A STUDY OF STUDENT RESPONSES TO TEXT-ONLY AND ILLUSTRATED CONCEPTTEST QUESTIONS RELATED TO PLATE TECTONICS: DIFFERENCES BY GENDER AND PRIOR ACHIEVEMENT

A Dissertation

Presented to

The Graduate Faculty of The University of Akron

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

Kyle Raymond Gray

August, 2009
A STUDY OF STUDENT RESPONSES TO TEXT-ONLY AND ILLUSTRATED CONCEPTTEST QUESTIONS RELATED TO PLATE TECTONICS: DIFFERENCES BY GENDER AND PRIOR ACHIEVEMENT

Kyle Raymond Gray

Dissertation

Approved: ________________________________
Advisor
Dr. Katharine D. Owens

Accepted: ________________________________
Department Chair
Dr. Bridgie A. Ford

______________________________
Committee Member
Dr. David N. Steer

______________________________
Committee Member
Dr. Xin Liang

______________________________
Committee Member
Dr. Catharine C. Knight

______________________________
Committee Member
Dr. Nidaa Makki

______________________________
Dean of the College
Dr. Mark D. Shermis

______________________________
Dean of the Graduate School
Dr. George R. Newkome

______________________________
Date
ABSTRACT

Many introductory undergraduate science courses use active learning pedagogies to improve student learning. Peer instruction is one such methodology that research suggests improves student learning. At the heart of peer instruction is the asking of conceptest questions. These conceptually-based, multiple-choice questions are presented to the class via a digital projector, and students individually respond using a personal response system.

Numerous research studies tout the benefits of peer instruction and personal response systems, however little work has been done on the conceptest questions. The few published examples of student responses do not consider whether different groups of students (such as men and women or students of different achievement levels) select different answers. Similarly, prior research assumes that all conceptest questions are equally valid and that students do not exhibit differences in their responses based on whether the question consists of only text or contains an illustration.

This study analyzed responses to conceptest questions asked in nine introductory earth science courses offered at a large Midwestern university. Conceptest questions asked during classes covering topics associated to plate tectonics were categorized as containing only text or containing an illustration. A repeated measures, mixed analysis of variance (ANOVA) determined that the 561 students gave a similar, non-significant percentage of correct responses to text-only and illustrated questions. There was no
significant difference in student responses when analyzed within a given gender, but when the data were compared between the genders, it was found that men answered a significantly higher percentage of text-only and illustrated concept test questions.

Additional analyses comparing prior achievement (student ACT-Reading, ACT-Math, and ACT-Science scores) indicated that students at any level of prior achievement correctly answered a similar, non-significant percentage of text-only and illustrated concept test questions. Comparisons between levels of achievement found significant differences in student responses.

Given the visual nature of geoscience inquiry and a demonstrated gender difference in spatial ability, the results from this study suggest that student responses to concept test questions may be influenced by their ability to visualize or mentally manipulate images or data. Such a finding provides one explanation for why men had higher responses for both types of questions; however further research is needed to establish this connection. The results from the ACT tests suggest that student achievement in reading, math, or science does not influence one’s ability to answer the different types of concept test questions and further suggests that the concept test questions included in this study served as a valid formative assessment of student comprehension of the course content.
ACKNOWLEDGEMENTS

A study of this size requires the help of many people. First I would like to acknowledge those helped to academically prepare me for this challenge. Thanks to my former teachers who pushed me to never stop learning including Pam Owens, Carolyn Johnson, and Scott Wilson at Kittitas High School, Dr. Raymond Preiss, Dr. Bryan Smith, Dr. Barry Goldstein, and Dr. Al Eggers at the University of Puget Sound, Dr. Gary Smith at the University of New Mexico, and Dr. Barbara Sanders at Whitworth University. Without your help, guidance, and encouragement, I would not have been prepared to earn this degree. I would also like to thank all of my friends and colleagues over the years who forced me to see the world from a different perspective and encouraged me to persevere when the learning became difficult.

I would also like to thank those at the University of Akron who were instrumental in my completion of this degree. Thank you to Dr. David McConnell for allowing access to his courses as well as the many stimulating discussions on geoscience education. Thanks also to Elaine Butcher for all her assistance in completing this manuscript and to Tom Quick for helping with technical issues. I would also like to thank the member of my committee. Thanks to Dr. Nidaa Makki for your outsider’s view and constructive comments on my research; to Dr. Catharine Knight for inspiring me to consider the findings from educational psychology; to Dr. Xin Liang for your tutelage in the art of inferential statistics and for enduring my numerous questions; to Dr. David Steer for
providing the funds for this project as well as a mentor who has shaped my identify as a geoscience educator; and to Dr. Katharine Owens – my advisor. From the first day we talked, you have done nothing but help me achieve this goal whether that was a constructive comment on my draft or a reminder that an important deadline was looming. Thank you.

Lastly, I would like to thank my family for all of their love and support over the years. To my parents, Ray and Wilma Gay, who opened my eyes to the wonders of the natural world and encouraged me to never stop learning. Thanks also to my children, Oksana, Elena, Lydia, and Alex for patience as dad worked to “finish his book”. May you all achieve great things in your lives. Lastly, thank you to Carla, my faithful companion who has encouraged me every step of the way and provided the emotional support needed to complete this task. Without your love and strength, I would not have completed this journey.
TABLE OF CONTENTS  

| LIST OF TABLES | xiii |
| LIST OF FIGURES | xv |

CHAPTER  

I. OVERVIEW OF THE STUDY .................................................................1  
   Introduction .................................................................................1  
   Purpose of the Study .................................................  
   General Research Questions .............................................3  
   Significance of the Study .................................................4  
   Lack of Previous Research on Student Responses to Conceptest Questions .......5  
   Gender Differences .................................................................6  
   Prior Achievement .................................................................8  
   Student Responses to Different Types of Conceptest Questions .......................8  
   Research Methods .................................................................9  
   Definition of Terms Used in the Study ...........................................10  
   Delimitations of the Study ..........................................................13  

II. REVIEW OF RELATED LITERATURE ..........................................15  
   Introduction .................................................................................15  
   The Nature of Geoscience Inquiry .............................................16  
   Visual Nature of Geoscience .........................................................17  
   Temporal and Dynamic Frameworks Within Earth System Science ................19
Student difficulties with the temporal nature of earth system science ..........20
Student difficulties with the dynamic nature of earth system science ..........21
Educational Implications of Temporal and Spatial Scales in Geoscience .............24
Difficulties teaching temporal time scales .......................................................25
Difficulties teaching very large or very small spatial scales ............................27
Impacts on Geoscience Education ...............................................................................29
Active Learning Strategies ........................................................................................29
Peer Instruction ........................................................................................................31
Personal Response Systems .......................................................................................37
Development of personal response systems .....................................................38
Research on personal response systems ...........................................................38
Literature reviews .....................................................................................................39
Student attitudes towards personal response systems ........................................40
Student learning and personal response systems .............................................43
Conceptest Questions ..............................................................................................47
Terms and usage ........................................................................................................47
Characteristics of a good conceptest question ....................................................48
Research on Conceptest Questions ........................................................................50
Studies with aggregated student responses to conceptest questions ..........53
Studies with student response data to individual conceptest questions ..........57
Studies that report individual student responses to conceptest questions ......60
Summary of previous studies that report student responses to conceptest questions.................................................................61
Conceptest Questions and Gender .........................................................................62
Gender, peer instruction, and course achievement .............................................62
Differing attitudes towards using a personal response system ...................171
Differences in spatial ability ........................................................................173
The possible influence of other psychological and social variables ............176
Prior Achievement ........................................................................................178
Prior achievement and response differences between text-only and
illustrated questions .......................................................................................178
Response differences within different achievement groups ......................181

Applications to Geoscience Education .......................................................182

VI. SUMMARY OF THE STUDY .................................................................186

REFERENCES ..............................................................................................188

APPENDICES ................................................................................................203

APPENDIX A. INSTITUTIONAL REVIEW BOARD CONSENT LETTER........204
APPENDIX B. STUDENT DEMOGRAPHIC DATA ...........................................206
APPENDIX C. TEXT-ONLY CONCEPTTEST QUESTIONS USED IN THIS
STUDY ..........................................................................................................207
APPENDIX D. ILLUSTRATED CONCEPTTEST QUESTIONS USED IN
THIS STUDY ....................................................................................................222
APPENDIX E. BIVARIATE SCATTERPLOTS TO TEST MULTIVARIATE
NORMALITY ..................................................................................................260
APPENDIX F. Q-Q PLOTS TO TEST MULTIVARIATE NORMALITY ............261
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Summary of Published Studies that Mention Conceptest Questions</td>
<td>52</td>
</tr>
<tr>
<td>2. Student demographics for the nine earth science classes included in this study</td>
<td>100</td>
</tr>
<tr>
<td>3. Ranges of ACT Scores for each Quartile Defined in this Study</td>
<td>102</td>
</tr>
<tr>
<td>4. Defining Characteristics and Examples of Text-Only Conceptest Questions</td>
<td>106</td>
</tr>
<tr>
<td>5. Defining Characteristics and Examples of Illustrated Conceptest Questions</td>
<td>107</td>
</tr>
<tr>
<td>6. Reliability and Standard Error Values For Three ACT Tests Used in This Study</td>
<td>109</td>
</tr>
<tr>
<td>7. Variables and Statistical Tests Associated With Each Research Hypothesis</td>
<td>114</td>
</tr>
<tr>
<td>8. Corresponding Values of $d$ and Partial Eta Squared for Different Levels of Effect Sizes</td>
<td>125</td>
</tr>
<tr>
<td>9. Student Demographics Used to Determine Study Sample Size</td>
<td>128</td>
</tr>
<tr>
<td>10. Numbers of Students From Each Class Included in This Study</td>
<td>130</td>
</tr>
<tr>
<td>11. Number of Class Sessions Devoted To Each Content Topic</td>
<td>132</td>
</tr>
<tr>
<td>12. Number and Types of Conceptest Questions Asked During Each Class Session</td>
<td>133</td>
</tr>
<tr>
<td>13. Normality Data For Text-Only and Illustrated Conceptest Questions Analyzed in This Study</td>
<td>138</td>
</tr>
<tr>
<td>14. Normality Data Categorized by Gender For the Percentage of Correct Responses to Text-Only and Illustrated Conceptest Questions</td>
<td>139</td>
</tr>
<tr>
<td>15. Normality Data Categorized by Student ACT-Reading Achievement Categories For the Percentage of Correct Responses to Text-Only and Illustrated Conceptest Questions</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>Title</td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>16.</td>
<td>Normality Data Categorized by Student ACT-Math Achievement Categories For the Percentage of Correct Responses to Text-Only and Illustrated Conceptest Questions</td>
</tr>
<tr>
<td>17.</td>
<td>Normality Data Categorized by Student ACT-Science Achievement Categories For the Percentage of Correct Responses to Text-Only and Illustrated Conceptest Questions</td>
</tr>
<tr>
<td>18.</td>
<td>Tests for Sphericity for Each Repeated Measures Analysis</td>
</tr>
<tr>
<td>19.</td>
<td>Bivariate Correlation Coefficients Between the Covariate Variable (Instructor), the Independent Variable (Gender), and the Two Dependent Variables (Text-Only and Illustrated Questions)</td>
</tr>
<tr>
<td>20.</td>
<td>Adjusted and Unadjusted Means For the Percentage of Correct Responses For Text-Only and Illustrated Questions</td>
</tr>
<tr>
<td>21.</td>
<td>Adjusted and Unadjusted Means For the Percentage of Correct Responses To Text-Only and Illustrated Questions Sorted by ACT Achievement</td>
</tr>
<tr>
<td>22.</td>
<td>Bonferroni Post Hoc Pairwise Comparisons of Correct Student Responses Categorized by ACT Achievement Category</td>
</tr>
<tr>
<td>23.</td>
<td>Paired T-Test Post Hoc Analysis Comparing Student Responses to Different Question Types For Each Achievement Category</td>
</tr>
<tr>
<td>24.</td>
<td>Summary of Analytical Results For Each Research Hypothesis</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Examples of text-only conceptest questions that required students to access knowledge of vocabulary (Question T-33) or evaluate criteria (Question T-13)</td>
</tr>
<tr>
<td>2.</td>
<td>Examples of illustrated conceptest questions that required the interpretation of a map and diagram or referred to a physical specimen that the student could manipulate.</td>
</tr>
<tr>
<td>3.</td>
<td>Differences by gender of the percentage of correct responses for each type of question</td>
</tr>
<tr>
<td>4.</td>
<td>Percentage of correct student responses for each type of question compared to their ACT-Reading achievement categories adjusted for instructor as a covariate. P-values indicate whether the responses by question type are significantly different.</td>
</tr>
<tr>
<td>5.</td>
<td>Percentage of correct student responses for each type of question compared to their ACT-Math achievement categories. P-values indicate whether the responses by question type are significantly different.</td>
</tr>
<tr>
<td>6.</td>
<td>Percentage of correct student responses for each type of question compared to their ACT-Science achievement categories. P-values indicate whether the responses by question type are significantly different.</td>
</tr>
</tbody>
</table>
CHAPTER I

OVERVIEW OF THE STUDY

Introduction

“Geology was called a descriptive science, and with its pitted outwash plains and drowned rivers, its hanging tributaries and starved coastlines, it was nothing if not descriptive” (McPhee, 1981, p. 24).

