not specifically designed to evaluate the exact mechanisms behind successful self-explanation or analogical comparison.
In particular, the success of self-explanation can be attributed either to the development of useful inferences and gap-filling steps within the worked examples, or mental model revision. Future work in distinguishing between these cases is merited. Depending on the results, those findings could also have profound implications for how to best structure analogical comparison prompts.

Another key limitation of this series of experiments is that they do not account for potential interactions with feedback during training. In particular, the effectiveness of analogical comparison might increase when students are provided immediate feedback on their comparisons, either through peer-mediated feedback in a group work setting or via individualized tutoring in computer-based instruction. In fact, this may suggest possible extensions for the work described in the previous volume of this thesis.

Finally, these experiments did not explicitly manipulate the presentation of the worked example prompts or solutions. It is possible that the addition of expert-like explanations or highlighted presentation of the worked examples may help reduce cognitive load, and subsequently improve the effectiveness of any invoked comparisons. To this, one possibility is to scaffold the presentation of the worked examples and their comparisons in the context of a computer-based tutorial. Unlike a traditional paper task, implementation in a computer-based tutorial would allow for portions of the problem statement and provided solution to be unveiled at different times to the student. For example, there may be additional benefits if students are first asked to compare, contrast, and predict whether a second physics problem required the same physics concepts as another provided example – before they are presented with the corresponding solutions. There may be similar opportunities for students to fill-in masked sections of an incomplete worked example based on the reference example. Analogical comparison prompts could then be included to scaffold this prediction and subsequent comparisons between the two full worked solutions along with corresponding feedback. As a result, such interventions may help students to more actively compare the structure of the worked examples – while also reducing the overall cognitive load in ways that the current paper-based implementation cannot. These considerations also represent potentially valuable avenues for further research.
BIBLIOGRAPHY


Gentner, D., Loewenstein, J., and Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95*(2), 393–405.


176


Appendix A

Computer-based Training and Test Materials

A.1 FVA Test Questions

The following question bank was developed and validated by Rosenblatt and Heckler (2011) and is included here for convenient reference.

1. At exactly 2:31PM, a boat is moving to the north on a lake. Which statement best describes the forces on the boat at this time?
   (a) there may be several forces to the north and to the south acting on the boat, but the forces to the north are larger.
   (b) there may be several forces to the north and to the south acting on the boat, but the forces to the south are larger.
   (c) there may be several forces to the north and to the south acting on the boat, but the forces to the south are equal in magnitude to those to the north.
   (d) both a and b are possible.
   (e) both a and c are possible.
   (f) a, b, and c are possible.

2. A car is moving to the right and speeding up. Which statement best describes the acceleration of the car at this instant?
   (a) the car's acceleration is to the right.
   (b) the car's acceleration is to the left.
   (c) the car's acceleration is zero.
   (d) both a and b are possible.
   (e) both a and c are possible.
   (f) a, b, and c are possible.

3. A student and a dog are playing tug of war with a rubber toy. If at a particular time the student is pulling on the toy to the right and the dog is pulling to the left with an equal force, which statement best describes the motion of the toy at this time?
   (a) it is moving toward the dog.
   (b) it is moving toward the student.
   (c) it is not moving.
   (d) both a and b are possible.
   (e) a, b, and c are possible.

4. A car is on a hill and the direction of its acceleration is uphill. Which statement best
describes the motion of the car at that time?
(a) it is moving uphill.
(b) it is to moving downhill.
(c) it is not moving.
(d) both a and b are possible.
(e) both a and c are possible.
(f) a, b, and c are possible.

5. A group of workers is pushing on a car in a driveway. There may be several forces on the car but those toward the street are greater. Which statement best describes the acceleration of the car at this instant?
(a) its acceleration is toward the street.
(b) its acceleration is away from the street.
(c) its acceleration is zero.
(d) both a and b are possible.
(e) both a and c are possible.
(f) a, b and c are possible.

6. A wagon is rolling east along the sidewalk. What can you say about the acceleration of the wagon at this time?
(a) it is accelerating east.
(b) it is accelerating west.
(c) it is not accelerating.
(d) both a and b are possible.
(e) both a and c are possible.
(f) a, b, and c are possible.

7. At exactly $t_0 = 3.0 \text{sec}$, a student is pulling with a force $F_{\text{student}}$ on a box which is connected to a spring, as in the diagram below. The spring is exerting a force $F_{\text{spring}}$ on the box. At this exact time, the box is moving toward the right and slowing down. Assuming the friction is negligible, which statement best describes the forces at this time?
(a) $F_{\text{student}} > F_{\text{spring}}$.
(b) $F_{\text{student}} < F_{\text{spring}}$.
(c) $F_{\text{student}} = F_{\text{spring}}$.
(d) both a and b are possible.
(e) both a and c are possible.
(f) both b and c are possible.
(g) a, b and c are possible.

8. The direction of acceleration of an object is to the right. What is the most you can say about the motion of the object at this time?
(a) it is moving to the right and its speed is increasing.
(b) it is moving to the right and its speed is decreasing.
(c) it is moving to the left.
(d) both a and b are possible.
(e) both a and c are possible.
(f) a, b, and c are possible.
9. As a wood block slides past a tick-mark on a smooth level surface, it has a velocity of 2 m/s to the right. There is a small amount of friction between the block and the surface, and eventually the block comes to rest. What are the forces acting on the block a few moments after it passes the mark?
(a) weight (down) and normal force (up)
(b) force of block (right) and friction (left)
(c) there are no forces on the block
(d) weight (down), normal force (up), and force of block (right)
(e) weight (down), normal force (up), and friction (left)
(f) weight (down), normal force (up), force of block (right), and friction (left)

10. A force sensor is attached inside a soccer ball that is used during a match. The force sensor measures the forces acting on the ball. At a randomly chosen instant during the game, the sensor detects that there is only one horizontal force on the ball, and that force is directed toward the home-team goal. Which statement best describes the motion of the ball at this instant?
(a) the ball is moving toward the home-team goal.
(b) the ball is moving away from the home-team goal.
(c) the ball is not moving.
(d) both a and b are possible.
(e) both a and c are possible.
(f) a, b, and c are possible.

11. At a particular instant of time during a kickball game, a ball on the playground is accelerating to the right. What can you say about the forces on the ball at this time?
(a) there may be several forces to the right and to the left acting on the ball, but the forces to the right are larger.
(b) there may be several forces to the right and to the left acting on the ball, but the forces to the left are larger.
(c) there may be several forces to the right and to the left acting on the ball, but the forces to the right are equal in magnitude to those to the left.
(d) both a and b are possible.
(e) both a and c are possible.
(f) a, b, and c are possible.

12. A boy rolls a ball toward the east on level ground into the wind. The ball rolls eastward against the wind and slows down, and after a short time the ball stops and rolls westward (with the wind) and starts to speed up. At the moment the ball turns around, the velocity is zero. Which statement best describes the forces on the ball at this moment?
(a) there may be several forces to the east and to the west acting on the ball, but the forces to the east are larger.
(b) there may be several forces to the east and to the west acting on the ball, but the forces to the west are larger.
(c) there may be several forces to the east and to the west acting on the ball, but the forces to the west are equal in magnitude to those to the east.
(d) both a and b are possible.
(e) both a and c are possible.
(f) both b and c are possible.
(g) a, b, and c are possible.
13. A block is attached between two springs as in the diagram below, and oscillates back and forth. At an instant of time depicted in the diagram, the acceleration of the block is to the left. Which statement best describes the motion of the block at this instant?
(a) it is moving left.
(b) it is moving right.
(c) it is not moving.
(d) both a and b are possible.
(e) both a and c are possible.
(f) both b and c are possible.
(g) a, b, and c are possible.

14. At a particular instant of time, there are several forces acting on an object in both the positive and negative direction, but the forces in the negative direction (to the left) are greater. Which statement best describes the motion of the object at this instant?
(a) it is moving to the right.
(b) it is moving to the left.
(c) it is not moving.
(d) both a and b are possible.
(e) both b and c are possible.
(f) a, b, and c are possible.

15. A child is playing with a toy car. At one instant, the acceleration of the toy car is to the north. Which statement best describes the toy car's motion?
(a) its speed is increasing.
(b) its speed is decreasing.
(c) both a and b are possible.

16. At exactly 10:02 A.M., a man is pushing to the right on a box with a force, \(F\). There is also a friction force \(f\) between the box and the floor. If at that exact moment, the box is moving to the right, which statement best describes the forces on the box at that time?
(a) \(F > f\).
(b) \(F < f\).
(c) \(F = f\).
(d) both a and b are possible.
(e) both a and c are possible.
(f) a, b and c are possible.

17. A train going from Columbus to Cleveland passes a telephone pole. What can you say about the acceleration of the train when it passes the pole?
(a) it is accelerating toward Cleveland.
(b) it is accelerating toward Columbus.
(c) it is not accelerating.
(d) both a and b are possible.
(e) both a and c are possible.
(f) a, b, and c are possible.
A.2 Experiment #1: Force & Motion Training Questions (Multiple Choice Version)

1. One of Ohio State’s football players is on the field during a game. At a particular instant, his acceleration is directed towards the offensive line. What do you know about the direction of his velocity at that instant?
   (a) His velocity is directed towards the offensive line.
   (b) His velocity is directed away from the offensive line.
   (c) His velocity is zero.
   (d) His velocity could be directed either towards or away from the offensive line.
   (e) His velocity could be directed towards the offensive line, or it could be zero.
   (f) His velocity could be directed towards the offensive line, away from the offensive line, or it could be zero.