John McPhee illustrates how earth science is a highly visual discipline (Clary, 2003; Gobert, 2005; Manduca, Mogk, & Stillings, 2004; Rudwick, 1976) by vividly describing the geologic setting of the western United States. Besides educating the reader on the geology of Nevada and Utah, McPhee uses rich, visual, descriptions to teach such fundamental principles of geology as plate tectonics and erosion.

Undergraduate earth science instructors, like John McPhee, also use visual imagery to teach geologic concepts. They show students photographs illustrating the different types of volcanic eruptions, diagrams illustrating Earth’s internal layers, maps showing ocean currents, and graphs comparing global ocean temperature with time. Research on the role of images and diagrams in the learning process confirms this visual pedagogy by demonstrating that using visual aids foster greater student comprehension (Fletcher & Tobias, 2005; Hegarty & Just, 1993; Mayer, 2001).

In the classroom of the 21st Century, earth science instructors often use multimedia technologies to effectively present visual aids designed to illustrate and teach geologic
concepts. Computerized presentation software such as Microsoft PowerPoint allows instructors to illustrate geoscience concepts such as plate tectonics. For example, block diagrams show the characteristics of tectonic plate boundaries, digital photographs display rock layers in the Grand Canyon, computerized animations illustrate the three types of earthquake waves, and data-based simulations show global earthquake locations. The result is a multimedia learning environment that accesses both verbal and visual cognitive processes to foster student comprehension of geoscience concepts.

Instructors also use advances in computer technology to assess student conceptual understanding. Personal response systems (colloquially known as ‘clickers’) are one popular technology that allows instructors to poll their classes to determine whether the students understand a given concept (Crouch & Mazur, 2001; Mayer et al., 2009; Trees & Jackson, 2007). Students use a small hand-held infra-red or radio frequency transmitter to answer multiple choice questions posed by the instructor. The student responses are collected by a receiver attached to a computer where the class-wide results may be displayed as a histogram. Many undergraduate science instructors have adopted Mazur’s (1997a) peer instruction pedagogy as an effective means of using personal response systems to evaluate student comprehension (Fagen, Crouch, & Mazur, 2002; McConnell et al., 2006). With peer instruction, students use the response system to answer conceptually based, multiple-choice questions called ‘conceptest’ questions before discussing the problem with their peers. This process encourages students to evaluate their own conceptual understanding and provide the necessary feedback for the instructor to monitor student progress during the lecture (Fies, 2005; Mazur, 1997a).
Conceptest questions are also widely used in undergraduate science courses (see Fies & Marshall, 2006, MacArthur & Jones, 2008) including geoscience courses (McConnell et al., 2006). Many of these questions use illustrations to visually present data that students must interpret before arriving at the correct response, whereas other conceptest questions only use text to ask the question. Therefore, students in an undergraduate, earth science course that utilizes conceptest questions must possess the requisite conceptual knowledge as well as the ability to interpret different types of conceptest questions. To date, few research studies on this topic have investigated student responses to conceptest questions, and even fewer studies report responses by gender, prior achievement, or question type. This study was designed to fill in this gap by identifying differences in correct student responses to conceptest questions based on student gender or prior achievement.

Purpose of the Study

This study investigated responses to conceptest questions given by students taking an introductory earth science course at a large, public, Midwestern university. Specifically, the study investigated whether students respond differently to text-only questions compared to illustrated conceptest questions. Furthermore, the study identified whether student gender or prior achievement affected the percentage of correct student responses to either type of conceptest question. Any observed differences in student responses were interpreted using conclusions drawn from research in cognitive psychology.

General Research Questions

This study investigated differences in student responses to two different types of conceptest questions.
Research Question #1: In an undergraduate earth science course, do men and women correctly answer a different percentage of text-only and illustrated conceptest questions?

Research Question #2: Does the percentage of correct student responses to text-only and illustrated conceptest questions vary as a function of prior achievement as measured by student ACT scores?

Significance of the Study

Conceptest questions are conceptually-based, multiple-choice questions that require students to utilize their understanding of underlying concepts presented in the class (Mazur, 1997a). Typically an instructor displays the question during class using commercial presentation software such as Microsoft PowerPoint, which the students answer, often using an electronic personal response system (i.e. clickers). The results are often displayed as a histogram so the instructor can formatively assess the pacing of the lecture and the students can evaluate how well they comprehend the material. In the United States, the use of conceptually-based (conceptest) questions, in conjunction with personal response systems, has grown dramatically over the past five years (Fies & Marshall, 2006). Despite the rising popularity of conceptest questions and personal response systems, few studies have identified factors that might influence student responses to these questions. For example, Mayer et al. (2009) assessed whether using a personal response system fostered student comprehension and King and Joshi (2008) discussed possible gender differences in classes that used the technology, but neither study presented student response data or considered variations in student responses to conceptually-based questions. This study evaluated differences in student responses to
conceptest questions by comparing correct responses across student subpopulations based on gender and prior achievement.

Lack of Previous Research on Student Responses to Conceptest Questions

Publications on personal response systems or their associated conceptest questions tend to focus on assessing the course pedagogy, student attitudes, or student learning gains throughout the course rather than investigating differences in student responses to the conceptest questions themselves. Many papers provide advice on implementing personal response systems (Beekes, 2006; d’Inverno, Davis & White, 2003) or present self-reported survey data on student attitudes towards personal response systems (Bombaro, 2007; Uhari, Renko, and Soini, 2003). Other studies report using conceptest questions in class but assess gains in student comprehension by using independent data instruments such as the Force Concept Inventory (Crouch & Mazur, 2001; Hestenes, Wells, & Swackhamer, 1992), midterm or final exams (Mayer et al., 2009), or final course grades (Lasry, 2008). Other studies report conceptest questions to illustrate either the format of a well-written question (Debourgh, 2007; Nicol & Boyle, 2003) or a typical student response pattern (Brewer, 2004; Graham, Tripp, Seawright, & Joeckel III, 2007). To date, the literature focus has been on the implementation of the technology or the evaluating the impact on student achievement or attitudes rather than reporting data on student responses to conceptest questions.

The few studies that report student responses to conceptest questions typically focus on evaluating student comprehension of the course material. In addition, those studies that report data on individual conceptest questions assume that all questions elicit uniform student responses. For example, Eggert, West, and Thomas (2004), Pradhan,
Sparano, and Ananth (2005), and Stowell and Nelson (2007) reported aggregated response data that used a single number to summarize answers over an entire semester, and other studies (Crouch, Watkins, Fagen, & Mazur, 2007; DeMaria, 2005; Gray, Steer, McConnell, & Owens, in press; Piepmeier Jr, 1998) reported student responses to individual questions. A few studies report response data that compare gender with course performance (King & Joshi, 2008), or prior achievement with the time needed to answer a conceptest question (Gonzalez-Espada & Bullock, 2007). Yet, these studies appear to assume that responses from all student sub-populations as well as the conceptest questions themselves are uniform and equal across the study and do not investigate the possibility that student answers may vary based on factors pertaining to each student or question. An extensive search of three internet databases and 110 peer-reviewed journal articles yielded no studies that considered the possibility that men and women might correctly answer a different percentage of conceptest questions or that student prior knowledge might impact the selection of a correct response. This study adds to the existing research on conceptest questions by investigating whether men and women respond equally to different types of conceptest questions and provides data on how individual students respond to different types of conceptest questions.

*Gender Differences*

Few published studies consider how men and women respond to conceptest questions. Mazur’s Harvard Physics Education Research Group has shown that using peer instruction and conceptest questions eliminates the gender gap in introductory physics classes (Crouch et al., 2007; Lorenzo, Crouch, & Mazur, 2006); however, these studies use results from the Force Concept Inventory (Hestenes et al., 1992; Hestenes &
Halloun, 1995) rather than responses to conceptest questions to determine gender
differences. A study by Gonzalez-Espada and Bullock (2007) reported gender differences
in the time it takes students to answer conceptest questions, but this study analyzed a very
small sample population (n = 27 students), and thus the authors were hesitant to suggest
that the observed differences were actually present. A more recent study (King & Joshi,
2008) related gender differences in student final exam scores to the number of times
students used a personal response system. Even though these studies included conceptest
questions, they relied on measures other than student responses to assess gender
differences in student comprehension and assumed that question formatting or style did
not influence how men and women responded to a set of conceptest questions.

Research on gender differences and learning suggests that gender may influence
responses to the different types of conceptest questions (text-only versus illustrated).
Studies suggest that women tend to perform better than men on tests of verbal ability
(Halpern et al., 2007) but men outperform women on assessments of mental rotations and
spatial ability (Geary, 1995; Halpern et al., 2007). Gender differences have also been
observed in science problem-solving (Moreno & Mayer, 1999) and student reading
ability (Walla, Hufnagel, Lindinger, Deecke, & Lang, 2001). The format of a question
may also influence how men and women perform in a science course (Hamilton, 1998).
Given the need to correctly understand the words in a text-only question and to correctly
interpret the maps and diagrams found in illustrated conceptest questions, a gender
difference may exist in student responses to the two types of conceptest questions, as yet
prior research has not addressed this topic. If student responses are influenced by spatial
or reading ability, then the answers given in class may not accurately assess of student learning.

Prior Achievement

Like gender, few research studies have investigated the connection between prior achievement and student responses to conceptest questions. Research from cognitive psychology suggests that student prior knowledge or ability within a given content domain may influence how much information a student learns from a multimedia presentation (Hegarty & Just, 1993; Mayer, 2001). Interactive lectures that utilize peer instruction and conceptest questions provide learning opportunities through both the visual and auditory senses, and as such, are a form of multimedia learning (Mayer, 2001). Mayer and Gallini (1990) found that students with low prior knowledge experienced a greater benefit from material that contained both text and explanatory illustrations than students with a high level of prior knowledge. These studies suggest that a link may exist between student prior achievement and understanding of textual material. If a similar dynamic exists when students answer conceptest questions, then prior achievement may influence student responses to conceptest questions. To date, the few studies that have investigated the relationship between prior achievement and conceptest questions have not compared student responses to levels of prior achievement (Gonzalez-Espada & Bullock, 2007; Griff & Matter, 2008).

Student Responses to Different Types of Conceptest Questions

Research from cognitive psychology suggests that the way a conceptest question is presented (i.e. text-only versus illustrated) may impact how students process the information and may influence student responses. For example, if a text teaches the
relationship between two concepts or requires the reader to solve a problem, the inclusion of a diagram improves comprehension when compared to reading text without the accompanying diagram (Hegarty & Just, 1993). However, reading a text containing a diagram does not improve learning when the reader answers questions that require the repetition of known facts (Fiore, Cuevasa, & Oser, 2003; Mayer, 1989). Conceptest questions require students to evaluate new information or solve a problem using concepts presented in class and the results can be used by the instructor and students to formatively assess student comprehension (Crouch & Mazur, 2001; Mazur, 1997a; McConnell et al., 2006). It is possible that students may respond differently to conceptest questions that include an illustration compared to questions that only contain text. This study investigated this question by comparing the percentage of correct responses to text-only and illustrated conceptest questions.

Research Methods

This study used an ex post facto, causal-comparative research design to evaluate differences in student responses to two types of conceptest questions. According to Gay and Airasian (2000), causal-comparative research attempts to determine “the cause of reason, for preexisting differences in groups or individuals” (p. 349). In addition, the independent variables typically cannot be manipulated. For this study, the independent variables were gender and prior achievement in reading, math, and science, all of which cannot be experimentally manipulated. The dependent variables were the percentages of correct responses for two different types (text-only versus illustrated) of conceptest questions that were asked during the first month of an undergraduate earth science course.
The participants were 561 undergraduate students enrolled in an introductory earth science course at a large Midwestern university. During each class session, the students answered one to six conceptest questions pertaining to topics related to plate tectonics and the solid earth. The conceptest questions are categorized as containing only text (text-only) or containing an illustration (illustrated). For each student, the percentage of correct responses is calculated for each of the two types of questions.

This study assessed only those conceptest questions asked during the nine to fifteen lessons on topics pertaining to plate tectonics and the geosphere. Plate tectonics is a foundational topic that shapes the core of earth science (Dodick & Orion, 2003a). In addition, the associated topics of volcanoes, earthquakes, mountain-building, rocks, and minerals provided the framework for understanding plate tectonics.

Differences in correct student responses between the two types of questions were assessed using a repeated measures (Stevens, 2007) analysis of variance. The same test was used to identify any gender differences in the percentage of correct student responses for the two types of conceptest questions. A repeated measures mixed-ANOVA analysis compared correct student responses to text-only and illustrated questions to student prior achievement (as measured by student ACT scores). The results were then interpreted through research from cognitive psychology.

Definition of Terms Used in the Study

This study uses terms that may have multiple definitions based on usage, or a formal definition may not exist. For the purposes of this study, the following terms are operationally defined as follows.
Personal Response System – An electronic system consisting of a receiver connected to a computer and a set of hand-held transmitters through which students can answer instructor-initiated questions.

Conceptest Question – A clearly worded, conceptually-based, multiple-choice question that requires student comprehension of given concept (Mazur, 1997a). When applicable, incorrect answers are based on common misconceptions (Crouch et al., 2007). These questions can be answered by a variety of techniques including an electronic personal response system, printed answer cards, bubble sheets, or a show of hands (Lasry, 2008; Mayer et al., 2009). In addition, not all questions answered using personal response systems are conceptest questions. For example, student responses to survey questions such as “Have you ever experienced an earthquake?” are not conceptest questions.

Text-Only Conceptest Question – A conceptually-based, multiple-choice question answered by students using a personal response system that only contains text and does not contain any accompanying illustration, diagram, graph, or graphic.

Illustrated Conceptest Question – A conceptually-based, multiple-choice question answered by students using a personal response system that contains text and one or more accompanying illustrations, diagrams, photographs, or graphs. The illustrations are a necessary, integral part of the question that students must correctly interpret in order to successfully answer the question. Questions that only contain text but refer the student to observe a physical object (such as a rock, a map, or a graph) are considered to be illustrated questions.
Peer Instruction – An active-learning pedagogical methodology designed by Harvard University’s Eric Mazur (1997a) where the instructor lectures for 10-20 minutes on a given concept, then asks the class a conceptest question. If fewer than 70% of the students correctly answer the question, the students discuss the question in small work groups before answering the conceptest question a second time. Conceptest questions form the cornerstone of peer instruction (Crouch et al., 2007); however the two are not synonymous. Peer instruction is an interactive pedagogical methodology whereas conceptest questions are a tool used to assess student comprehension of the concepts presented in a lecture.

Active Learning – For this study, active learning includes any pedagogy that fosters student engagement and participation while constructing and understanding new knowledge. The emphasis is on student attention to learning rather than physical movement of the body. Active learning is often contrasted with passive learning where students listen to a didactic lecture.

American College Test (ACT) – A standardized test written and administered by the American College Test, Inc. that is primarily taken by high school seniors. The ACT is aligned with a typical college-preparatory curriculum and is purported to measure student academic achievement (American College Test, 2008a). The test consists of sub-tests covering reading (ACT-Reading), English (ACT-English), math (ACT-Math), and science (ACT-Science), as well as one comprehensive score (ACT-Comprehensive) calculated as the mean of the four sub-tests. All five test scores are measured on a scale from 0 – 36.
Prior Achievement – For the purposes of this study, prior achievement is operationally defined as the previous demonstration that a student comprehends a body of content knowledge (such as earth science) and/or has demonstrated proficiency in a body of academic skills (such as reading). This study uses three of the four tests from the American College Test (reading, math, and science) to measure student prior achievement in reading, language, math, science.

ACT Achievement Category – One of four categories based on the quartile rankings of student scores on an ACT test. The categories are labeled high, high-average, low-average, and low and correspond to the four quartile rankings for each ACT test used in this study.

Geosphere – This term is used by geoscientists to denote the inanimate portion of the Earth system that is composed of rock and includes the crust, mantle, and core (McConnell, Steer, Knight, Owens, & Park, 2008), and is also referred to as the “solid earth”. It does not include the air that surrounds the planet (atmosphere), the water on and within the planet (hydrosphere), or the sum total of all living organisms (biosphere).

Delimitations of the Study

Results from this study were delimited in the following manner. The students investigated in this study enrolled in an earth science course at a Midwestern university. Approximately 12% of the students included in this study were self identified as a racial or ethnic minority. It is possible that students from other racial, ethnic, or cultural backgrounds may have responded differently to the two types of conceptest questions. In addition, this study investigated conceptest questions asked during a specific content
domain (plate tectonics, volcanoes, earthquakes, mountain building, rocks, and minerals) within a specific course (earth science). Questions asked in other scientific disciplines (such as physics), or other content domains (such as global climate change) might not yield the same results. Lastly, this study included student responses collected from classes taught by two instructors who worked to coordinate their lessons so the same topics were taught on the same days using a similar set of conceptest questions. Despite these efforts to ensure uniformity of instruction across the classes, it is possible that inherent differences in teaching style may have influenced the student responses to the text-only or illustrated conceptest questions.
CHAPTER II
REVIEW OF RELATED LITERATURE

Introduction

This study categorized student responses by gender and prior achievement and compared the percentage of correct responses to text-only and illustrated concept test questions asked during an introductory, undergraduate earth science course. A topic such as this must utilize knowledge from a variety of scholarly disciplines including science education, assessment, and educational technology as well as educational and cognitive psychology. Chapter II is divided into four sections that discuss the published findings related to this study.

The first section in this chapter reviews the nature of geoscience inquiry including its visual nature and the need for geoscience instructors to convey abstract theories that incorporate both temporal and spatial scales that dwarf human experience. Plate tectonics and geologic time are two such constructs that utilize graphic illustrations, yet even with such visual teaching aides, instructors continue to find these concepts difficult to teach (Gobert & Clement, 1999).

The second section discusses the ways in which the visual nature of geoscience impacts teaching and learning. Within the context of teaching earth science, instructors use several active-learning pedagogies to teach such abstract and unobservable phenomena such as plate tectonics. Two such teaching tools that work well within an
active-learning environment are personal response systems (clickers) and conceptest questions. Conceptest questions are discussed by numerous authors but few researchers report student responses to questions. In addition, this section reviews the current understanding of how text and images used to convey conceptual material are interpreted by the students.

The third section reviews the literature on variations of spatial ability and the possible impacts on science learning. Studies on the composition of text and illustrations suggest that how scientific content is presented may impact student understanding of the material. Differences in the spatial ability of men and women have also been observed, and several studies have discussed the possible educational implications of such variations. Given the visual nature of the geosciences, several studies have recently demonstrated how student spatial ability may impact their ability to learn complex geologic concepts.

The fourth section describes the American College Test (ACT) and how previous studies have used this instrument. Of particular interest for this study is the ongoing discussion over gender differences pertaining to ACT test results and how student scores on the ACT can be used to assess student achievement.

The Nature of Geoscience Inquiry

The geosciences embody a set of knowledge and explanatory theories that traditionally are taught using a visually descriptive language accompanied by detailed illustrations (Clary, 2003). Two of the fundamental concepts in geoscience (plate tectonics and geologic time) remain difficult to teach because they describe phenomena that operate on temporal and spatial scales much larger than experienced during a human lifetime (Gobert & Clement, 1999).
Visual Nature of Geoscience

Researchers on the history and nature of geoscience inquiry have repeatedly commented on the visual nature of the discipline. Rudwick (1976) described how modern geoscientists make extensive use of maps, cross-sections, and diagrams when giving a scholarly talk whereas earlier geologists did not rely on such visual aids. Geoscience texts from the late eighteenth century primarily contained few illustrations and tended to rely on verbal descriptions to convey scientific concepts, but by the 1830s most geoscience texts contained numerous color illustrations. Rudwick suggested that the change to a more visual form of communication was fostered by the development of newer printing techniques such as lithography and was essential in the emergence of geology as a distinct scientific enterprise. Furthermore, he described the interpreting of geoscience illustrations as a form of visual literacy requiring an understanding of the complex rules and conventions inherent in the map, cross-section, or diagram. In this context, geoscience illustrations must be interpreted before they can be understood.

Similarly, Hineline (1993) documented the visual nature of geoscience using three case studies from works published prior to the 1960s covering topics in paleontology, geomorphology, and tectonics. He noted how geoscience used pictures to convey important scientific concepts. Hineline’s analysis of paleontological illustrations culminated in a dichotomous classification of geoscience illustrations as either proxies or diagrams. He defined proxy illustrations as graphics that illustrate or copy nature and do not contain any interpretation by the originator. Drawings of fossils in paleontological textbooks are examples of proxy illustrations. In contrast, Hineline reserved the term diagram to represent illustrations that include inferences or generalizations about the
subject of the illustration. Using this definition, a geologic map is an example of a diagram.

Of importance to this study, Hineline (1993) concluded that diagrams were pivotal in the development of the tectogene theory as a means of explaining observed gravity anomalies. Tectogenes were conceived as long, narrow folds found below ocean trenches and were thought to be a precursor to mountain building. Hineline argued that diagrams were a vital component in the acceptance of this theory. Eventually a more comprehensive theory (plate tectonics) emerged from the scientific inquiry and supplanted the tectogene model.

Similarly, Clary (2003) described the development of geoscience as a visual discipline and observed that introductory earth science textbooks contained a variety of graphics including maps, graphs, photographs, and diagrams. Her analysis of 72 texts from 1788 to 1840 yielded a significant correlation between the number of graphic illustrations and publication year ($r = 0.42$, $t_{(70)} = 3.88$, $p < 0.01$) suggesting that visual aids provided a crucial role in the development of geoscience inquiry. Clary also detailed the innovative illustrations of Henry de la Beche including his accurate portrayal of rock sections and the creation of illustrations that encapsulated deep time. Many of his illustrative conventions are still used by practicing geoscientists. Clary concluded that de la Beche could be considered the “Father of Visual Geology.”

Geoscience illustrations also provide the author with a visual means of conveying an abstract scientific concept. Manduca et al. (2004) summarized the nature of geoscience inquiry and described how geoscience is rich in multidimensional representations. They concluded that the visual nature of geoscience dictates how experts in the field must
possess an array of visualization techniques including the ability to interpret different types of maps or comprehend three-dimensional views of geologic data or phenomena. In addition, the successful geoscientist must cultivate the ability to “read” the rocks of a given area and use those observations to construct its geologic history. These skills require the use of complex cognitive and spatial processes that have yet to be fully investigated, and Manduca et al. recommended further study in this area.

Historically, the use of illustrations and visualizations to augment explanatory text has grown over time. Both Rudwick (1976) and Clary (2003) documented the concurrent growth in the number of scientific illustrations and development of geoscience inquiry. Hineline (1993) extended the contribution and influence of scientific illustrations through much of the 20th Century and provided an example of how visualizations influenced geoscience thinking. Manduca et al. (2004) echoed the findings of those authors by commenting on the visual nature of geoscience and concluded that earth scientists use a variety of mental visualization techniques to comprehend geoscience phenomena. They also recommended further research into the area of cognitive abilities and the visual nature of the discipline.

**Temporal and Dynamic Frameworks Within Earth System Science**

Earth system science is an integrative framework for organizing geoscience knowledge and concepts, and considers Earth as one large, complex system that should be studied from a multi-disciplinary perspective (MacKenzie, 2003). This inclusive, earth-system viewpoint was summarized by NASA (1988) as a scientific goal to understand the entire earth system by studying the complex interactions of its component systems. Earth system science requires expertise from many scientific disciplines and has
become a widely accepted conceptual framework within the geoscience community (e.g. Johnson, Ruzek, & Kalb, 2000). Two fundamental concepts with in earth system science are the depth of geologic time and the dynamic motions of Earth’s surface. Research evidence suggests that in both cases, K-12 teachers struggle with these concepts and in turn struggle to pass this knowledge on to their students.

*Student difficulties with the temporal nature of earth system science.*

The temporal nature of earth system science describes how change occurs over time. Both physical systems, such as the uplift of a mountain range, and biologic systems, such as the workings of an ecosystem, are not static, unchanging entities, but continually evolve and change. When viewed historically, this process of constant evolution means that Earth has a natural history which can be interpreted from the evidence left in the fossil and rock records. Geologic history (also called “deep time”, McPhee, 1981) incorporates an enormous time scale that is beyond human experience (Trend, 2000), and as such, is a difficult concept to teach (Trend, 2001).

Research on perceptions of geologic time indicates that both students and teachers hold inaccurate beliefs about the ordering of geologic events. Trend (2000) found that pre-service elementary school teachers grouped all geologic events into three informal categories and often confused the relative timing of key events such as the timing of Earth’s formation relative to the sun or the universe. Trend (2001) reported that in-service elementary teachers struggled with the concept of geologic time and believed that time considerations were irrelevant when discussing geologic specimens. To counter these inaccurate conceptions, Trend developed a framework for teaching key geologic events such as the extinction of the dinosaurs. Trend (2002) reiterated that geologic time is a
crucial geoscience concept and suggested that it may be a critical barrier to student comprehension of the geosciences. Other research (Dodick & Orion, 2003a; Dodick & Orion, 2003b) has also indicated that high school students struggle to use logic to interpret problems involving geologic time; however, field-based experiences can strengthen student understanding of geologic time.

_Student difficulties with the dynamic nature of earth system science._

Besides having a history, Earth is a dynamic planet with a surface that is continually in a state of physical change (MacKenzie, 2003). For example, ocean currents continually circulate water, heat, and nutrients across the planet, and air masses constantly move air from region to region. So too, Earth’s lithosphere is not a stationary, rigid, unmoving mass of rock. Rather the tectonic plates shift and move against one another in a complex, dynamic system that continually reshapes the surface of the planet (McConnell et al., 2008). This process is best described in the theory of plate tectonics.

The theory of plate tectonics is based on Alfred Wegner’s older continental drift hypothesis which stated that the less dense continental crust was moving over the more dense oceanic crust. Plate tectonic theory states that Earth’s surface is divided into approximately eight major, and several smaller, rigid plates which are formed at mid-ocean ridges and consumed at oceanic trenches (McConnell et al., 2008). This theory provides a unifying framework that geoscientists use to explain a wide range of observations including volcanic activity, earthquakes, mountain building, and ocean floor topography.