2. In a laboratory experiment, a metal disk accelerates towards the right side of a table. What is the most you can say about the magnitude of the forces (i.e., forces directed to the right vs. forces directed to the left) on that object?
   (a) There may be several forces acting on the disk, but those to the right are greater.
   (b) There may be several forces acting on the disk, but those to the left are greater.
   (c) There may be several forces acting on the disk, but the net sum of the forces must be zero.
   (d) There may be several forces acting on the disk, but either those to the right are greater, or those to the left are greater.
   (e) There may be several forces acting on the disk, but either those to the right are greater, or the net sum of the forces must be zero.
   (f) There may be several forces acting on the disk, but those to the right could be greater, those to the left could be greater, or the sum of the forces must be zero.

3. The net force acting on a particle at time t points in the positive direction. What is the most you can say about the direction of the particle’s velocity at time t?
   (a) The particle’s velocity points in the positive direction.
   (b) The particle’s velocity points in the negative direction.
   (c) The particle’s velocity is zero.
   (d) The particle’s velocity could point towards either the positive or negative directions.
   (e) The particle’s velocity could point in the positive direction or it could be zero.
   (f) The particle’s velocity could point in either the positive direction, the negative direction, or it could be zero.

4. A hockey puck is moving rapidly across a frictionless surface towards the opponent’s net. What is the direction of the net force on the puck at that instant?
   (a) The net force is directed towards the opponent’s net.
   (b) The net force is directed away from the opponent’s net.
   (c) The net force is zero.
   (d) The net force could be directed towards the opponent’s net or it could be directed away from the opponent’s net.
   (e) The net force could be directed towards the opponent’s net or it could be zero.
   (f) The net force could be directed towards the opponent’s net, away from the opponent’s net, or it could be zero.

5. A turtle is moving east as it crosses a desert highway. Describe what is known about the direction of the acceleration of the turtle at that instant.
   (a) The acceleration of the turtle is zero.
   (b) The acceleration is directed in the direction the turtle is moving.
(c) The acceleration is directed opposite the direction the turtle is moving.
(d) The acceleration could be zero or directed in the direction the turtle is moving.
(e) The acceleration is either directed in the direction the turtle is moving, or opposite the direction.
(f) The acceleration could be zero, or it could be in the direction the turtle is moving, or it could be opposite the direction the turtle is moving.

6. A student is driving her car to school. At a particular instant, the net force acting on her car is pointing away from a stop sign. What can you say about the velocity of her car at that instant?
(a) The velocity of her car is directed towards the stop sign.
(b) The velocity of her car is directed away from the stop sign.
(c) The velocity of her car is zero.
(d) The velocity of her car could be directed towards or away from the stop sign.
(e) The velocity of her car could be directed towards the stop sign or zero.
(f) The velocity of her car could be directed towards or away from the stop sign, or it could be zero.

7. A dolphin is moving and has an acceleration directed downward towards the sea floor. What is the most you can say about its motion (what direction is the dolphin moving and how is its speed changing)?
(a) The dolphin is moving upward and its speed is decreasing.
(b) The dolphin is moving downward and its speed is decreasing.
(c) The dolphin is moving downward and its speed is increasing.
(d) The dolphin could either be moving upward with a decreasing speed, or it could be moving downward with a decreasing speed.
(e) The dolphin could either be moving upward with a decreasing speed, or it could be moving downward with an increasing speed.
(f) The dolphin could be moving upward or downward and its speed could be either increasing or decreasing.

8. There may be several forces acting on the space shuttle, but the forces directed towards the moon are greater. What is the most you can say about its acceleration?
(a) The acceleration of the space shuttle is directed towards the moon.
(b) The acceleration of the space shuttle is directed away from the moon.
(c) The acceleration of the space shuttle is zero.
(d) The acceleration of the space shuttle could be directed towards or away from the moon.
(e) The acceleration of the space shuttle could be directed towards the moon or it could be zero.
(f) The acceleration of the space shuttle could be directed towards the moon, away from the moon, or it could be zero.
9. A wooden block hangs on a spring suspended from a rigid support as shown. At a particular instant, the velocity of the block is directed downward. What is the most you can say about the acceleration of the block at that time?
(a) The acceleration of the block is directed downward.
(b) The acceleration of the block is directed upward.
(c) The acceleration of the block is zero.
(d) The acceleration of the block could either be directed downward or directed upward.
(e) The acceleration of the block could either be directed downward or zero.
(f) The acceleration of the block could be directed downward, directed upward, or zero.

10. A dog is enjoying a day at the park. At an instant in time the dog’s velocity is directed towards a small group of tulips. What is the direction of the net force on the dog at that instant?
(a) The net force on the dog is directed towards the tulips.
(b) The net force on the dog is directed away from the tulips.
(c) The net force on the dog is zero.
(d) The net force on the dog could be directed towards or away from the tulips.
(e) The net force on the dog could be directed towards the tulips or it could be zero.
(f) The net force on the dog could be directed towards the tulips, away from the tulips or it could be zero.

11. Wooden block 'A' lies on top of a rough surface, connected by a string to another block which has been hung from a pulley as shown in the diagram. There is an additional applied force on block A to the left as shown. At an instant in time, the sum of the forces on block A equals zero. What is the most you can say about the acceleration of block A at that instant?
(a) The acceleration of the block is directed towards the right.
(b) The acceleration of the block is directed towards the left.
(c) The acceleration of the block is zero.
(d) The acceleration of the block could be directed towards either the right or the left.
(e) The acceleration of the block could be directed towards the right or it could be zero.
(f) The acceleration of the block could be directed towards the right, the left, or it could be zero.

12. An object is moving in the positive direction and slowing down. What is the most you can say about the forces on that object?
(a) There may be several forces in the positive and negative directions, but the forces in the positive
(b) There may be several forces in the positive and negative directions, but the forces in the negative direction are greater.
(c) There may be several forces in the positive and negative directions, but the forces in the positive and negative directions are equal in magnitude.
(d) There may be several forces in the positive and negative directions, but the forces in the positive direction could be greater or the forces in the negative direction could be greater.
(e) There may be several forces in the positive and negative directions, but the forces in the positive direction could be greater or the forces in the positive and negative directions could be equal.
(f) There may be several forces in the positive and negative directions, but the forces in the positive direction could be greater, the forces in the negative direction could be greater, or the forces in the positive and negative directions could be equal.

A.3 Experiments #2-4: Force & Motion Training Questions (Multiple Choice Versions, Initial Question Only)

A.3.1 Training Questions

The following questions are presented in the order they were administered in Study #2. The secondary number (in parentheses) indicates the order used in Studies #3 and #4. Only the initial base questions are presented. Follow-up questions included in Studies #3 and #4 followed the structure described in the experimental design.

1 (1). One of Ohio State’s football players is on the field during a game. At a particular instant, his acceleration is directed towards the offensive line. What do you know about the direction of his velocity at that instant?
(a) His velocity is directed towards the offensive line.
(b) His velocity is directed away from the offensive line.
(c) His velocity is zero.
(d) His velocity could be directed either towards or away from the offensive line.
(e) His velocity could be directed towards the offensive line, or it could be zero.
(f) His velocity could be directed towards the offensive line, away from the offensive line, or it could be zero.

2 (2). Consider the following question and responses from three different students.
Question. A boy and a dog are playing tug of war with a length of rope. The boy is pulling on the rope to the left, while the dog pulls to the right. At an instant in time, the rope is moving to the left. How do the magnitudes of the forces on the rope compare at that instant in time?

Student A. Since the rope is moving to the left, the forces must be equal.
Student B. Since the object is moving to the left, the force from the boy must be greater than the force from the dog.
Student C. Not enough information is giving about what is meant by moving.

Which do you think is the better response? Select the best answer.

(a) Student A, because in tug of war there must be equal and opposite forces on the rope.
(b) Student B, because in order for the rope to move the force must be greater on one side and hence the force on the rope from the boy must be greater than the force from the dog.
(c) Student C, because although the rope is moving to the left at that instant, the student doesn’t
know if it is moving at a constant speed or accelerating in a certain direction.

3 (3). In a laboratory experiment, a metal disk accelerates towards the right side of a table. What is the most you can say about the magnitude of the forces (ie. forces directed to the right vs. forces directed to the left) on that object?
(a) There may be several forces acting on the disk, but those to the right are greater.
(b) There may be several forces acting on the disk, but those to the left are greater.
(c) There may be several forces acting on the disk, but the net sum of the forces must be zero.
(d) There may be several forces acting on the disk, but either those to the right are greater, or those to the left are greater.
(e) There may be several forces acting on the disk, but either those to the right are greater, or the net sum of the forces must be zero.
(f) There may be several forces acting on the disk, but those to the right could be greater, those to the left could be greater, or the some of the forces must be zero.

4 (4). The net force acting on a particle at time t points in the positive direction. What is the most you can say about the direction of the particle’s velocity at time t?
(a) The particle’s velocity points in the positive direction.
(b) The particle’s velocity points in the negative direction.
(c) The particle’s velocity is zero.
(d) The particle’s velocity could point towards either the positive or negative directions.
(e) The particle’s velocity could point in the positive direction or it could be zero.
(f) The particle’s velocity could point in either the positive direction, the negative direction, or it could be zero.

5 (7). Consider the statements of the two students below.
Student A. If there is a net force on an object, the object will move.
Student B. An object will continue to move in a straight line if no net force acts upon it.
Do the students each mean the same thing when they say ’move’? Select the best answer.
(a) Yes, they each mean that the position of the object changes with time.
(b) No, Student A means ’change from being stationary to starting to move’ which refers to Newton’s Second Law. Student B means ’keep in the status of moving’ which refers to Newton’s First Law.
(c) No, Student A means acceleration, Student B means a constant velocity.

6 (6). A hockey puck is moving rapidly across a frictionless surface towards the opponent’s net. What is the direction of the net force on the puck at that instant?
(a) The net force is directed towards the opponent’s net.
(b) The net force is directed away from the opponent’s net.
(c) The net force is zero.
(d) The net force could be directed towards the opponent’s net or it could be directed away from the opponent’s net.
(e) The net force could be directed towards the opponent’s net or it could be zero.
(f) The net force could be directed towards the opponent’s net, away from the opponent’s net, or it could be zero.

7. Can an object have a nonzero net force and not be moving? Select the best answer.
(a) No, this cannot be true. If there is a net force then the object is moving.
(b) No, by Newton’s Second Law, if there is a nonzero net force, there is a nonzero acceleration.
(c) Yes, but only at a single instant in time. At a particular instant in time, a nonzero net force does not require that the object be moving at all (that is, it could have zero instantaneous velocity).
The question above was used in Study #2. In Studies #3 and #4, the question below was used instead.

(5). In a laboratory experiment, a metal disk is subject to multiple forces. At a particular instant in time, the forces directed to the right are greater than those directed to the left. Select the best description of the motion of the object at that instant, given the above information.
(a) The disk is moving to the right.
(b) The disk is moving to the right and speeding up.
(c) The disk is accelerating to the right.
(d) The disk is accelerating to the left.
(e) The disk is moving to the left and slowing down.
(f) The disk is at rest but accelerating to the right.
(g) There is not enough information to describe anything about the motion.

8 (8). A wooden block hangs on a spring suspended from a rigid support as shown. At a particular instant, the velocity of the block is directed downward. What is the most you can say about the acceleration of the block at that time?
(a) The acceleration of the block is directed downward.
(b) The acceleration of the block is directed upward.
(c) The acceleration of the block is zero.
(d) The acceleration of the block could either be directed downward or directed upward.
(e) The acceleration of the block could either be directed downward or zero.
(f) The acceleration of the block could be directed downward, directed upward, or zero.

9 (9). Wooden block 'A' lies on top of a rough surface, connected by a string to another block which has been hung from a pulley as shown in the diagram. In addition, there is an applied force acting on block A to the left as shown. At an instant in time, the sum of the forces on block A equals zero. What is the most you can say about direction of the velocity of block A at that instant?
(a) The direction of the velocity of the block is to the right.
(b) The direction of the velocity of the block is to the left.
(c) The direction of the velocity of the block is zero.
(d) The direction of the velocity of the block could be to the right or to the left.
(e) The direction of the velocity of the block could be to the right or zero.
(f) The direction of the velocity of the block could be to the right, to the left, or zero.

10 (10). An object is moving in the positive direction and slowing down. What is the most you can
say about the forces on that object.

(a) There may be several forces in the positive and negative directions, but the forces in the positive direction are greater.
(b) There may be several forces in the positive and negative directions, but the forces in the negative direction are greater.
(c) There may be several forces in the positive and negative directions, but the forces in the positive and negative directions are equal in magnitude.
(d) There may be several forces in the positive and negative directions, but the forces in the positive direction could be greater or the forces in the negative direction could be greater.
(e) There may be several forces in the positive and negative directions, but the forces in the positive direction could be greater or the forces in the positive and negative directions could be equal.
(f) There may be several forces in the positive and negative directions, but the forces in the positive direction could be greater, the forces in the negative direction could be greater, or the forces in the positive and negative directions could be equal.

A.3.2 Short Answer Test Questions

1. In a laboratory experiment, a metal disk is subject to multiple forces. At a particular instant in time, the forces directed to the right are greater than those directed to the left. Describe the motion of the object at that instant.

2. Consider the statements of the two students below.
Student A. If there is a net force on an object, the object will move.
Student B. An object will continue to move in a straight line if no net force acts upon it.
If you were to improve their statements to be more precise, what would you change? Explain your reasoning.

3. The graph at the right shows the net force acting on a 1.0 kg object over a period of time. The positive direction is defined to be to the right. The initial velocity of the object at time \( t=0 \) was 0 m/s. Describe the motion of the object at time \( t=10s \).

A.4 Experiment #5: Training Questions (Varied-type)

A.4.1 First session training questions

1. One of Ohio State's football players is on the field during a game. At a particular instant, his acceleration is directed towards the offensive line. What, if anything, do you know about the direction of his velocity?

2. The net force acting on a particle at time \( t \) points in the positive direction. What is the most you can say about the direction of the particle's velocity at time \( t \)?
3. In a laboratory experiment, a metal disk accelerates towards the right side of a table. What is the most you can say about the direction of the net force on the disk at that instant?

4. A rocket ship is initially traveling towards a distant star. Over the course of three days, its velocity changes at the same steady rate. At a point in time on day two, the rocket ship is momentarily at rest and then begins to move away from the star with an increasing speed. What can you say about the magnitude and direction of the net force acting on the rocket ship over the three day time period?

5. At time t=0 a cart is at rest on a frictionless surface. From t=0s to t = 10s, a constant force is applied toward the right on the cart. Qualitatively describe the motion of the cart during that interval.

6. At time t = 0, a lab cart is sliding across a frictionless rail as shown with an applied force of 2N in the direction of motion. At t=10s, the applied force is suddenly removed (the net force becomes zero and remains zero). Qualitatively describe the motion of the cart after the applied force is removed.

7. Consider the following question and responses from three different students.
   Question. A boy and a dog are playing tug of war with a length of rope. The boy is pulling on the rope to the left, while the dog pulls to the right. At an instant in time, the rope is moving to the left. How do the magnitudes of the forces on the rope compare at that instant in time?
   Student X. Since the rope is moving to the left, it must be speeding up.
   Student Y. Since the rope is moving to the left, the force from the boy must be greater than the force from the dog.
   Student Z. Not enough information is given about how the object is moving. Which do you think is the better response? Explain your reasoning.

8. Consider the statements of the two students below.
   Student Y. If there is a net force on an object, the object will move.
   Student Z. An object will move in a straight line if no net force acts upon it.
   Do the students each mean the same thing when they say 'move'? Explain your reasoning.

9. The velocity and acceleration of an object are shown at a specific time t. At that instant, what can you say about the direction of the net force on the object?

10. The velocity and the net force acting on an object are shown at a specific time t. At that instant, what can you say about the direction of the acceleration of the object?
11. The velocity and the net force acting on an object are shown at a specific time \( t \). At that instant, what can you say about the direction of the acceleration of the object?

12. A student is standing on a scale in an elevator as shown. The only forces acting on the student are gravity and the normal force from the scale. (A scale measures the magnitude of the normal force). When the elevator remains at rest for a period of time, what must be true about the normal force reported by the scale, and the force of gravity on the student? Explain your reasoning.

13. A student is standing on a scale in an elevator as shown. The only forces acting on the student are gravity and the normal force from the scale. (A scale measures the magnitude of the normal force). The elevator makes multiple trips up and down. During a random period of time, a student finds that the upward normal force from the scale and the force of gravity on the student are equal over that period. Do they know for sure they are at rest during that time interval? Explain your reasoning.

14. A student is standing on a scale in an elevator as shown. The only forces acting on the student are gravity and the normal force from the scale. (A scale measures the magnitude of the normal force). The elevator makes multiple trips up and down. At a random point in time, the student notices that the upward normal force is larger than the force of gravity on the student. Given only this information, can the student determine which direction the elevator is moving at that instant? State Yes or No, and explain your reasoning.

**A.4.2 Second session training questions**

1. An astronaut on the international space station uses a steel ball and a can of compressed air as part of a demonstration. The can of compressed air is used to provide a force on the ball. Assume the force from the compressed air is strong enough that ambient air resistance can be ignored. Initially, the steel ball is moving towards the right. When the astronaut activates the compressed air, the speed of the steel ball slows down at a steady rate until the ball reaches its furthest point, momentarily stops, and then speeds up to the left at the same steady rate. What can you say
about the magnitude and direction of the force due to the compressed air acting on the steel ball throughout the entire period?

2. You are recording the motion of a distant jet-powered satellite. For the first 2 days, you notice that the object speeds up in a straight line towards the west until it reaches a maximum speed for the rest of the observed time. What can you conclude about the net force on that object after the object has reached its maximum speed?