Most students are introduced to plate tectonics in elementary school. Prior research suggests that this topic is difficult to teach at any age level. Gobert and colleagues
(Gobert, 2000; Gobert, 2005; Gobert & Clement, 1999; Gobert & Pallant, 2004) have published several studies characterizing the process of attaining an accurate understanding of plate tectonic processes. They noted that plate tectonics consists of hidden, unobserved mechanisms that lie outside the typical human experience (Gobert, 2000; Gobert & Clement, 1999). In order to fully comprehend plate tectonic processes, earth science students must also incorporate spatial, causal, and dynamic knowledge (Gobert & Clement, 1999). For example, students must comprehend the spatial arrangement of plates or rock layers, the causes for plate movement, and the dynamic interactions that occur at plate boundaries. Such attributes make plate tectonics difficult to conceptualize.

Studies from this research group have focused on using student drawings to assess improved conceptual understanding. Gobert and Clement (1999) discussed differences in student learning when asked to only read a passage on plate tectonics compared to verbally summarizing the key points or drawing their conceptions. The results indicated that summarizing the text produced more sophisticated responses during the experiment, but those students who drew their concepts before and after the lesson experienced larger learning gains on a posttest assessment. Also, student spatial understanding was significantly enhanced by drawing their conceptions.

Another study (Gobert, 2000) with 47 fifth grade students investigated the role of student prior knowledge on the development of scientifically accurate mental models. Students were interviewed and asked to draw their understanding of several topics related to plate tectonics (e.g. volcanoes, Earth’s layers, and earthquakes). They reported that prior conceptual or mental models held by students can either inhibit or accentuate new
mental model development. Some students clung to inaccurate models and continued to incorporate these inaccuracies even after direct instruction on the topic. For example, one student continued to place the crust at the north pole, the mantle at Earth’s center, and the core at the south pole even when tutored on the spatial arrangement of these three layers. By contrast, if students possessed a correct spatial model, they appeared to use this knowledge as a tool for constructing a newer, more accurate mental model.

Gobert and Pallant (2004) evaluated increases in student learning when using a new, cognitively-based curriculum to teach plate tectonic concepts. The study asked 360 fifth grade students to draw a model of their conceptual understanding of the target concept and write a summary of their thinking. After these products were critiqued by a student from another class, students had an opportunity to revise their summaries and drawings. This curriculum yielded significant learning gains pretest to posttest for five classes ($F = 16.0 – 75.5, p = 0.002$).

Lastly, Gobert (2005) investigated whether explaining or diagramming student conceptions affected student learning. In this study, students from two fifth-grade classes explained what would happen in a given situation, whereas in a previous study (Gobert & Clement, 1999), 58 fifth-grade students were asked to summarize their thoughts. The results suggested that there was no significant difference in learning when students must explain their understanding versus diagramming their understanding ($F = 5.72, p = 0.001$). The author suggested that both diagramming and explaining (but not summarizing) activated deep-processing of the content material resulting in improved learning.
The preceding studies only assessed student learning of elementary students. Steer, Knight, Owens, and McConnell (2005) used Gobert’s methodology to investigate student conceptual change among their undergraduate earth science students. The instructor used a conceptual change model approach to teaching the plate tectonic concepts. For each lesson, the 97 students completed a pre-class reading assignment and drew their initial understanding of Earth’s layers. After a short lecture on Earth’s internal structure, the students created a physical model of the layers using string and markers and viewed an expert example as a target model. After comparing and contrasting their models to the expert model, the students drew a new diagram of their conceptual understanding and wrote an explanation of their thinking. The typical student drew a diagram showing a rudimentary concentric structure with layers of uniform thickness. In contrast, student posttest drawings were significantly different and displayed differential layer thicknesses that approximated the target model. Scores for both the pre- and post-instruction student drawings exhibited no significant gender differences. The authors noted that students bring naïve preconceptions to class which appear to be difficult to modify. They tentatively concluded that using physical models within a conceptual change pedagogy facilitated student learning about earth’s internal layers.

Educational Implications of Temporal and Spatial Scales in Geoscience

Besides struggling to teach the complex, dynamic nature of Earth processes and the history of these events, geoscience instructors must also convey the enormous temporal and spatial scales over which geologic processes operate.
Difficulties teaching temporal time scales.

Any discussion of geologic time includes events that occur millions or billions of years in the past and represent time periods that dwarf a human lifetime. For example, the Big Bang occurred approximately 11 to 20 billion years ago, Earth formed 4.6 billion years ago, and the dinosaurs became extinct 65 million years ago. Even geologically “recent” events such as the latest glacial retreat approximately 12 thousand years ago or the periodic eruption of a volcano such as Mount Vesuvius occur at scales that outlast personal human memory.

Recent studies illustrate the complex conceptual nature of the geologic time concept and describe why students struggle to understand this foundational issue. Marques and Thompson (1997) described alternative conceptions held by elementary and middle school students concerning geologic time, and used their results to develop teaching guidelines that foster conceptual understanding. Additional work by Trend (2002) documented how teachers struggle with the concept of geologic time and consequently struggle to teach their students. More recently, Dodick and Orion (2003c) described student understandings of geologic time and found that students who study earth science over a two-year time frame, coupled with hands-on fieldwork, yielded a better understanding of geologic time than students who took just one year of earth science.

Marques and Thompson (1997) surveyed and interviewed 493 Portuguese students (10 – 15 years old) concerning the origin of the planet, the age of Earth, and the sequence in which various organisms first appeared. Their results suggested that students held common alternative conceptions. For example, student answered that water and other planets existed before Earth’s formation, and more students selected earthworms ahead of
fish as the oldest organism. In addition, many of the alternative conceptions were not altered after formal teaching on the topic of geologic time. The authors concluded that students exit high school with an inaccurate view of the subject and these inaccurate conceptions may produce barriers to learning as undergraduate students.

If students enter adulthood with inaccurate conceptions of geologic time, it is possible that they carry these ideas into the classroom as science educators. Trend (2002) summarized his findings to date including Trend (1998, 2000, 2001) on the alternative conceptions held by students and teachers. Like Marques and Thompson (1997), Trend’s research indicated that people at all age levels cluster geologic events into discrete time categories such as ancient, moderately ancient, less ancient, and geologically recent with teachers demonstrating a more complex understanding of the concept than the elementary and secondary students. His data also suggested that people possess a wide range of loosely-held geoscience conceptions, but these ideas are not understood in detail. Similarly, the introduction of a numeric time scale lowered the accuracy of people’s understanding of events throughout geologic time. Therefore, Trend recommended that Earth events and processes be taught in conjunction with the appropriate temporal framework.

Dodick and Orion (2003c) identified cognitive factors in middle school and high school students that affect their understanding of geologic time. They studied 507 students in grades 7-12 using three instruments designed to evaluate student spatial abilities as well as their understanding of geoscience concepts pertaining to geologic time. They found that all participants understood basic principles such as superposition if the rocks were drawn as horizontal layers. In contrast, all participants struggled to
interpret drawings with tilted rock layers, with the younger students struggling more than the older students. Like Marques and Thompson (1997), Dodick and Orion (2003c) found that students in grades seven most struggled to comprehend geologic time, so the authors suggested that the principles behind a scientific understanding of geologic time should be taught starting in grade eight.

These papers all demonstrate that geologic time remains a difficult set of concepts to understand and that people carry many alternative ideas into adulthood. The studies also illustrate how altering these conceptions is a difficult task that requires a synthesis of factual knowledge about our planet and the theoretical framework within which we interpret geologic phenomena. For the undergraduate earth science instructor, these studies indicated that students will not easily adopt the scientific viewpoint, but learning gains can be achieved with a comprehensive, earth systems teaching approach.

*Difficulties teaching very large or very small spatial scales.*

In the same manner, Earth’s active processes often take place at spatial scales that are much larger than human experience (Jacobi, Bergeron, & Malvesy, 1996). For example, Earth’s radius is greater than 6300 km, the oceans are up to 10 km deep, and one volcanic eruption can eject several cubic kilometers of ash and debris into the atmosphere (McConnell et al., 2008). These features and processes are difficult to comprehend when compared to common events that humans readily experience (Jacobi et al., 1996). Therefore, geoscience students must be able to conceptualize events, processes, and concepts that operate over a wide range of spatial scales.

Tretter, Jones, and Minogue’s (2006) study on accurate scale conceptualizations illustrated why students struggle with earth science concepts such as plate tectonics. They
asked 215 students from elementary (5th, 7th, and 9th grades) and graduate school (doctoral students) to name objects as small as one nanometer to as large as one billion meters in length. Participant interviews provided additional insight into student thinking about scale. The results indicated that over 80% of all participants, regardless of age, could accurately name objects as small as one millimeter, but this percentage sharply declined for micrometer and nanometer-sized objects. Conversely, all age groups exhibited a gradual decline in accuracy when naming objects larger than one meter. The results suggested that students of all ages who did not have direct experience with a given scale face the same struggles when asked to think at an extreme scale. Yet, those students who successfully conceptualized at the smallest and largest scales relied on a strategy of converting a given scale into a more familiar unit of measurement. For example, many students mentally converted large units such as one million meters into a more manageable unit of one thousand kilometers and compared that number to distances traveled in a car.

The results from Tretter et al.’s (2006) study suggested that students in an introductory earth science course will also struggle to conceive of very large or very small objects. These students most likely have few experiences with extremely large objects such as tectonic plates or very small objects such as silica molecules. Consequently, when these students encountered scientific concepts such as plate tectonics or mineral crystallization, their lack of experience most likely hindered their ability to accurately comprehend the underlying process. To address this difficulty for students to comprehend processes that operate at very large and small scales, geoscience instructors
must find ways to bridge this conceptual gap and assist students in visualizing very large and very small objects.

Impacts on Geoscience Education

This study investigated student responses to different types of conceptest questions. As such, conceptest questions were the focus of this study; however, the conceptest questions were asked within a larger classroom context and were inexorably entwined with both the instructor’s choice of teaching strategies as well as the technological tools used to assess the student responses. Therefore, this section discusses the use and effectiveness of active-learning strategies in a large, undergraduate earth science course and research on electronic personal response systems before discussing the literature on conceptest questions. The section ends with a discussion on how students interpret and decode meaning from the text and diagrams used in conceptest questions.

Active Learning Strategies

Active learning describes a teaching style and method of course organization that seeks to engage the student and foster student comprehension of the material (Uekart & Gess-Newsome, 2006). There is no formal, agreed-upon definition of active-learning. Several studies (Braxton, Jones, Hirschy, & Hartley III, 2008; Hamer, 2000; Keyser, 2000; Sarason & Banbury, 2004; Steiner, 2000; Wilke, 2003) cite Bonwell and Eison’s (1991, p. 2) definition of active learning as instruction which “involves students doing things and thinking about the things that they are doing.” Active learning is grounded in a constructivist, social-learning framework which states that students learn best when they actively construct their own knowledge by incorporating new experiences with their prior knowledge on the subject (Bransford, Brown, & Cocking, 2000). From a constructivist
perspective, students do not passively receive new knowledge handed down from an authoritative teacher. Active learning strategies such as the ones described below promote student engagement of the material.

Meyers and Jones (1993) presented active learning in terms of two fundamental assumptions in that learning is an active endeavor and that different people learn in different ways. They further stated that active learning is typically viewed in opposition to the traditional, didactic, instructor-centered lecture. They proposed that active learning involves three factors: Basic elements such as reading and writing; learning strategies such as cooperative learning; and teacher resources such as educational technology. Meyers and Jones further elaborated on this definition by comparing active learning to Socrates’ student-centered method of asking questions and Piaget’s theory that learning occurs when the student actively constructs new meanings.

More recently, Keyser (2000) noted the lack of a precise definition of active learning in the educational research literature, and suggested four general characteristics of active learning. First, students do more than listen to a lecture. Second, emphasis is placed on developing student skills rather than the transmission of information. Third, students are engaged in classroom activities such as reading, writing, and discussing the course concepts. Fourth, students in active classes have opportunities to explore their own attitudes and values. In all cases, active learning is viewed as a style of teaching that focuses on active student participation in the learning process and acquisition of skills rather than a passive transmission of facts that the student memorizes.

Active learning only positively impacts student learning when instructors utilize one or more teaching strategy that incorporates the goals of active learning. A number of
teaching strategies aligned with the active-learning philosophy have been developed for use in an undergraduate setting. For the purpose of this study, the discussion is limited to cooperative learning and peer instruction. Other examples not discussed in this review of the literature include: just-in time teaching (Guertin, Zappe, & Kim, 2007), problem-based teaching (Keller, 2002); and case studies (Meyers & Jones, 1993). In effect, any methodology that promotes student interaction with the course content can be viewed as a component of active learning.

Cooperative learning is one active-learning strategy wherein students are placed in small groups to maximize student learning (Keyser, 2000). Not all group work is cooperative learning. Cooperative learning uses student groups to focus on students helping one another achieve the desired learning objective. With this in mind, Keyser cautioned that student understanding is fostered when instructors carefully choose group exercises that engage all students within the group. Think-pair-share is one common example of cooperative learning (Cooper & Robinson, 2000; Lyman, 1981; Yuretich, Khan, Leckie, & Clement, 2001) in which the instructor poses a question and students discuss the question in pairs. If time permits the instructor invites students to share their conclusions with the whole class (Cooper & Robinson, 2000). Peer instruction (Mazur, 1997a) is a version of the think-pair-share methodology that has been widely adopted in undergraduate science courses.