3. At time t=0, a hockey puck is moving towards the right. An accelerometer measures the puck to have a constant acceleration directed toward the left. Assuming the acceleration remains constant for a very long period of time, qualitatively describe the subsequent motion. Your answer should explicitly describe how the direction and magnitude of the puck's velocity change over time.

4. A student is standing on a scale in an elevator as shown. The only forces acting on the student are gravity and the normal force from the scale. (A scale measures the magnitude of the normal force). The elevator makes multiple trips up and down. At a random instant, the student notices that the scale reads a normal force smaller than the force of gravity on the student. Given only this information, can the student determine which direction the elevator is moving at that instant? State Yes or No, and explain your reasoning.

5. A student is standing on a scale in an elevator as shown. The only forces acting on the student are gravity and the normal force from the scale. (A scale measures the magnitude of the normal force). The elevator makes multiple trips up and down. At a random instant, the student notices that the scale reads a normal force larger than the force of gravity on the student, and they can tell from looking through a small window that they are moving down at that instant. Is the elevator speeding up, slowing down, or moving at a constant speed? Explain your reasoning.

6. A hockey puck is moving rapidly across a frictionless surface towards the opponent’s net. What is the direction of the net force on the puck at that instant?

7. Wooden block 'A' lies on top of a rough surface, connected by a string to another block hung from a pulley as shown in the diagram. There is an applied force acting on block A to the left as shown. At an instant in time, the sum of all forces on block A equals zero. What is the most you can say about the direction of the velocity of block A at that instant?

8. At exactly 3:04 PM, the net force on an automobile is directed towards an intersection. What, if anything, do you know about the direction of the automobile’s acceleration at that moment?
9. An object is moving in the positive direction and slowing down. What is the most you can say about the direction of the net force on that object at that instant?

10. As part of a homework problem, a student needed to analyze the motion of a car, where the acceleration of the car at an instant in time was labeled as shown. Given the diagram, the student remarked:
   "The car is moving to the right and speeding up."
   Do you agree or disagree with this student? Explain your reasoning.

11. Consider the following statements from two different students.
   Student Y. A net force (with magnitude greater than zero) causes motion.
   Student Z. A net force (with magnitude greater than zero) causes change in motion.
   Which do you think is the better statement? Explain your reasoning.
A.5 Experiment #6: Voltage & Potential Questions

A.5.1 Training Questions

1. Consider the graph of electric potential shown to the right. At which point is the electric field greater? Justify your answer.
   (a) Point A, because the potential is less at A than at B.
   (b) Point B, because the potential is greater at B than at A.
   (c) Point A, because the slope of the potential graph at point A is less than the slope at B.
   (d) Point B, because the slope of the potential graph at point B is greater than the slope at A.

2. Consider the graph of electric potential shown. Define the positive x-direction to be to the right. What is the direction of the electric field at point B?
   (a) Left
   (b) Right

3. Two equal positive charges are located a fixed distance apart as shown in the figure. What do you know about the electric field at point P?
   (a) The electric field is to the right.
   (b) The electric field is to the left.
   (c) The electric field is positive.
   (d) The electric field is negative.
   (e) The electric field is zero.
4. Two equal positive charges are located a fixed distance apart as shown in the figure. What do you know about the electric potential at point P?
(a) The electric potential is to the right.
(b) The electric potential is to the left.
(c) The electric potential is positive.
(d) The electric potential is negative.
(e) The electric potential is zero.

5. The electric potential is nonzero and positive at a given point. Based on this information, what do you know about the electric field at that point? Explain your reasoning.
(a) The electric field is oriented in the positive direction, since the potential is positive.
(b) The electric field is oriented in the negative direction, since the potential is positive.
(c) The electric field is nonzero, since the potential is nonzero.
(d) There is not enough information since we need to know the rate of change of the electric potential at that point.

6. The diagram shows a pair of very large parallel plates, which are kept at the electric potentials shown. At which point is the magnitude of the electric field greater? Justify your answer.
(a) A, because it is closer to the higher potential.
(b) B, because it is farther from the higher potential.
(c) The electric field is the same at points A and B.

7. Consider the discussion between three students below.
Student A. To calculate unknown electric fields or potentials use $\Delta V = -Ed$ because it is the easiest to manipulate.
Student B. You can’t use that. Always use $E = -\frac{dV}{dx}$ or equivalently $\Delta V = -\int Edx$.
Student C. I’m pretty sure the first formula is a special case of the other two.
Explain why Student C is correct. Under what circumstances are the methods the same?
8. Consider the graph of electric potential shown to the right. At which point is the electric field greater? Justify your answer.
(a) Point A, because the slope of the potential graph at point A is less than the slope at B.
(b) Point B, because the slope of the potential graph at point B is greater than the slope at A.
(c) They are the same, because the potential is the same at both A and B.

9. Consider the graph of electric potential shown to the right. Which point(s) have an electric field oriented in the positive direction?
(a) A
(b) B
(c) C
(d) D
(e) A and B
(f) C and D

10. Two equal but opposite charges are located a fixed distance apart as shown in the figure. What do you know about the electric field at point P?
(a) The electric field is to the right.
(b) The electric field is to the left.
(c) The electric field is positive.
(d) The electric field is negative.
(e) The electric field is zero.

11. Two equal but opposite charges are located a fixed distance apart as shown in the figure. What do you know about the electric potential at point P?
(a) The electric potential is to the right.
(b) The electric potential is to the left.
(c) The electric potential is positive.
(d) The electric potential is negative.
(e) The electric potential is zero.
12. The diagram shows two pairs of very large parallel plates, which are kept at the electric potentials shown. At which point is the magnitude of the electric field greater? Justify your answer.
(a) A, because the plates are closer together.
(b) B, because the plates are farther apart.
(c) The electric field is the same at points A and B, because the difference in potential is the same.

A.5.2 Test Questions

1. Consider the electric potential graph shown to the right. At which point is the electric field greater?
(a) A
(b) B
(c) The electric field is the same at A and B.

2. Consider the electric potential graph shown to the right. At which point(s) is there an electric field oriented in the -x direction?
(a) A
(b) B
(c) C
(d) D
(e) A and B
(f) C and D
3. Consider the electric potential graph shown to the right. At which point(s) is there no electric field?
   (a) A  
   (b) B  
   (c) C  
   (d) A and C

4. Consider the electric potential graph shown to the right. At which point is the electric field greater?
   (a) A  
   (b) B  
   (c) The electric field is the same at A and B.

5. Two equal negative charges are located a fixed distance apart as shown in the figure. What do you know about the electric field at point P?
   (a) The electric field is to the right.  
   (b) The electric field is to the left.  
   (c) The electric field is positive.  
   (d) The electric field is negative.  
   (e) The electric field is zero.

6. Two equal negative charges are located a fixed distance apart as shown in the figure. What do you know about the electric potential at point P?
   (a) The electric potential is to the right.  
   (b) The electric potential is to the left.  
   (c) The electric potential is positive.  
   (d) The electric potential is negative.  
   (e) The electric potential is zero.
7. Two equal but opposite charges are located a fixed distance apart as shown in the figure. What do you know about the electric field at point P?
(a) The electric field is to the right.
(b) The electric field is to the left.
(c) The electric field is positive.
(d) The electric field is negative.
(e) The electric field is zero.

8. Two equal but opposite charges are located a fixed distance apart as shown in the figure. What do you know about the electric potential at point P?
(a) The electric potential is to the right.
(b) The electric potential is to the left.
(c) The electric potential is positive.
(d) The electric potential is negative.
(e) The electric potential is zero.

9. The diagram shows a pair of very large parallel plates, which are held at the electric potentials shown. How does the magnitude of the electric field compare at the three points?
(a) A > B > C
(b) B > A > C
(c) C > B > A
(d) A = C > B
(e) B > A = C
(f) A = B = C

10. The diagram shows a pair of very large parallel plates, which are held at the electric potentials shown. How does the electric potential compare at the three points?
(a) A > B > C
(b) B > A > C
(c) C > B > A
(d) A = C > B
(e) B > A = C
(f) A = B = C
11. The diagram shows two sets of very large parallel plates, which are held at the electric potentials shown. At which point is the magnitude of the electric field greater?
(a) A  
(b) B  
(c) The electric field is the same at A and B  
(d) The electric field is zero at both A and B

12. The diagram shows two sets of very large parallel plates, which are held at the electric potentials shown. At which point is the magnitude of the electric field greater?
(a) A  
(b) B  
(c) The electric field is the same at A and B  
(d) The electric field is zero at both A and B

13 (Short Answer). There is a region of space where the electric potential only varies based on position along one direction; in that region, the electric potential increases linearly in the positive x-direction. Based on that information, what do you know about the electric field in that region? Explain your reasoning.

14 (Short Answer). A student reads the following true statement in a physics textbook: The electric field inside a hollow, uniformly charged sphere is zero. The student reasons: If the electric field inside a uniform charged sphere is zero, the electric potential is also zero inside the sphere. Do you agree or disagree with this student? Explain your reasoning.