**Peer Instruction**

Eric Mazur of Harvard University’s Physics Research Group developed the peer instruction strategy for use in introductory, undergraduate physics courses (Mazur, 1997a). This methodology has been adopted across many science and non-science
disciplines including biology (Knight & Wood, 2005), earth science (McConnell et al., 2006), mathematics (Lucas, 2007), and economics (Maier & Simkins, 2008). In addition, many of the courses that use personal response systems do so within the context of a peer instruction paradigm.

In the peer instruction model (Mazur, 1997a) a lecture is divided into small, 10-minute sections. The instructor then displays a conceptually-related, multiple choice question (a concepttest question) and the students individually answer the question using a variety of means such as raising their hands, using flashcards, or using a personal response system (Lasry, 2008). The students then discuss the question with their neighbors for two to four minutes before answering a second time. Crouch and Mazur (2001) provided a summary of the methodology and report results from a decade of use. Several different versions of this basic technique have been reported. Mazur (1997a) and Crouch and Mazur (2001) did not specify whether they assigned students to groups but rather indicated that students were encouraged to discuss their answers with their neighbors. In contrast, Gray et al. (in press) assigned their earth science students to permanent, four-person groups that remained intact throughout the semester.

Research on the peer instruction model suggests that students who experience this interactive learning environment learn more than students who experience a traditional didactic lecture. An early study by Hake (1998) compared achievement data from over 6,542 students from 62 introductory physics courses to demonstrate that interactive courses, such as those taught using peer instruction, resulted in significantly improved student learning. Starting in 1992, Hake solicited pretest/posttest data from institutions that administered the Force Concept Inventory and/or the Mechanics Baseline Test and
received data from 14 high schools, 16 colleges, and 32 universities courses. Hake
determined student learning for semester-long physics courses using pretest/posttest data
from two validated instruments (the Force Concept Inventory and the Mechanics
Baseline Test) and calculated student learning gains by normalizing the posttest results to
the pretest results using the equation \( g = \frac{\text{Posttest} - \text{Pretest}}{1 - \text{Pretest}} \). This value
described the maximum possible gain that each student can possibly achieve.

The results from Hake’s (1998) analysis suggested that students who experienced an
interactive learning environment (i.e. peer instruction) exhibited greater learning gains
than students from traditionally taught lecture courses. The average traditional lecture
course yielded a normalized gain of 0.28 but the average interactive learning course
yielded a normalized learning gain of 0.48. This increase was two standard deviations
higher than the control (traditional) courses. Hake did not compute whether this
difference was significant, yet the large difference in the two means suggests that the
difference was significant. This pattern was consistent across all three class levels (high
school, college, and university courses). Hake concluded that instructors should adapt
their courses to include this form of active learning and suggested that an apprenticeship
program could provide a forum where additional instructors could learn this
methodology.

Crouch and Mazur’s (2001) paper summarized their results from ten years of
implementing and using peer instruction in their introductory physics courses. Their data
(collected from 1990 to 2000) indicated that student pretest scores on the Force Concept
Inventory remained constant over the years, but students who experienced peer
instruction attained significantly higher posttest values (\( p = 0.001 \), effect size = 0.34) on
the same assessment. The authors also used the Mechanics Baseline Test to measure student problem-solving ability in a physics context. Their results indicated that students in a peer instruction course achieved an additional 6 – 13 percentage points above those students who took a traditional lecture course. The authors noted that they had modified the original peer instruction methodology in three ways. First, they have replaced in-class reading quizzes with web-based assignments that promoted student interaction with the text. Second, they used a physics textbook designed to be read before class rather than serve as a reference for use after the lecture. Third, they incorporated group activities in their course discussion sessions. They ended with a note that implementing peer instruction required a radical shift in instructor focus and proper use of the peer instruction model demanded a significant time commitment on the part of the instructor. Yet, these barriers were offset by gains in student comprehension.

Lenaerts, Wieme, and Van Zele (2003) presented a qualitative and quantitative case study of the peer instruction model as implemented in an introductory magnetism course designed for engineers taught at a Belgium university. They compared 200 Belgium engineering students who experienced a traditional lecture course with 200 engineering students who took a physics course with peer instruction. The paper included data from classroom observations however the authors did not describe how the data were collected. The results indicated that students who experienced peer instruction provided more detailed descriptions of the target phenomena than students who experienced a traditional lecture. Results from a pretest/posttest course assessment (the Magnetism Concept Inventory) further suggested that students who experienced a peer instruction course achieved higher learning gains than students who experienced a traditional lecture
course. The students from the traditional lecture course achieved an average normalized learning gain of 0.39 (out of a possible 1.00) but students from the peer instruction course achieved a learning gain of 0.59. The authors did not report if the earning gains were significantly different but did report that the values were consistent with the findings from Hake (1998). Lenaerts et al. also noted that the peer instruction course had a higher attendance rate than the traditional lecture course.

Nicol and Boyle (2003) tested the effectiveness of the peer instruction model by comparing it to an alternative methodology called “class-wide discussion”. Both models used personal response systems to ask conceptually-based questions and both systems made use of student discussions of the questions. The difference lies in the sequence in which the students answered and discussed the concept test questions. In peer instruction students individually answer a question and then discuss the answer in small groups. In class-wide discussion, the students first discuss the question in small groups before answering the question. They then engage in a class-wide discussion before voting a second time.

The Nicol and Boyle (2003) study investigated 117 students in an undergraduate engineering mechanics course offered at the University of Strathclyde in the United Kingdom. Over the course of the semester, the students experienced the class-wide discussion model for two weeks followed by peer instruction for an additional two weeks. Data sources included 30 student interviews, a 36-question survey using themes and ideas that emerged from the interviews, and a questionnaire with five open-ended questions that probed student experiences and attitudes towards peer instruction. Results from the study suggested that peer instruction was a more effective model than class-wide
discussion. A majority of students (82%) indicated that they preferred to individually answer a concept test question before discussing it with their small group, whereas only 13% of the students preferred small group discussion before answering the question. The students also indicated that they valued the opportunities to talk about the answers within a small group setting and they remained confused over the target concept if the discussion failed to address their struggles with the question.

Lastly, Lasry (2008) investigated whether the mode of collecting student responses affected the effectiveness of the peer instruction model. Lasry compared data from two physics courses as a two-year Canadian community college. One course (n = 41) used peer instruction and a personal response system and the second course (n = 42) used flashcards to respond to the concept test questions. Like other studies, student gains were measured using the Force Concept Inventory. Both courses showed large normalized learning gains using the method from Hake (1998). The course that used the flashcards produced a 52% improvement compared to a 49% improvement for the course that used the personal response system, yet, this difference was not significant (p = 0.745). Lasry interpreted the results as suggesting that peer instruction was an effective teaching strategy regardless of the method by which the instructor solicited student responses. He did note that using a personal response system allowed the instructor to more quickly collect, display and interpret the results than using the flashcards, and the data could be archived for later evaluation. These differences affected the instructor more than the student, and any instructor looking to implement the peer instruction model need not incur the expense of using a personal response system to improve student learning.
Available research demonstrated that the peer instruction teaching methodology improved student performance when compared with traditional lecture. Possible reasons for this learning gain included increased student attention by dividing the lecture into smaller, more manageable units as well as providing an opportunity for students to formally assess their own conceptual understanding and discuss the matter with their peers. In addition, peer instruction as a methodology often was used in conjunction with personal response systems; however, instructors can successfully use peer instruction with student flashcards or raising their hands. This study included classes that used personal response systems to collect formative data as part of the peer instruction methodology.

**Personal Response Systems**

Personal response systems are electronic voting systems that allow students to answer multiple-choice questions and instructors to observe the results. Researchers and adherents to personal response systems do not agree on a common term for this technology. Past publications have used clickers (Lasry, 2008), audience response systems (Graham et al., 2007), classroom response systems (Fies & Marshall, 2006), personal response systems (Griff & Matter, 2008), student response systems (Trees & Jackson, 2007), electronic response systems (Hatch, Jensen, & Moore, 2005), and i-clickers (Lucas, 2007). Weiman et al. (2008) provided a summary on the best ways to implement personal response systems in an undergraduate course.

The term *personal response system* is used throughout this study because this investigation focused on students and their responses to concepttest questions. The term audience response system connotes a focus on the class as whole, and terms such as
clicker, i-clicker, and electronic response system imply a focus on evaluating the technology. These terms do not suggest research on individual students and as such are not used. The term student response system would also imply research on individual students, but the term personal response system is more prevalent in the literature and implies the possibility of more general uses outside of the formal classroom.

*Development of personal response systems.*

Personal response systems have a long, forty-year history, including changes in how they have been implemented in the classroom and different impacts on student learning. Variants of this technology have been used since the 1960s (Bessler & Nisbet, 1971; Froelich, 1969; Judson & Sawada, 2002). Judson & Sawada (2002) and Abrahamson (2006) both provided overviews on the historical use of personal response systems. Early research results indicated that using personal response systems did not significantly improve student learning (Judson & Sawada, 2002 and references therein). Yet research on classes that use the newer, infra-red or radio-frequency transmitters yielded significant gains in student learning (Crouch & Mazur, 2001; Judson & Sawada, 2002). The improvements in student learning were not attributed to the improved technology, but rather to the embrace of constructivist pedagogies such as peer instruction (Judson & Sawada, 2002, Mayer et al., 2009).

*Research on personal response systems.*

Personal response systems are currently used in a variety of disciplines including physics (Crouch & Mazur, 2001); biology (Brewer, 2004); earth science (McConnell, Steer, & Owens, 2003); mathematics (d’Inverno et al., 2003), economics (Ball & Eckel, 2004), and nursing (Debourgh, 2007). Much of the research on personal response systems
originated from the physics education community. Three published literature reviews (Penuel, Roschelle, & Abrahamson, 2005; Fies & Marshall, 2006; MacArthur & Jones, 2008) have summarized the findings from this body of literature and highlight key published findings. These reviews have also noted that studies such as (Hake, 1998) claim that using a personal response system improves student learning, yet Clark (2001) and Mayer et al. (2009) warn that personal response systems do not cause learning, but rather the improved teaching methods (such as peer instruction) are the root cause of observed learning gains.

*Literature reviews.*

Three recent published literature reviews have described the findings on personal response systems. Penuel, Roschelle, and Abrahamson (2005) summarized their findings after investigating 41 research studies pertaining to personal response systems. They noted how previous studies suggest that using a personal response system improves student engagement but individual instructors may find it difficult to adjust their courses to the new format. Penuel et al. concluded that instructors must have a wide array of questions that assess all aspects of their course. They also noted how data from physics courses (Hake, 1998; Crouch & Mazur, 2001) suggested that using a personal response system improved student learning, improved classroom discussion, and defined student difficulties with the course content.

A second literature review (Fies & Marshall, 2006) analyzed 24 studies and discussed the findings in terms of pedagogical theory or implementation of a personal response system. Fies and Marshall noted that research suggested that using a personal response system helped students engage the material and made the course more enjoyable. They
also noted that personal response systems provided an anonymous format from which shy students were empowered to participate in the class. Fies and Marshall agreed with Penuel et al. (2005) that studies such (Crouch & Mazur, 2001) suggested a link between the use of this technology and improved student learning, but cautioned that differences in pedagogy may influence the observed learning gains.

The third literature review (MacArthur & Jones, 2008) discussed their findings after reviewing 56 publications on the use of personal response systems in college courses. They noted how research studies suggested that using personal response systems provided an avenue for formative assessment and improved student attitudes towards the course, but required a significant time commitment. The authors did not discuss student learning gains in connection with the use of a personal response system.

These three studies illustrated some of the common themes that have emerged from research on personal response systems. First, when students use this technology, they remain engaged with the course material and develop positive feelings for the course. The technology also affords shy students an opportunity to actively participate in the class without fear of retribution over an incorrect answer. Second, some studies suggest that personal response systems promote improved student learning, but not all studies agree with this conclusion. In addition, some of the studies cited by these reviews may have confounded the technology with changes in instructional methodology.

Student attitudes towards personal response systems.

All three published literature reviews described above report that using personal response systems promoted positive student attitudes towards the course and the course content. These conclusions are typically based on studies which relied on student self-
reported assessments on surveys administered at the end of the semester. Three studies are discussed to illustrate this body of literature.

Graham et al. (2007) investigated student attitudes towards using a personal response system. Their study encompassed 688 students from 10 courses that participated in a pilot program. Courses included chemistry, physics, biology, psychology, statistics, education, as well as marriage, family, and human development. After using a personal response system for a semester, students in these courses completed a 14-question survey on student attitudes towards the technology. The overall findings indicated that a majority of students thought the system was useful in helping them learn. A majority of the students agreed or strongly agreed that the system made the class more enjoyable (77%), improved student interest in the course (78%), and fostered student participation (87%). The authors also determined that students who were reluctant to actively participate in the class held similar attitudes towards the response system than students who were not reluctant. This study provided data on student attitudes towards courses that used a personal response system; however, the results were qualified by the fact that the courses did not use a common pedagogy and no control classes were assessed.

Crossgrove and Curran (2008) investigated the attitudes of major and non-major biology students towards the use of a personal response system. They compared survey data from 196 students in a 100-level biology course for non-majors with data from 46 biology majors taking a 200-level genetics course. The survey included 11 Likert-scale questions and four open-ended questions on student experiences using the response system. Chi-square analysis of the survey responses produced non-significant differences between the majors and non-majors. These included positive attitudes towards
involvement, attention, and understanding the course contents. Two questions produced significant differences between the two groups. The biology majors more strongly agreed that the response systems did not help them on the course exams, and the non-majors were more likely to recommend the continued use of the personal response system.

Similar results were found by Greer and Heaney (2004) who asked 582 earth science students from four classes to complete a 21-item survey administered as part of the midterm exam and write open-ended responses to a prompt on the effectiveness of the response system. The authors noted an increase in average student attendance (30% - 40%) in previous traditional lecture courses compared to 81% - 90% in a class using a personal response system. The survey results suggested that a majority of the students felt the response system reinforced the course contents (71% - 85%), was an effective learning tool (54% - 57%), and would recommend continued use of the system (67% - 71%). Answers to the open-ended responses suggested that the response system motivated shy students to participate in class, made the class more enjoyable, and helped students learn the course material. Negative comments included dislike for the price of the transmitters, the time taken away from lecture, and not receiving credit when the student failed to bring the transmitter.

These studies illustrated how similar student attitudes are observed across a variety of course disciplines. Most students enjoyed using the response systems and used them to formatively assess their learning. The technology also provided a forum in which shy students could actively participate in the course without incurring unwanted attention. As discussed below, these results must be tempered with the understanding that student
attitudes towards the course may be a reflection of the altered pedagogy rather than the technology itself.

*Student learning and personal response systems.*

Previous studies on the relationship between using a personal response system and student learning have been inconclusive. Results from two studies described in a previous section (Hake, 1998; Crouch & Mazur, 2001) suggested that students who use personal response systems earned higher gains on measures of student comprehension. Both studies reported significant learning gains from classes that used personal response systems compared to classes that used a traditional lecture. Yet these studies also included a change in pedagogy as well as the introduction of a new technology. It is uncertain whether the observed gains are due to the implementation of an active learning pedagogy such as peer instruction or the use of a new technological device. Mayer et al. (2009) noted that teaching methods and not technology cause learning, and cautioned that studies which purport to demonstrate a causal relationship between use of a personal response system and student learning gains actually describe the more efficacious active learning methods used in conjunction with this technology.

Knight and Wood (2005) also compared student learning gains to the use of personal response systems. In two courses designed for biology majors, the authors used a pretest/posttest assessment and homework scores to evaluate student learning. One course (n=73) used a traditional lecture whereas a second course (n=73) used a personal response system as part of an interactive class. Students from the traditional (control) class produced a 46% normalized gain on the posttest assessment, but the students from the interactive class had a higher, significant, learning gain (62%). In addition, 17 of the
19 graded homework scores from the interactive class were significantly higher than the scores from the control course. Knight and Wood concluded that using personal response systems as part of an interactive class did improve student learning.

Another paper by Mayer et al. (2009) demonstrated that using personal response systems within a peer instruction setting and not the conceptest questions themselves can positively impact student learning. The subjects of this study were predominately female upper-level students taking an educational psychology course. Data from three classes were included in the study. One class used a personal response system to answer conceptest questions during 18 lectures. A second class answered the same questions printed on a paper handout, and the third class served as a control and did not answer any of the conceptest questions. Students from all three classes were assessed using the midterm and final exams for the course. Questions from the exams were similar to the conceptest questions given in class. The results indicated that students who used the personal response systems performed significantly better (effect size = 0.38) on exam questions that were similar and dissimilar to the conceptest questions. This difference was equivalent to approximately one-third of a letter grade. There was no difference in exam scores for students who answered on paper and students who never saw the conceptest questions. Even though this paper provided data supporting the claim that using a personal response system in a peer instruction setting improved student achievement, the authors did not report any response data associated with answering the conceptest questions, so it is not known how the student responses related to their exam scores.
Not all studies yielded a positive connection between personal response systems and student learning. Suchman, Uchiyama, Smith, and Bender (2006) compared students taking two microbiology classes. Students in one class (n = 84) experienced a traditional lecture format but used personal response systems to answer a single pretest question at the beginning of each class; whereas, students in another section of the same course (n = 143) answered questions throughout each lecture period. Their results indicated that the students who actively used the personal response systems throughout the lecture earned significantly higher grades (p < 0.001) on the final exam than their counterparts in the control class. Yet the students in the interactive course did not earn a significantly higher score on those final exam questions which were similar to in-class conceptest questions. The authors had expected the students who answered a larger number of conceptest questions to correctly answer a larger percentage exam questions when compared to students in the control class. The fact that both classes answered a similar percentage of exam questions related to the conceptest questions led the authors to conclude that their results must be cautiously interpreted and suggested that differences in student demographics or instructor implementation might have affected the student responses.

Research results on the positive connection between personal response systems and improved student learning imply that using this technology is one way to optimize student learning. Yet, two warnings from the research on multimedia education are applicable to this setting. Clark (2001) warned that researchers should not confound the method of instruction with the medium in which it is delivered. Both Mayer, Almeroth, Bimber, Chun, Knight, and Campbell (2006) and Mayer et al. (2009) echoed this warning by advising that researchers who investigate educational technologies should “consider
the fundamental distinction between instructional media and instructional method” (Mayer et al., 2006). When considering studies that purportedly demonstrate a connection between using personal response systems and either improved student learning or improved student attitude towards the course material, one must not confuse the pedagogical method with technological delivery system.

Mayer et al. (2009) attempted to distinguish learning gains in courses that used an interactive pedagogy such as peer instruction from courses that use a personal response system and peer instruction. They assessed student learning in three educational psychology classes that used a different combination of technology and pedagogy. The first class (n = 139) served as a control class by not using a personal response system and did not use group discussions. The second class (n = 111) used a personal response system to answer concept test questions and participated in group discussions. The third class (n = 135) engaged in group discussions but answered the concept test questions on a sheet of paper. Using the final exam as the measure of student learning, the authors found that the class who used the personal response system answered a significantly higher percentage of exam questions (p < 0.05) than those students from the other two classes with moderate effect sizes ranging from d = 0.38 to 0.44. The same students also earned higher grades than their counterparts in the other two classes. In addition, students in the paper-only class performed at the same level as the students in the control class. The same pattern held true for questions on the final exam that were not asked as concept test questions. The authors concluded that using the personal response systems did contribute towards improving student learning gains, and they noted that the response system appeared to stimulate the student-instructor interactions.
My study was not designed to assess whether personal response systems result in improved student learning and did not evaluate gains associated with using conceptest questions. Instead, this study assessed correct student responses to different types of conceptest questions. This research focus was not concerned with evaluating the technology used to deliver the conceptest questions, but rather investigated the method by which the questions were presented. As such, this study heeded the warnings described by Clark (2001) and Mayer et al. (2009) by investigating variations in student responses and interpreting the results within the context of an active learning environment rather than ascribing differences in student responses to the technology itself.

*Conceptest Questions*

Conceptest questions form the foundation of Mazur’s (1997a) peer instruction pedagogy. Without these questions, it would be difficult for students to assess their comprehension of the course material and would have fewer opportunities to discuss the concepts with their peers. Instructors also use the answers to these questions to formatively assess student learning. Therefore, the successful implementation of peer instruction requires successful writing an implementing of conceptest questions. Yet, the extant research literature has focused on student learning gains measured over the entire course with few studies reporting student answers to conceptest questions.

*Terms and usage.*

The term *conceptest* was introduced by Mazur’s Harvard University’s Physics Education Research Group (Mazur, 1997a) and was originally spelled ConcepTest. Many of the recent articles on interactive learning pedagogies such as peer instruction do not capitalize the “T” in test (e.g. McConnell et al., 2006; Nagy-Shadman & Desrochers,
2008; Piepmeier Jr., 1998); however, Mazur’s research group continues to use the original spelling (Crouch, Fagen, Callan, & Mazur, 2004; Crouch & Mazur, 2001; Fagen et al., 2002), and some studies do not use the term conceptest at all (e.g. Hake, 1998; Suchman et al., 2006).

In most prior studies, the term conceptest was denoted those conceptually-based multiple-choice questions used during peer instruction. However, Mazur (1997a) and his associates (Rosenberg, Lorenzo, & Mazur, 2006) broadly defined conceptests to include the entire questioning process, not just the questions themselves. Mazur’s book on peer instruction (Mazur, 1997a, p. 10) stated that conceptests included the initial individual student answers, the group discussions, follow-up student answers, and instructor explanation of the correct answer. When writing about these conceptually-based questions, Piepmeier Jr. (1998) used the term “conceptest question” to denote the questions, and Gray et al. (in press) also used the term conceptest questions to indicate the questions that were asked as part of the peer instruction methodology. This study investigated correct student responses to different types of questions. It did not investigate variations in the peer instruction methodology such as different ways in which the questions were presented to the students, differences in how the students responded, or differences in how the groups discussed the questions. Therefore, this study operationally defined the term conceptest questions to denote the conceptually-based, multiple-choice questions that were asked during a conceptest.

*Characteristics of a good conceptest question.*

Several published articles have described the characteristics of a good conceptest question. Mazur (1997a) stated that these questions should be “short conceptual questions
on the subject being discussed” (p. 10). He further stated that a concept test question, “(a) forces the students to think through the arguments being developed, and (b) provides them (as well as the teacher) with a way to assess their understanding of the concept.” (p. 10). A uniform set of criteria for developing a good concept test question does not exist, however Crouch et al. (2007) provided the Harvard Group’s criteria as… “a good ConcepTest, must satisfy a number of basic criteria:

- Focus on a single important concept, ideally corresponding to a common student difficulty
- Require thought, not just plugging numbers into equations
- Provide plausible incorrect answers
- Be unambiguously worded
- Be neither too easy nor too difficult.” (p. 9-10)

Crouch and Mazur (2001) also noted that a well written concept test question should have an initial correct response rate of 35% - 70%.

Other researchers outside of Mazur’s research group have described the characteristics of good conceptually-based, multiple-choice questions. Even though these articles did not specifically discuss concept test questions, their analyses are relevant to this discussion. Greer and Heaney (2004) noted that writing good questions for a personal response system can be very challenging. They also noted how well-written questions provided an impetus for redesigning their lectures to include more active learning strategies. Green (2002) also commented on the difficulty in writing good questions and described how good concept test questions facilitated student learning and provided feedback for formative assessment of student learning. Beatty, Gerace, Leonard, and
Dufresne (2006) summarized the characteristics for designing good questions and presented a general framework based on the question’s role, goal, and mechanism. They recommended that questions form the core of a science course, and should meet well-defined content, process, and metacognitive goals. Considering the purpose of the question (i.e. the goals) could drive the instructor to develop questions that focus on conceptual understanding rather than rote memorization. Beaty et al.’s analysis aligned with the definition provided by Mazur’s research group (Crouch et al., 2007) and emphasized the importance on using conceptually based questions.

**Research on Conceptest Questions**

Peer instruction and personal response systems have been the subject of numerous studies (see Crouch & Mazur, 2001; Fies & Marshall, 2006; Judson & Sawada, 2002; MacArthur & Jones, 2002 for reviews), yet these studies tended to focus on the connection between peer instruction and student learning gains. My study investigated differences in student responses to two different types of conceptest questions (text-only and illustrated questions) rather than student learning gains measured across the length of the course. A literature search was conducted to assess the number of studies that have reported data on student responses to conceptest questions. An internet search using Google-Scholar and the keyword “conceptest” located 107 distinct items, of which 47 were peer-reviewed journal articles, 25 items represented abstracts from conference proceedings, six were books or chapters from books, and four were doctoral dissertations or masters theses. The remaining entries were 24 websites and three citations of other works.
Some evaluations of the peer instruction pedagogy or studies on the usage of personal response systems included data or descriptions of conceptest questions without using “conceptest” as a keyword. For example, Suchman et al. (2006) evaluated the effect of using a personal response in an undergraduate biology course and provided data on student responses, yet they did not use the term “conceptest” despite the clear fact that their questions met Mazur’s criteria. To broaden the number of references considered for this literature review, the online searches included keywords such as “clicker” and “response system”. Searches were conducted using Google-Scholar and other online databases such as Ohiolink. In addition, key references cited in other papers were obtained and evaluated and a general internet search located any additional publications. The list of conceptest references from Vanderbilt University (Bruff, 2008) was particularly helpful.

A reference was included if it contained a description or example of one or more conceptest questions answered using a personal response system or included student response data from such questions. This search yielded a total of 81 references which are summarized in Table 1. A majority of references (66.7%) came from science, math, or engineering disciplines, and one-third of all the references were from physics. The predominance of science disciplines and physics in particular, was not unexpected given that conceptest questions and peer instruction were first developed and studied in undergraduate physics courses (Abrahamson, 2006; Mazur, 1997a). The wide range of disciplines that have published on conceptest questions indicated the widespread popularity of peer instruction and personal response systems.
Yet, when the references were categorized by the type of data presented, few papers reported student response data. Over half of the publications (54%) only mentioned conceptest questions when describing the peer instruction methodology and did not provide any examples. Similarly, 26% of the references provided examples of conceptest questions but did not include any student response data. Only 18 (21%) of the 81 papers identified for this study reported student responses to conceptest questions.