15 (Short Answer). The electric potential is zero at a point. What do you know about the electric field at that point? Justify your answer.
Appendix B

Synthesis Problem Solving Materials

For each of the subsequent experiments, worked examples were either coupled with analogical comparison prompts or a prompt to explain the provided worked example to a friend. The worked examples and sets of comparison prompts are included for reference.

The following instruction was provided to all of the analogical comparison conditions:

Consider the two pairs of problems and correct student solutions below. You will be asked to compare and evaluate the similarities and differences between these problems and the corresponding student solutions.

The following instructions were provided to conditions based on summarization:

Consider the problem and correct student solution below. On the next page, you will be asked to create a summary explaining how to solve this problem to a friend.

Take time to study the problem so that you fully understand it. Then, write a guide summarizing the important information necessary for a friend in an introductory physics class to fully understand this problem. You can assume your friend has a copy of the solution, so your guide can refer to elements of the problem, figure and line numbers within the solution.
B.1 Experiment #1 Training Materials

Synthesis Worked Examples

Problem 1. A block with mass \( M = 2.0 \) kg slides from a horizontal surface into a vertical circular track with a radius of \( R = 3.0 \) m. Assume that friction between the block and the track is negligible. What is the minimum speed the block must have at the bottom of the loop that will permit it to slide all the way around the circular track without leaving the track at the top?

![Diagram of block sliding on a circular track]

Student Solution 1

At the top of the loop:

\[
\begin{align*}
F_N & \rightarrow \ F_g \\
F_{\text{net}} & = F_g + F_N = ma_c \\
F_{\text{net}} & = mg + 0 = \frac{mv^2}{R} \\
g & = \frac{v^2}{R} \\
v & = \sqrt{gR} \\
\frac{1}{2}mv_f^2 & = \frac{1}{2}mv^2 + mgh \\
\frac{1}{2}mv_f^2 & = \frac{1}{2}m(gR) + mg(2R) \\
v_f^2 & = gR + 4gR \\
v_f & = \sqrt{5gR} = \sqrt{5\left(9.81 \, \text{m/s}^2\right)(3.0 \, \text{m})} \\
v_f & = 12.1 \, \text{m/s} \\
\end{align*}
\]

Problem 2. You are tasked with designing a rollercoaster. As part of the design, the track gradually descends until it comes to a small semi-circular hill of height \( R \).

You know that the speed of the rollercoaster cart will be 18.5 m/s when the cart is 20 m above the height of the oncoming hill. What is the minimum possible hill height for which the cart does not leave the surface of the track at the top? Ignore friction.

![Diagram of rollercoaster track]

Student Solution 2

\[
\begin{align*}
\frac{1}{2}mv_f^2 + mgh & = \frac{1}{2}mv_f^2 \\
v_f^2 & = v_i^2 + 2gh \\
v_f & = \sqrt{\left(18.5 \, \text{m/s}\right)^2 + 2\left(9.81 \, \text{m/s}^2\right)(20 \, \text{m})} \\
v_f & = 27.1 \, \text{m/s} \\
\end{align*}
\]

At the top of the hill:

\[
\begin{align*}
F_N & \rightarrow \ F_g \\
F_{\text{net}} & = F_g - F_N = -ma_c \\
F_{\text{net}} & = 0 - mg = -\frac{mv_f^2}{R} \\
v_f^2 & = gR = \frac{(27.1 \, \text{m/s})^2}{9.81 \, \text{m/s}^2} \\
R & = 75 \, \text{m} \\
\end{align*}
\]

---

Figure B.1: Experiment #1: Synthesis worked examples

204
### Single-Concept Worked Examples

**Problem 1.** A block with mass $M = 2.0 \text{ kg}$ slides around a vertical circular track with a radius of $R = 3.0 \text{ m}$. Assume that friction between the block and the track is negligible. What is the minimum speed the block must have at the top of the track in order to ensure that it does not leave the track at the top?

![Figure of a block sliding on a circular track](image)

**Student Solution 1**

At the top of the loop:

\[
\begin{align*}
F_N & \leq mg \quad \text{(1)} \\
F_g & = mg \quad \text{(2)} \\
F_{net} & = F_g + F_N = ma_c \quad \text{(3)} \\
F_{net} & = mg + 0 = \frac{mv^2}{R} \quad \text{(4)} \\
v & = \sqrt{\frac{9.8 \text{ m/s}^2 \times 3.0 \text{ m}}{}} \quad \text{(5)} \\
v_b & = 5.4 \text{ m/s} \quad \text{(6)}
\end{align*}
\]

**Problem 2.** You are tasked with designing a rollercoaster. As part of the design, the track includes a semi-circular hill of height $R$. You know that the speed of the rollercoaster cart will be $27 \text{ m/s}$ at the top of the hill. What is the minimum possible hill height for which the cart does not leave the surface of the track at the top? Ignore friction.

![Figure of a rollercoaster track](image)

**Student Solution 2**

At the top of the hill:

\[
\begin{align*}
F_N & \leq mg \quad \text{(1)} \\
F_g & = mg \quad \text{(2)} \\
F_{net} & = F_N - F_g = -ma_c \quad \text{(3)} \\
F_{net} & = 0 - mg = -\frac{mv^2}{R} \quad \text{(4)} \\
v & = \sqrt{\frac{9.8 \text{ m/s}^2 \times 27 \text{ m}}{}} \quad \text{(5)} \\
R & = 74 \text{ m} \quad \text{(6)}
\end{align*}
\]

**Problem 1.** A block with mass $M = 1.5 \text{ kg}$ slides from a horizontal surface into a circular track with a radius of $R = 2.0 \text{ m}$. Assume that friction between the block and the track is negligible. If the speed of the block at the top of the track is $3.0 \text{ m/s}$, what was the speed of the block at the bottom before entering the loop?

![Figure of a block sliding on a circular track](image)

**Student Solution 1**

\[
\begin{align*}
\frac{1}{2}mv_b^2 & = \frac{1}{2}mv_r^2 + mgh \quad \text{(1)} \\
\frac{1}{2}mv_r^2 & = \frac{1}{2}mv_b^2 + mg(2R) \quad \text{(2)} \\
v_r & = v_b + 4gR \quad \text{(3)} \\
v_b & = \sqrt{\left(3.0 \text{ m/s}^2\right)^2 + 4 \times 9.81 \text{ m/s}^2 \times 2.0 \text{ m}} \quad \text{(4)} \\
v_b & = 9.4 \text{ m/s} \quad \text{(5)}
\end{align*}
\]

**Problem 2.** You are tasked with designing a rollercoaster. As part of the design, the track gradually descends until it comes to a small semi-circular hill of height $R$. You know that the speed of the rollercoaster cart will be $15 \text{ m/s}$ when the cart is $20 \text{ m}$ above the height of the oncoming hill. What is the velocity at the top of the oncoming hill?

![Figure of a rollercoaster track](image)

**Student Solution 2**

\[
\begin{align*}
\frac{1}{2}mv_r^2 + mgh & = \frac{1}{2}mv_f^2 \quad \text{(1)} \\
v_f^2 & = v_r^2 + 2gh \quad \text{(2)} \\
v_f & = \sqrt{\left(15 \text{ m/s}^2\right)^2 + 2 \times 9.81 \text{ m/s}^2 \times 20 \text{ m}} \quad \text{(3)} \\
v_f & = 42.8 \text{ m/s} \quad \text{(4)}
\end{align*}
\]
# Mastery Prompts (Synthesis Version)

Each of the following questions refers to the two different problem-solution pairs on page 1.

A) Consider the diagram in line 1 of Solution 1 and line 6 of Solution 2. Explain any similarities and differences between the identified forces in terms of the two physical situations.

B) Consider line 3 of Solution 1 and line 8 of Solution 2. Is the direction of the $\frac{mv^2}{R}$ term the same or different in the two solutions? Explain your reasoning.

C) In Problem 1 line 3, the student substitutes in zero for $F_N$. Explain why.
   
   Is the situation in Problem 2 similar or different to Problem 1? Explain your reasoning.

D) Consider line 6 in Solution 1 and line 1 in Solution 2.
   
   Identify the sources of energy considered in the two solutions.
   
   Why is the term $mgh$ on the left side of Solution 2, but on the right side in Solution 1? Where did the student in Solution 1 assign $y = 0$? Explain your reasoning.

E) A friend in an introductory physics class at another institution asks for help in understanding and tackling similar problems. What will you tell your friend so that s/he can fully understand the problem?

## Recognition Prompts

Each of the following questions refers to the two different problem-solution pairs on page 1.

A) What are the main physical concept(s) used in both of the students’ solutions?

B) For Problem 1, identify the elements of the problem statement or diagram which indicate the need to use each of the physical concept(s) you mentioned in Part A. List specific elements and explain your reasoning.

C) For Problem 2, identify the elements of the problem statement or diagram which indicate the need to use each of the physical concept(s) you mentioned in Part A. List specific elements and explain your reasoning.

D) Are the elements you identified above similar between the two problems? Explain your answer highlighting any differences and similarities.

E) Another student compared these two problems and solutions, and remarked: “I get why the initial and final velocity in Problem 2 are both non-zero, but I don’t understand why the final velocity in Problem 1 must be non-zero.”
   
   i) Using the physical concept(s) you identified in Part A, what would you say to help this student?
   
   ii) What elements of Problem 1 indicate that the final velocity should be non-zero?