Table 1. Summary of Published Studies that Mention Conceptest Questions

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Number of References</th>
<th>Percent of References</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Examples or Data</td>
<td>44</td>
<td>54.3</td>
</tr>
<tr>
<td>Examples Only</td>
<td>21</td>
<td>25.9</td>
</tr>
<tr>
<td>Aggregated Data</td>
<td>4</td>
<td>4.9</td>
</tr>
<tr>
<td>Question Data</td>
<td>7</td>
<td>8.6</td>
</tr>
<tr>
<td>Student Data</td>
<td>3</td>
<td>3.7</td>
</tr>
<tr>
<td>Other Data</td>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>Totals</td>
<td>81</td>
<td>100</td>
</tr>
</tbody>
</table>

A majority of the references from this analysis only mentioned the term “conceptest” but did not provide any examples or student response data. Many of these papers, such as Draaijer, Hartog & Hofstee (2007) or Tanner and Allen (2004) presented conceptest questions as a means of improving one’s pedagogy. Other authors mentioned conceptest questions in the context of evaluating a personal response system (e.g. Beekes, 2006; Lasry, 2008; Pradhan et al., 2005) or assessing the effectiveness of an active-learning methodology such as peer instruction (e.g. Crouch et al., 2004; Fagen et al. 2002; Zolan, Strome, & Innes, 2004). The second largest group of papers provided examples of conceptest questions but did not provide student response data. These publications spanned a wide range of disciplines including geoscience (Greer & Heaney, 2004;
McConnell et al., 2006), chemistry (Butcher, Brandt, Norgaard, Atterholt, & Salido, 2003), math (Schlatter, 2002), physics (Beatty et al., 2006), Biology (Brewer, 2004), nursing (Debourgh, 2007), packaging (Auras & Bix, 2007), and economics (Maier & Simkins, 2008). The publications in both of the groups described above did not provide data on student responses to any of the questions used during their studies. Even in papers that described research results (such as Lasry, 2008 and Knight & Wood, 2005), student response data were not reported.

Less than one quarter (22%) of the references included any type of student response data. Those studies that reported student responses did so in four different ways: 1) as aggregated data; 2) as responses to individual questions; 3) as answers by individual students; 4) and other types of response data. Even when response data were clearly presented, nearly all of the studies used the student answers to augment conclusions drawn from other data instruments and did not use the student responses as their primary data source. The publications for each of these four ways of reporting student response data are presented below.

*Studies with aggregated student responses to conceptest questions.*

Four papers presented student response data as a single percentage aggregated across all of the questions and all of the students. The earliest paper (Hatch et al., 2005) described their experiences in implementing peer instruction and conceptest questions and used the results from the conceptest questions to modify their teaching. In the text they reported how they spent more time discussing energy flow through ecosystems because only 51% of the students correctly answered the associated conceptest question, but they spent less time on other topics because over 80% of the students answered
correctly. They also described how 16% of their students correctly identified the trophic level for a given ecosystem but after three additional questions 100% of the students provided a correct response to a similar question and one week later 80% of the students continued to give a correct response. This study focused on the implementation of the personal response technology rather than assessing student responses to concept test questions. The reported data were provided to illustrate instructor use of the system and improvements in student understanding after experiencing the peer instruction pedagogy.

Freeman et al. (2007) assessed gains in student learning for an undergraduate biology course. This study measured student performance using final course grades and questions from an in-class exam, and compared the results to different teaching methodologies including the use of a personal response system, flashcards, and student study groups. They grouped students into high- and low-risk categories based on a multiple regression equation designed to predict the final course grade for each student. High-risk students were defined as students who were predicted to earn a D+ or lower. Their results suggested that using a personal response system significantly benefited only low-risk students, yet their analysis was based on scores from an in-class exam rather than the student responses. For one academic term they provided data on student responses to concept test questions but the results were presented as the percentage of correct responses across the entire semester for high-risk (HR) and low-risk (LR) students. Their results suggested that all students answered a higher percentage of questions when their answers are graded for correctness (59%) versus graded for participation (55%). Low-risk students also answered a significantly higher percentage of questions than their high-risk counterparts. Significant differences were observed between the low-risk and high-risk
groups when points were awarded for correct answers (LR = 63%, HR = 52%, p = 0.021) as well as when points were awarded for participation (LR = 59%, HR = 47%, p = 0.003).

Freeman et al. also observed that when comparing common questions on the midterm exam, students who earned participation points earned a higher, but non-significant, grade than students who earned points for correct responses. In addition, results from a previous semester suggested that high-risk students who used a personal response system earned a higher, but not significant (p = 0.08) score than high-risk students who used flashcards, but the low-risk students who used a personal response system earned a significantly lower score (10 points out of 335 possible, p = 0.034) on the midterm exam than low-risk students who used flashcards.

A third paper by Stowell and Nelson (2007) also reported aggregated student response data. Their study investigated the benefits of a personal response system on student participation, learning, and emotion. They tested 140 undergraduate psychology students to determine whether using low-tech alternatives to a personal response system (such as raising hands or using flashcards) were as effective as the high-tech option. Students completed a pretest and posttest quiz as well as a survey on their attitudes and emotions towards using each type of technology. The only student response data reported in this study were the percentage of correct responses to the conceptest questions (called “formal review questions” by the authors) asked in class. Like Freeman et al. (2000), Stowell and Nelson found that students who used a personal response system correctly answered a lower percentage (82%) of the course review questions than the students who used flashcards (92%) or raised their hands (98%); however those students who used the electronic devices did score the highest on the post-lecture quiz. The authors noted that
the high values for the low-tech options might reflect social conformity rather than actual differences in student learning.

A fourth paper by Gray et al. (in press) investigated whether using student-manipulated models during an earth science lecture significantly improved student learning. This article used the same student population as the study described in this dissertation, but used data from a smaller, separate set of conceptest questions. Gray et al. reported the percentage of correct responses from five model-use and two control classes and found that students who used the models answered a significantly larger percentage (11.5% with p < 0.001) of conceptest questions than those students who just listened to an interactive lecture that included peer instruction and conceptest questions. The individual conceptest questions were listed in this article, but responses are not segregated by student or by question.

In each of the four studies described above, the authors combined data into a single percentage of correct responses and used this number to evaluate the effectiveness of a given pedagogy. The data were aggregated over time periods ranging from a single class session (Gray et al., in press) to an entire semester (Freeman et al., 2007), but in all cases these studies focused on student learning gains after experiencing a given instructional strategy such as comparing the use of a personal response system to using flashcards or a show of hands (Stowell & Nelson, 2007). The results from these studies were also reported at the class level rather than the question level, so possible variation in student responses between individual questions or between types of questions were not addressed.
Studies with student response data to individual conceptest questions.

The largest category of studies examined for this review of the literature (41%) reported student response data as mean percentages to individual questions. In most cases the data were presented were not the primary data source for each study but augmented conclusions drawn from other data sources such as student surveys or course exam scores. The earliest published work to include student response data was Piepmeier Jr.’s (1998) work on using peer instruction and conceptest questions in a large pharmacy course. Data from 19 conceptest questions were collected as a pretest before instruction, posttest after lecture but before student discussion, and post-peer after student discussion. Piepmeier Jr. had to discontinue the study because students complained about using the system and requested a return to a typical lecture. He also provided an appendix with all 19 questions. The results suggested that students learned most from the lecture than from their peers. Student scores improved an average of 25% after lecture and 6.8% after discussion. One should note that Piepmeier Jr. asked each question three times over one class period. It is possible that the results suffered from an adverse treatment effect (Gall et al., 2007) because students may have become familiar with the question as time progressed. In addition, he did not account for a diminished range of improvement after peer discussion. For example, 65% of the students correctly answered question number 11 on the pretest and 97% answered correctly after the lecture. Piepmeier Jr. did not normalize the scores, so for this question the students could only achieve a additional 3% after discussion. The remainder of this paper qualitatively discussed how to implement peer instruction and a personal response system in a pharmacy course.
DeMaria (2005) described the use of a personal response system to collect survey data from physicians attending a debate on surgical methods to relieve obesity. Even though DeMaria did not collect data from an educational setting, his study was included in this discussion to illustrate the range of papers that present data from personal response systems. The 88 participants were provided three scenarios they would likely encounter and were asked to choose one of five medical procedures for each scenario. The physicians were polled both before and after a debate on using bariatric surgery to combat obesity and the results indicated that physician opinions varied depending on patient health factors.

Another study (Suchman et al., 2006) described the impact of using a personal response system in a microbiology course. Like Piepmeier Jr. (1998), Suchman et al. provided data from 19 conceptest questions and listed the percentage of students who correctly answered each question. Their data indicated that student scores improved by 22% after discussing the conceptest question with a peer. No other response data were provided.

Bombaro (2007) provided an example of using a personal response system and conceptest questions in a library science course. For a lesson on plagiarism, Bombaro presented a series of scenarios and had students indicate whether or not the situation was an example of plagiarism. She found that students were correct 46% - 92% of the time. While this study presented student response data, the results were not applicable to this study. Bombaro used a system that could only ask true/false questions, so the questions asked had only two options.
In a synthesis paper, Crouch et al. (2007) summarized the success of the peer instruction methodology by presenting student response data for seven conceptest questions and listing the data as the percentage of students who correctly answered each question. Like Piepmeier Jr. (1998) and Suchman et al. (2006), Crouch et al. found that the percentage of students who correctly answered a conceptest question improved from 46% to 72% after discussing the problem with a peer. On a subsequent exam, 61% of the students correctly answered nearly identical questions. Their data further suggested that not all of the questions elicited similar student responses. Of the seven questions reported, the percentage of students who correctly answered a question before any peer discussion ranged from a low of 9% to a high of 63%. Crouch et al. commented that the data were intended to demonstrate that students learned the concepts during peer instruction and retained them on the exam. Even though Crouch et al. reported data on student responses, the focus of the paper was evaluating the effectiveness of the peer instruction methodology rather than analyzing differences in student response rates.

Lucas’ (2007) paper was the final paper that reported student response data. This paper discussed the need to guide student discussions during peer instruction and presented data from an undergraduate calculus course. Lucas used the histograms from one conceptest question to illustrate how non-guided peer discussion may have produced an incorrect group consensus. Similarly, Lucas presented the results from a second question to illustrate how guided peer discussion led to meaningful student consideration of the topic. In this paper, the student responses were not presented in a systematic manner, but rather were used to illustrate the author’s main hypothesis.
Studies in this category did report variations in student responses to individual questions; however papers like Demaria (2005) and Lucas (2007) used the data as illustrations of the salient points in their papers and did not investigate differences in responses between different types of questions. The other papers that listed student responses to individual questions (Crouch et al., 2007; Piepmeier Jr., 1998; Suchman et al., 2006) evaluated the peer instruction methodology and used the response data to assess student learning. These studies focused on the instructional methodology rather than variations in student responses to individual questions or groups of questions. They also did not discuss possible differences in student responses based on student demographics.

*Studies that report individual student responses to conceptest questions.*

Only three published articles presented student response data plotted for individual students, yet none of these studies used student responses as a significant data source. Besides the histograms described above, Lucas (2007) also plotted an aggregate student score from all conceptest questions versus course grade and calculated a positive correlation ($r=0.57$) between the two variables.

(Mazur, 1997b) described how to implement the peer instruction methodology and used student response data to illustrate the effectiveness of this pedagogy. For each student, Mazur plotted the percentage of correct responses of questions asked before versus after the discussion. The plot indicated that students answered a larger percentage after discussion than before discussion. The text did not report the actual numbers associated with these graphs, rather the reader was left to qualitatively interpret the graph. Similarly, Rosenberg et al. (2006) used the same graph to present data from their study.
indicating that a higher percentage of students correctly answered conceptest questions after peer discussion.

The remaining study (Crouch & Mazur, 2001) reported student response data as a graph showing the percentages of students who answered incorrectly before group discussion but correctly after discussion. Yet they did not analyze any differences in responses produced by individual students.

These studies were all based on results from Mazur’s Harvard Physics Education Research Group. The results described above presented response data at the student level, but did not provide a detailed examination of variations between students. In addition, these studies did not consider possible variations in responses between questions.

**Summary of previous studies that report student responses to conceptest questions.**

All of the studies described above either ignored the student response data associated with the peer instruction methodology, or used the results to illustrate effective peer instruction. In most cases the response data did not take into account possible differences in student responses between groups of questions. They also did not discuss possible differences in responses between different student populations or between courses taught by different instructors. Of the 81 publications reviewed for this study, only Gray et al. (in press) and King and Joshi (2007, described below) used student response data to evaluate a research hypothesis. Given the paucity of reported data on possible variations in student responses between questions or students, it is not known whether student responses vary according to how the question is worded or whether different student populations provide different responses.
Conceptest Questions and Gender

As shown above, most of the research studies that included conceptest questions focused on overall student achievement throughout the entire course. These studies often administered a diagnostic exam as a pretest and posttest. For example, Crouch and Mazur (2001) used the Force Concept Inventory as a pre- and posttest instrument to assess gains in student comprehension. Similarly, Mayer (2009) used final exam grades to evaluate student comprehension in a psychology course. Most studies that included student responses to conceptest questions did not report separate scores by gender but rather considered the class as a whole (such as Piepmeier Jr., 1998; Suchman et al., 2006). Those studies that included gender differences did not connect gender with student responses to conceptest questions.

Gender, peer instruction, and course achievement.