F) A friend in an introductory physics class at another institution asks for help in understanding and tackling similar problems. What will you tell your friend so that s/he can fully understand the problem?

---

Figure B.3: Experiment #1: Mastery & recognition prompts
Figure B.4: Experiment #2: Synthesis worked examples - same variable version
Figure B.5: Experiment #2: Synthesis worked examples - switched variable version

Problem I. Consider the circuit shown below. Let $r_1 = 0.50$ m, $r_2 = 0.75$ m, and $r_3 = 1.25$ m. Given that the magnetic field at point $B$ is $4.0 \times 10^{-7}$ T into the page, find the voltage of the battery.

**Student Solution I**

1. Define $I_1$ to be current through arc of radius $r_1$, $I_2$ current through arc of radius $r_2$, and $I_3$ current through arc of radius $r_3$.

\[ I_1 = \frac{3}{4} \left( \frac{H_1}{2r_1} \right) \text{ into page} \]

\[ I_2 = \frac{1}{2} \left( \frac{H_2}{2r_2} \right) \text{ into page} \]

\[ I_3 = \frac{1}{4} \left( \frac{H_3}{2r_3} \right) \text{ into page} \]

\[ I_{total} = \frac{3H_1}{8r_1} + \frac{H_2}{4r_2} \text{ into page} \]

\[ \frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} = \frac{1}{20\,\Omega} + \frac{1}{30\,\Omega} \]

\[ R_{eq} = 12\,\Omega \]

\[ R_{total} = R_1 + R_{eq} = 10\,\Omega + 12\,\Omega = 22\,\Omega \]

\[ I_1 + I_2 = I_3 \]

\[ I_1R_1 = I_3R_3 \]

From lines (11) and (12):

\[ I_1 = \left( 1 + \frac{2}{5} \right)I_3 \]

\[ I_2 = \frac{3}{5}I_3 \]

\[ I_3 = \frac{V}{R_{total}} \]

\[ \frac{\mathbf{B}_{nulld}}{\mathbf{B}_{total}} = \frac{\frac{3}{8} \frac{3}{5} + \frac{2}{5} \frac{2}{5}}{\frac{3}{8} \frac{3}{5} + \frac{2}{5} \frac{2}{5}} \]

\[ V = \frac{8R_{total}}{3r_1 + 2r_2 + 2r_3} \mathbf{B}_{total} \]

\[ V = \left( 4\pi \times 10^{-7} \right) \left( 0.50 \text{ m} + 0.75 \text{ m} + 1.25 \text{ m} \right) \]

\[ V = 7.9V \]

Problem II. Consider the circuit shown below. You may assume that the horizontal wires are very long. Given that the magnetic field at point $B$ is $(1.5 \times 10^{-4})$ T, $I_1 = 0.25$ m, $I_2 = 0.5$ m, and $I_3 = 2.0$ m, find the voltage of the battery.

**Student Solution II**

\[ R_{eq} = R_1 + R_2 = 5\,\Omega + 15\,\Omega = 20\,\Omega \]

\[ \frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} = \frac{1}{10\,\Omega} + \frac{1}{20\,\Omega} \]

\[ R_{eq} = 6.7\,\Omega \]

\[ I_{total} = \frac{V}{R_{eq}} \]

\[ I_{load} = \frac{V}{R_3} \]

\[ I_{bottom} = \frac{V}{R_3} \]

From lines (6) and (7):

\[ \mathbf{B}_{load} = \frac{\mu_0 I_{load}}{2\pi r_1} \text{ into page} \]

\[ \mathbf{B}_{bottom} = \frac{\mu_0 I_{bottom}}{2\pi r_1} \text{ into page} \]

\[ \mathbf{B}_{total} = \mathbf{B}_{load} + \mathbf{B}_{bottom} \text{ into page} \]

\[ \mathbf{B}_{total} = \frac{\mu_0 I}{r_1} + \frac{\mu_0 I}{r_2} + \frac{\mu_0 I}{r_3} \text{ into page} \]

\[ V = \frac{2\pi}{4\pi} \left( \frac{3}{r_1} \frac{3}{r_2} + \frac{2}{r_3} \frac{2}{r_4} \right) \]

\[ V = \frac{(1.5 \times 10^{-4})}{1} \frac{(0.25 \text{ m})(0.75 \text{ m})}{1} \frac{(1.25 \text{ m})}{1} \]

\[ V = 9.4V \]
Problem 1. Consider the circuit shown below. Given that the battery provides 9.0 V, find the currents through $R_2$ and $R_3$.

**Student Solution 1**

\[
\begin{align*}
\frac{1}{R_{eq}} & = \frac{1}{R_2} + \frac{1}{R_3} + \frac{1}{50 \, \Omega} \\
R_{eq} & = 12 \, \Omega \\
R_{total} & = R_1 + R_{eq} = 10 \, \Omega + 12 \, \Omega = 22 \, \Omega \\

\text{Define } I_1 \text{ to be current through } R_1, \\
\text{I}_2 \text{ current through } R_2, \text{I}_3 \text{ current through } R_3 \\
V = I_1 R_{total} \\
I_1 = \frac{9.0 \, V}{22 \, \Omega} = 0.41 \, A \\
I_2 + I_3 = I_1 \\
I_2 R_2 = I_2 R_3 \\
\text{From lines (7) and (8): } I_1 = \left(1 + \frac{R_1}{R_2}\right)I_2 \\
I_2 = \frac{0.41 \, A}{1 + \frac{22 \, \Omega}{50 \, \Omega}} = 0.25 \, A \\
I_2 = I_1 - I_3 = 0.41 \, A - 0.25 \, A = 0.16 \, A
\end{align*}
\]

Problem 2. Consider the circuit shown below. Given that the battery provides 9.0 V, find the currents through $R_1$ and $R_2$.

**Student Solution 2**

\[
\begin{align*}
\frac{1}{R_{eq}} & = \frac{1}{R_2} + \frac{1}{R_3} + \frac{1}{15 \, \Omega} \\
R_{eq} & = 25 \, \Omega \\
R_{total} & = R_1 + R_{eq} = 5 \, \Omega + 25 \, \Omega = 30 \, \Omega \\

\text{Define } I_1 \text{ to be current through } R_1, \\
\text{I}_2 \text{ current through } R_2 \\
V = I_1 R_1 \\
I_1 = \frac{9.0 \, V}{15 \, \Omega} = 0.6 \, A \\
I_2 = \frac{V}{I_2 R_{eq}} = \frac{9.0 \, V}{I_2 \cdot 25 \, \Omega} = 0.36 \, A
\end{align*}
\]
Problem 1. Consider the current carrying wire shown below. Given \( r_1 = 0.5 \text{m}, r_2 = 2.0 \text{m}, \) and that the magnetic field is \( 2.0 \times 10^{-7} \text{T} \) out of page, find the current through the wire.

**Student Solution 1**

Define \( \vec{B}_1 \) due to circular arc of radius \( r_1 \) and \( \vec{B}_2 \) due to circular arc of radius \( r_2 \):

\[
\vec{B}_1 = \frac{1}{2} \left( \frac{\mu_0 I_1}{2r_1} \right) \quad \text{out of page (1)}
\]

\[
\vec{B}_2 = \frac{1}{2} \left( \frac{\mu_0 I_2}{2r_2} \right) \quad \text{out of page (2)}
\]

\[
\vec{B}_{\text{total}} = \vec{B}_1 + \vec{B}_2 \quad \text{(3)}
\]

\[
\vec{B}_{\text{total}} = \frac{\mu_0}{4} \left( \frac{1}{r_1} + \frac{1}{r_2} \right) \quad \text{(4)}
\]

\[
I = \frac{4}{4\pi \times 10^{-7} \text{T} \cdot \text{m}} \left( \frac{2.0 \times 10^{-7} \text{T}}{0.5 \text{m}} + \frac{1 \times 10^{-7} \text{T}}{2.0 \text{m}} \right) \quad \text{(5)}
\]

\[
I = 0.25 \text{A} \quad \text{(6)}
\]

Problem 2. Consider the long, current carrying wires shown below. The magnetic field at point P due to all three wires is \( 7 \times 10^{-7} \text{T} \) into the page. Given that \( I_1 = 0.25 \text{A}, I_2 = 0.5 \text{A}, \) and \( I_3 = 2.0 \text{A}, \) find the magnitude of the unknown current \( I_4. \)

**Student Solution 2**

\[
\vec{B}_1 = \frac{\mu_0 I_1}{2r_1} \quad \text{into page (1)}
\]

\[
\vec{B}_2 = \frac{\mu_0 I_2}{2r_2} \quad \text{into page (2)}
\]

\[
\vec{B}_3 = \frac{\mu_0 I_3}{2r_3} \quad \text{out of page (3)}
\]

\[
\vec{B}_{\text{total}} = \vec{B}_1 + \vec{B}_2 - \vec{B}_3 \quad \text{(4)}
\]

\[
\vec{B}_{\text{total}} = \left( \frac{4 \pi \times 10^{-7} \text{T} \cdot \text{m}}{\mu_0} \right) (1.5 \text{A})
\]

\[
\vec{B}_1 = \left( \frac{4 \pi \times 10^{-7} \text{T} \cdot \text{m}}{\mu_0} \right) (1.5 \text{A}) = 6 \times 10^{-7} \text{T} \quad \text{(5)}
\]

\[
\vec{B}_2 = \left( \frac{4 \pi \times 10^{-7} \text{T} \cdot \text{m}}{\mu_0} \right) (2.0 \text{A}) = 3 \times 10^{-7} \text{T} \quad \text{(6)}
\]

\[
\vec{B}_3 = \left( \frac{4 \pi \times 10^{-7} \text{T} \cdot \text{m}}{\mu_0} \right) (4.0 \text{A}) = 0 \text{A} \quad \text{(7)}
\]

\[
I_4 = \frac{2 \pi I_3 (0.25 \text{m}) (4 \times 10^{-7} \text{T})}{4 \pi \times 10^{-7} \text{T} \cdot \text{m}} \quad \text{(8)}
\]

Figure B.7: Experiment #2: Single concept worked examples - switched variable version

Problem 1. Consider the circuit shown below. Given that the current through \( R_3 \) is 0.25A, find the voltage of the battery.

**Student Solution 1**

\[
\frac{1}{R_{\text{eq}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} = \frac{1}{20 \Omega} + \frac{1}{50 \Omega} \quad \text{(1)}
\]

\[
R_{\text{eq}} = 12 \Omega \quad \text{(2)}
\]

\[
R_{\text{total}} = R_1 + R_{\text{eq}} = 10 \Omega + 12 \Omega = 22 \Omega \quad \text{(3)}
\]

Define \( I_1 \) to be current through \( R_3, \)

\[
I_2, \text{ current through } R_2, \quad I_3, \text{ current through } R_3
\]

\[
I_3 \cdot R_3 = I_2 \cdot R_2 \quad \text{(5)}
\]

\[
I_3 = 20 \Omega (0.25 \text{A}) = 0.17 \text{A} \quad \text{(6)}
\]

\[
I_1 = I_2 + I_3 = 0.25 \text{A} + 0.17 \text{A} = 0.42 \text{A} \quad \text{(7)}
\]

\[
V = I_2 R_2 = (0.42 \text{A}) (22 \Omega) \quad \text{(8)}
\]

\[
V = 9.2 \text{V} \quad \text{(9)}
\]

Problem 2. Consider the circuit shown below. Given that the current through \( R_3 \) is 0.25A, find the voltage of the battery.

**Student Solution 2**

\[
R_{\text{eq}} = R_2 + R_3 = 5 \Omega + 15 \Omega = 20 \Omega \quad \text{(2)}
\]

Define \( I_2 \) to be current through \( R_3, \)

\[
I_2, \text{ current through } R_2
\]

\[
I_2 R_2 = V \quad \text{(3)}
\]

\[
V = (0.25 \text{A})(20 \Omega) = 5 \text{V} \quad \text{(4)}
\]
<table>
<thead>
<tr>
<th>Comparison Prompts (Synthesis Switched Version)</th>
</tr>
</thead>
<tbody>
<tr>
<td>For each of the following questions, refer to the problem-solution pairs on pages 1 &amp; 2.</td>
</tr>
<tr>
<td>1) i) Consider lines 1-6 in solution one. What lines in solution two serve the same purpose? Explain your reasoning.</td>
</tr>
<tr>
<td>ii) Consider lines 7-16 in solution one. What lines in solution two serve the same purpose? Explain your reasoning.</td>
</tr>
<tr>
<td>2) i) In solution one, why does the student transition from line 6 to line 7?</td>
</tr>
<tr>
<td>ii) In solution two, is the transition from line 7 to line 8 similar? Explain.</td>
</tr>
<tr>
<td>3) Consider the order of the two solutions. Which solution order do you think is better? Explain why.</td>
</tr>
<tr>
<td>4) A friend in an introductory physics class at another institution asks for help in solving similar problems involving magnetic fields and circuits. Write a short (2-4 sentences) guide to help him/her.</td>
</tr>
</tbody>
</table>

Figure B.8: Experiment #2: Comparison prompts
B.3 Experiment #3 Training Materials

**Synthesis Worked Examples**

### Problem I
A square circuit of length $L = 40\text{cm}$ and unknown mass $m$ is falling (gravity directed as shown) through a region with a uniform and perpendicular magnetic field $B = 2.3\text{ T}$ directed out of the page. The resistors have resistances of $R_1 = 10\Omega$ and $R_2 = 25\Omega$, and the voltage of the battery is $9\text{V}$. At the instant shown, when part of the circuit is still in the region with a magnetic field, the velocity of the object is constant and equal to $10.0\text{ m/s}$ downward. Find the mass of the circuit.

**Student Solution I**

\[ F_B = F_g = ma \]  
\[ F_{Net} = F_B - F_g = ma \]  
\[ ILB - mg = 0 \]  
\[ m = \frac{ILB}{g} \]  
\[ |e| = \frac{d\Phi}{dt} \]  
\[ |e| = \frac{d(\Phi_A)}{dt} = \frac{d(BL)}{dt} = B\frac{dy}{dt} = BLv \]  
\[ \epsilon = BLv \text{ counter-clockwise} \]  
\[ \sum V = I_{total}R_{total} \]  
\[ V + \epsilon = I(R_1 + R_2) \]  
\[ I = \frac{V + \epsilon}{R_1 + R_2} \]  
\[ m = \frac{ILB(V + BLv)}{g(R_1 + R_2)} \]  
\[ m = \frac{(0.4\text{ m})(2.3\text{ T})(9\text{V} + (2.3\text{ T})(0.4\text{ m})(10.0\text{ m/s})}{9.81\text{ m/s}^2(35\Omega)} \]  
\[ m = 0.049\text{ kg} \]

### Problem II
A square circuit of length $L = 3.0\text{ m}$ is in a region with a changing magnetic field as shown. The magnetic field is directed into the page, and varies with time as $B(t) = B_0 t$, where $B_0 = 0.125\text{ T/s}$. The battery voltage is $5\text{V}$, $R_1 = 25\Omega$ and $R_2 = 75\Omega$. What is the magnitude and direction of the force on the top wire at time $t = 30.0\text{s}$?

\[ |\epsilon| = \frac{d\Phi}{dt} \]  
\[ |\epsilon| = \frac{d(BA)}{dt} = L^2\frac{dB}{dt} + L^2B_0 \]  
\[ \epsilon = L^2B_0 \text{ counter-clockwise} \]  
\[ \sum V = I_{total}R_{total} \]  
\[ V - \epsilon = I(R_1 + R_2) \]  
\[ I = \frac{V - \epsilon}{R_1 + R_2} \]  
\[ F(t) = B_0 R_2 \rightarrow F = 3.75\text{ T} \]  
\[ F = ILB = \frac{(V - L^2B_0)LB}{R_1 + R_2} \]  
\[ F = \frac{(5\text{V} - (3.0\text{ m})^2(0.125\text{ T/s})(3.0 \text{ m})(3.75 \text{ T})}{100\Omega} \]  
\[ F = 0.44\text{ N} \text{ towards the top of the page} \]

Figure B.9: Experiment #3: Synthesis worked examples
Single-Concept Worked Examples

Problem I. A square loop of conducting wire, length 40 cm is being pulled at a constant speed 10.0 m/s through a region with a uniform magnetic field \( \vec{B} = 2.3 \text{ T} \) directed out of the page as shown. At the instant shown, when part of the circuit is still in the region with a magnetic field, find the direction and magnitude of the induced emf in the loop.

Student Solution I

\[ |\vec{e}| = \frac{d\Phi_B}{dt} \quad (1) \]
\[ |\vec{e}| = \frac{d(S_{\Delta})}{dt} = B_L \frac{dy}{dt} \quad (2) \]
\[ |\vec{e}| = B \frac{dL}{dt} \quad (3) \]
\[ |\vec{e}| = (2.3 \text{ T})(0.4 \text{ m}) \quad (4) \]
\[ \varepsilon = 9.2 \text{ V counter-clockwise} \quad (5) \]

Problem II. A square loop of conducting wire, length 3.00 m is in a region with a changing magnetic field as shown. The magnetic field is into the page and varies with time as \( B(t) = B_0 t \), where \( B_0 = 0.125 \text{ T/s} \). What is the magnitude and direction of the induced emf in the loop?

Student Solution II

\[ |\vec{e}| = \frac{d\Phi_B}{dt} \quad (1) \]
\[ |\vec{e}| = \frac{d(S_{\Delta})}{dt} = \ell \frac{dB}{dt} \quad (2) \]
\[ |\vec{e}| = \ell^2 \sigma \quad (3) \]
\[ |\vec{e}| = (3.00 \text{ m})^2(0.125 \text{ T/s}) \quad (4) \]
\[ \varepsilon = 1.13 \text{ V} \quad (5) \]

Problem I. A current carrying wire of length \( L = 40 \text{ cm} \) is falling (gravity directed as shown) at a constant velocity through a region with a uniform and perpendicular magnetic field \( \vec{B} = 2.0 \text{ T} \) directed out of the page. If the current is 0.5A, find the mass of the wire.

Student Solution I

\[ F_B = F_B \quad (1) \]
\[ F_{\text{Net}} = F_B - F_T = ma \quad (2) \]
\[ 1LB - mg = 0 \quad (3) \]
\[ 7LB \quad (4) \]
\[ m = \frac{g}{9.81 \text{ m/s}^2} = 0.041 \text{ kg} \quad (5) \]

Problem II. A square loop of conducting wire, length \( l = 3.0 \text{ m} \) is in a region with a uniform magnetic field \( B = 1.2 \text{ T} \) as shown. If there is a current \( I = 0.25 \text{ A} \) in the loop, what is the magnitude and direction of the force on the top wire?

Student Solution II

\[ F = 1LB \quad (1) \]
\[ F = (0.25 \text{ A})(3.0 \text{ m})(1.2 \text{ T}) \quad (2) \]
\[ F = 0.9 \text{ N towards the top of the page} \quad (3) \]

Problem I. Given \( V_1 = 9 \text{ V} \), \( V_2 = 5 \text{ V} \), \( R_1 = 100 \Omega \) and \( R_2 = 150 \Omega \), find the magnitude of the current in the circuit shown.

Student Solution I

\[ \sum V = I_{\text{total}}R_{\text{total}} \quad (1) \]
\[ V_1 + V_2 = I(R_1 + R_2) \quad (2) \]
\[ V_1 + V_2 = 1 \quad (3) \]
\[ I = \frac{14}{35} = 0.4 \text{ A} \quad (4) \]

Problem II. There is a clockwise current \( I = 0.04 \text{ A} \) in the circuit shown. Given \( V_2 = 5 \text{ V} \), \( R_3 = 25 \Omega \) and \( R_2 = 75 \Omega \), find the voltage of the unknown battery \( V_2 \).

Student Solution II

\[ \sum V = I_{\text{total}}R_{\text{total}} \quad (1) \]
\[ V_1 - V_2 = I(R_1 + R_2) \quad (2) \]
\[ V_2 = V_1 - I(R_1 + R_2) \quad (3) \]
\[ V_2 = 5 \text{ V} - 0.04 \text{ A}(300 \Omega) \quad (4) \]
\[ V_2 = 1.0 \text{ V} \quad (5) \]

Figure B.10: Experiment #3: Single concept worked examples
### Mastery Prompts (Synthesis Version)

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each of the following questions refers to the two different problem-solution pairs on page 1.</td>
</tr>
<tr>
<td>A) What physical quantity is calculated in line 6 in Solution I and line 2 in Solution II?</td>
</tr>
</tbody>
</table>
| B) Explain any similarities and differences between line 6 in Solution I and line 2 in Solution II.  
   *Hint: As part of your answer, explain the \( \frac{dy}{dt} \) term in Problem I and the \( \frac{da}{dt} \) term in Problem II in light of the two physical situations.* |
| C) Consider line 7 in Solution I and line 3 in Solution II. Explain why both have a counterclockwise direction, even though the direction of the magnetic field is different in Problem I and Problem II. |
| D) Consider line 9 in Solution I and line 5 in Solution II. Explain why Solution I substitutes in \( V + \varepsilon \) while Solution II substitutes in \( V - \varepsilon \). |
| E) Why is the force due to the magnetic field on the loop in Problem I directed towards the top of the page (Solution I line 1)? Explain your reasoning. |
| F) The magnetic field in Problem II is in the opposite direction of Problem I. Why is the magnetic force on the top wire in Problem II also directed towards the top of the page (Solution II line 9)? Explain your reasoning. |
| G) A friend in an introductory physics class at another institution asks for help in understanding and tackling similar problems. What are the most important ideas or information that you can tell your friend so that s/he can solve a similar problem? |

### Recognition Prompts

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each of the following questions refers to the two different problem-solution pairs on page 1.</td>
</tr>
<tr>
<td>A) What are the main physical concept(s) used in both of the students’ solutions?</td>
</tr>
<tr>
<td>B) For Problem I, identify the elements of the problem statement and/or diagram which indicate the need to use each of the physical concept(s) you mentioned in Part A. List the elements and explain your reasoning for each of them.</td>
</tr>
<tr>
<td>C) For Problem II, identify the elements of the problem statement and/or diagram which indicate the need to use each of the physical concept(s) you mentioned in Part A. List the elements and explain your reasoning for each of them.</td>
</tr>
<tr>
<td>D) Are the elements you identified in Part B and C similar or different between the two problems? Explain your answer highlighting any similarities and/or differences.</td>
</tr>
</tbody>
</table>
| E) In Problem I line 9, the student substitutes in for the total emf for both circuits. Explain why the student includes more than just the battery voltage, using the physical concepts you identified in Part A.  
In what ways is the situation in Problem II line 5 similar or different from Problem I line 9? Explain your reasoning. |
| F) A friend in an introductory physics class at another institution asks for help in understanding and tackling similar problems. What are the most important ideas or information that you can tell your friend so that s/he can solve a similar problem? |

---

Figure B.11: Experiment #3: Mastery & recognition prompts
Combined Prompts

Each of the following questions refers to the two different problem-solution pairs on page 1.

A) What are the main physical concept(s) used in both of the students' solutions?

B) For Problem I, identify the elements of the problem statement and/or diagram which indicate the need to use each of the physical concept(s) you mentioned in Part A. List the elements and explain your reasoning for each of them.

C) For Problem II, identify the elements of the problem statement and/or diagram which indicate the need to use each of the physical concept(s) you mentioned in Part A. List the elements and explain your reasoning for each of them.

D) One of the important concepts is that of an induced emf due to a changing magnetic flux. Compare the physical reason for a changing magnetic flux in each of the problems. Explain your answer highlighting any differences between the two problems.

E) A changing magnetic flux induces an emf. Explain any similarities and differences between line 6 in Solution I and line 2 in Solution II. Hint: As part of your answer, explain the \( \frac{dy}{dx} \) term in Problem I and the \( \frac{dg}{dc} \) term in Problem II in light of the two physical situations and your answers to Part D.

F) Consider line 7 in Solution I and line 3 in Solution II. Explain why both induced emfs have a counter-clockwise direction, even though the direction of the magnetic field is different in Problem I and Problem II.

G) The battery is not the only source driving the current in both problems. Consider line 9 in Solution I and line 5 in Solution II. Explain why Solution I substitutes in \( V + \epsilon \) while Solution II substitutes in \( V - \epsilon \).

H) A friend in an introductory physics class at another institution asks for help in understanding and tackling similar problems. What are the most important ideas or information that you can tell your friend so that s/he can solve a similar problem?

Figure B.12: Experiment #3: Combined prompts
B.4 Experiment #4 Training Materials

![Synthesis Worked Examples + Annotation](image)

Figure B.13: Experiment #4: Presentation of worked example annotations

The concept(s) used in this section are:

The goal of this section is:

The concept(s) used in this section are:

The goal of this section is:

The concept(s) used in this section are:

The goal of this section is:

The concept(s) used in this section are:

The goal of this section is:

The concept(s) used in this section are:

The goal of this section is:

$F = 0.44 \text{ N}$ towards the top of the page
## Combined Prompts

<table>
<thead>
<tr>
<th>Prompt</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A)</td>
<td>What are the main physical concept(s) used in both of the students' solutions?</td>
</tr>
<tr>
<td>B)</td>
<td>Compare the main solution steps used to solve both problems. Explain in what ways they are similar and/or different.</td>
</tr>
<tr>
<td>C)</td>
<td>One of the important concepts in both problems is that of an induced $\text{emf}$ due to a changing magnetic flux. Compare the physical reason for a changing magnetic flux in each problem. Explain your answer highlighting any differences between the two problems.</td>
</tr>
<tr>
<td>D)</td>
<td>A changing magnetic flux induces an $\text{emf}$. Explain any similarities and differences between line 6 in Solution I and line 2 in Solution II. <em>Hint: As part of your answer, explain the $\frac{dx}{dt}$ term in Problem I and the $\frac{dB}{dt}$ term in Problem II in light of the two physical situations and your answers to Part C.</em></td>
</tr>
<tr>
<td>E)</td>
<td>Consider line 7 in Solution I and line 3 in Solution II. Explain why both induced $\text{emfs}$ have a counter-clockwise direction, even though the direction of the magnetic field is different in Problem I and Problem II. <em>Note: It is not enough to only state the name of a rule or physical law; you must explain and compare the application of that rule/law in the two cases.</em></td>
</tr>
<tr>
<td>F)</td>
<td>The battery is not the only source driving the current in both problems. Consider line 9 in Solution I and line 5 in Solution II. Explain why Solution I substitutes in $V + \varepsilon$ while Solution II substitutes in $V - \varepsilon$.</td>
</tr>
<tr>
<td>G)</td>
<td>Consider line 1 in Solution I and line 10 in Solution II. Explain why the force of the magnetic field on both current-carrying wires is directed towards the top of the page in both problems, even though the direction of the magnetic field is different in Problem I and Problem II. <em>Note: It is not enough to only state the name of a rule or physical law; you must explain and compare the application of that rule/law in the two cases.</em></td>
</tr>
</tbody>
</table>

---

**Figure B.14: Experiment #4: Combined prompts**