Peer instruction and conceptest questions have been observed to reduce an observed gender achievement gap. Lorenzo et al. (2006) described the impact of peer instruction on course achievement for men and women who took an undergraduate physics course. The authors used data collected over an eight-year period spanning the introduction and development of the peer instruction model. Results from the Force Concept Inventory suggested that peer instruction raised the achievement levels of both men and women, and it closed an observed pretest gender gap. Lorenzo et al. observed that men significantly outscored women at the beginning of the term but by the end of the term that gap had closed and scores by both genders were no longer significantly different. In addition, the percentage of women who failed the course dropped from 43% to zero percent. A similar study by Knight and Wood (2005) found that both men and women
experienced the same normalized learning gains after experiencing an interactive learning environment that included conceptest questions. These results suggested that using an interactive pedagogy such as peer instruction and conceptest questions engaged all students and promoted academic success in the female as well as male students.

**Gender and student attitudes towards using personal response systems.**

Recent papers from two research groups explored attitudes by men and women towards using a personal response system. The conflicting results illustrate the complexity of this issue and the need for further research in this area.

Two papers by MacGeorge and colleagues (MacGeorge et al., 2008a; 2008b) evaluated student attitudes towards using personal response systems in one of three introductory liberal arts courses. MacGeorge et al. (2008a) administered a custom audience response attitude survey to 703 students and found that women liked using the technology more than men. MacGeorge et al. (2008b) used a repeated measures design to analyze three administrations of the same questionnaire to 854 undergraduate students from the same three courses. This larger, more comprehensive study found no significant difference in men and women’s attitudes towards the technology. Their post hoc analyses suggested that student attitudes might vary by instructor, but the authors dismissed this conclusion by noting that their data did not tie student attitudes to instructional use of a personal response system.

Another recent study by Kay (2009) examined differences in attitudes held by 659 high school students from grades nine through twelve and yielded results that differ from MacGeorge et al.’s (2008a,b) conclusions. Kay’s survey explored student attitudes towards the personal response technology itself, learning using this technology, and
involvement in the class activities. The survey also assessed student perceptions towards computers in general. The results suggested that men gave higher ratings for nearly all measures including overall attitude, involvement, assessment, and learning. Kay confirmed these findings by analyzing the 714 comments students wrote on the survey. For these data, men and women gave similar responses except for references to thinking \((p < 0.005, d = 0.77)\), learning performance \((p < 0.005, d = 0.67)\), and references to the technology \((p < 0.001, d = 1.6)\). For these three attitudes, men provided more positive responses than the women. In addition, women predominately stressed negative aspects of the technology whereas men stressed positive aspects \((p < 0.001, d = 2.2)\). Further analysis found that these differences were removed if student comfort level towards using a computer was added as a covariate. From this analysis, Kay concluded that gender differences towards using computers may have interfered with student use of a personal response system. Yet it should be noted that Kay’s students were secondary students who only experienced a personal response system a handful of times over the course of a semester whereas the students in MacGeorge et al.’s studies were undergraduate students who answered one or more questions every class session. Kay’s participants did not experience a systematic, active learning environment such as the one described in my study, so any comparisons between the two projects must consider this key difference.

*Gender and responses to conceptest questions.*

Other studies have reported gender differences in how students answer conceptest questions. Gonzalez-Espada and Bullock (2007) compared the response time needed for students to answer a conceptest question to several independent variables including course grade, final exam score, and gender. The results suggested that response times for
men (r = -0.59, p = 0.077) were more strongly correlated to their final course scores than response times for women (r = -0.35, p = 0.186). The negative correlations suggested that students who took longer to answer conceptest questions earned lower grades than students who answered quickly and that this relationship was stronger for men than women. The authors did not report whether they observed any achievement differences based on gender. Their results must be viewed as preliminary because their small sample size (10 men and 17 women) greatly reduced the power of the study.

A more recent study investigated gender differences in answering conceptest questions (King & Joshi, 2007). Their participants included 715 engineering students taking a general chemistry class for engineers. Even though the study investigated differences in how men and women responded to conceptest questions, only 17% of the students were women. The data primarily consisted of the percentage of questions that students answered and were similar to Gonzalez-Espada and Bullock’s (2007) response rates. Students were considered to be “active participants” if they answered >75% of the questions asked during a given semester. They found that when the students were not graded for answering the conceptest questions, women answered a significantly higher percentage of conceptest questions than men (62% versus 48%), but when students were awarded participation points for answering the questions (whether correct or incorrect), the participation rates were not significantly different (women = 64% versus men = 54%) even though women continued to participate more often than men. They also observed that men who answered at least 75% of the conceptest questions earned 9-10% higher course grades than men who did not participate. A smaller, but non-significant difference was observed for women (4-5%). Overall, there was no difference in course performance
between men and women. Furthermore, students who answered fewer than 75% of the questions achieved similar grades as students who experienced a traditional didactic lecture but students who answered more than 75% of the conceptest questions scored 11 to 13 percentage points higher on the final exam than students who did not use a personal response system.

King and Joshi (2007) also found that the percentage of questions answered (regardless of whether the answer was correct or incorrect) by men significantly correlated to other measures of course performance such as scores on the final exam, other exams, recitation, labs, online homework, and course grade. By contrast, female participation rates did not significantly correlate (at a 99% confidence limit) to any of the other variables. The authors did not address why these differences were observed, which leaves room for further study.

No other studies were identified that addressed the issue of gender differences in answering conceptest questions. Results from studies like Lorenzo et al. (2006) suggested that using conceptest questions as part of an interactive pedagogy may have reduced or eliminated any observed gender differences in course achievement. Gonzalez-Espada and Bullock (2007) reported differences in how quickly men and women answered conceptest questions, yet their data must be cautiously interpreted because of their small sample size. The results by King and Joshi (2007) also suggested that peer instruction and conceptest questions may promote equitable learning between men and women, yet their sample may be biased given that the participants were engineering students with a prior interest and ability in science. The students observed for my study more closely mirrored the
general student population and better represented the student body typical of many introductory courses.

**Conceptest Questions and Prior Achievement**

Many of the studies described above (e.g. Crouch & Mazur, 2001; Mayer et al., 2009) attempted to connect the use of personal response systems to improved student achievement on a course exam or improved grades for the course. Few of the studies identified for this literature review discussed the role of student prior achievement in responding to conceptest questions. A few studies noted below used SAT scores as a measure of student aptitude, but their use of SAT scores was similar to this study’s use of ACT scores to measure prior achievement.

MacGeorge et al.’s (2008a) study included student composite SAT scores as a measure of student aptitude. They found that aptitude was a weak influence on student attitudes towards personal response systems and accounted for less than 2% of the variance. In addition, students with lower SAT scores reported a greater positive influence of personal response systems on their motivation for the course. These findings suggested that instructors need not worry about student opinions from any aptitude “strata” and that this technology may be beneficial in helping underachieving students succeed in class.

Similarly, Freeman et al. (2007) used SAT verbal and math scores as variables in a logistic regression analysis to identify at-risk students and predict the final course grade. They found that only the SAT-Verbal score and student undergraduate grade point average were needed to predict student course grades ($R^2 = 0.58$). These findings
suggested that standardized measures of student verbal ability or achievement (i.e., the ACT-Reading test) may explain student in-class achievement on conceptest questions.

Only one study (Gonzalez-Espada & Bullock, 2007) compared student ACT scores to student answers on conceptest questions. These authors compared student response times to several variables including three that measured student prior achievement (ACT- math, ACT-composite, grade point average). As will be discussed in a later section, the ACT assesses student achievement in four domains (math, science, reading, English). A fifth overall score called the ACT-composite score is calculated using the mean of the four domain tests. All five scored are reported using a 0-36 scale. Gonzalez-Espada and Bullock compared the student response time to the course conceptest questions answered during an undergraduate physical science course (n = 27). The results were correlated against a suite of variables including ACT-composite scores, ACT-math score, and student grade point average (GPA). Their results suggested that student response times were negatively correlated to the student ACT score (r = -0.50, p = 0.041). The authors did not state which ACT test they considered an “ACT score”, but given that they later reported data from the ACT-math test, the absence of a qualifier suggests that the results are from the ACT-composite test. No significant correlation was found between response time and student GPA or ACT-math-score. The authors also compared student response times on “easy” and “difficult” conceptest questions. This designation was determined using the ratio of the total number of correct responses and the total number of students who attempted the question; however, they do not specify the numeric value that defined the cut-off point delineating the two categories. Their analysis suggested that there was no correlation between student GPA and their response items to easy questions (r = 0.02;
p = 0.926), yet there was a significant correlation between response time and student GPA for difficult questions (r = -0.401; p = 0.038). These results suggested that the difficult questions may have been more discriminating in determining student answers than the easy questions. As noted earlier, this study suffered from a small sample size and failed to define the criteria by which questions were categorized as easy or difficult. Yet, the presence of a significant correlation between student ACT score and response time suggested that a similar pattern might be observed between ACT scores and student responses to conceptest questions.

*Conceptest Questions and Self-Efficacy*

Self-efficacy is an individual’s belief in their capability to perform a given task (Bandura, 1986) and was developed from a social cognitive perspective which argues that individuals develop new knowledge within a social environment. Furthermore, Bandura claimed that self-efficacy may be a better predictor of academic success than objective assessment of student ability. That is, students who believe they can succeed at a task (such as passing an earth science course) maybe motivated to perform well on that task (Zimmerman, 2000). Zimmerman and Kitsantas (2005) studied the self-efficacy of 179 girls from a parochial high school with a curriculum that emphasized homework. A path analysis showed that student self-efficacy significantly (p < 0.05) moderated their homework experiences and influenced their eventual grade point average, thus demonstrating that student achievement is likely influenced by their personal beliefs in their ability to succeed. Very few studies have discussed student self-efficacy as it pertains to using a personal response system or answering conceptest questions but other studies outline the possible role of self-efficacy in science achievement.
Fencl and Scheel (2005) compared the self-efficacy of 218 undergraduate physics students to ten different teaching styles. Their results indicated that improved student self-efficacy significantly correlated with answering in-class questions ($r = 0.33$), using electronic applications ($r = 0.27$) and completing conceptual problem assignments ($r = 0.24$). These three active learning strategies are all related to personal response systems or conceptest questions.

Kay (2008) conducted a meta-analysis of the literature on gender differences in attitudes towards using computers. In that study, Kay found that estimates of student ability to use computers remained stable from 1997 to 2007 with 45% of the studies favoring men, 10% favoring women, and 45% showing no difference. He also noted that men tended to give higher ratings of their abilities than women. His analysis found an insufficient number of studies to dissect self-efficacy, but those he did find suggested that men scored higher on measures of self-efficacy, but Kay warned that these findings may be indicative of differences in how men and women rate themselves rather than actual differences in self efficacy.

In a related study, Kay (2009) included self-efficacy as an explanatory variable when investigating gender attitudes towards personal response systems. In that study, student attitudes correlated with a measure of student comfort in using computers. Kay suggested that his measure of student comfort served as a proxy indicator for student self-efficacy for using computers. In other words, student self-efficacy towards using computers might influence their attitudes towards using a personal response system such that students with low measures of self-efficacy would possess negative attitudes towards the technology
and which in turn would lower student performance on the associated conceptest questions.

Conceptest Questions and Earth Science

Conceptest questions and peer instruction were first developed in undergraduate physics courses at Harvard University (Mazur, 1997a; Crouch & Mazur, 2001). Instructors have since adapted this pedagogy to earth science. Greer and Heaney (2004) presented conceptest questions that were successfully implemented in an earth science course, and McConnell et al. (2003) described several ways to use inquiry-based learning in an earth science course. They also highlighted conceptest questions and personal response systems as one of the successful means for improving student comprehension. In a later paper, McConnell et al. (2006) described the results of implementing peer instruction and personal response systems from nine colleges and universities. Since then, a number of earth science conceptest questions have been written and collected in an online database (SERC, 2008). Many of these questions were used by Gray et al. (in press) to evaluate student comprehension after manipulating a physical model.

Types of Conceptest Questions

A review of conceptest questions reported by Mazur (1997a), Piepmeier Jr. (1998), Crouch et al. (2007), Knight and Wood (2005), and McConnell et al. (2006) revealed that some conceptest questions used conceptual illustrations to support student understanding of the question. Interpreting these diagrams, graphs, photos, or maps may have been necessary in order to comprehend the question. In contrast, other conceptest questions contained only text and did not require any type of graphic illustration. As noted above, geoscience is a highly visual discipline and earth science instructors often use visual aids
to facilitate learning. Research from cognitive psychology on multimedia learning is reviewed below, but results from past studied have suggested that the presence of an illustration can improve student learning (Mayer, 2001). Therefore it seems likely that student responses to conceptest questions may be influenced by the inclusion or absence of an illustration. This study operationally defined two types of conceptest questions. Text-only questions contained text and did not include a graphic of any type. Illustrated questions contained text and a diagram, graph, photo, or map. Inclusion of the illustration was considered to be vital in determining the answer to the conceptest question.

Summary

Peer instruction and its associated conceptest questions are one method of incorporating an active-learning pedagogy into a large, undergraduate course. Much of the literature on conceptest questions did not discuss the questions themselves but rather assessed either the peer instruction methodology or the introduction of a personal response system into an undergraduate course. Other studies focused on student attitudes towards using a personal response system. Student learning gains were typically measured using in-course exams or independent measures such as the Force Concept Inventory. Even those studies that reported individual student responses to conceptest questions (e.g. King & Joshi, 2007; Lucas, 2007; Piepmeier Jr., 1998) did not consider variations in student responses based on how the question was asked. Other research suggests that men and women may hold different attitudes towards using this technology and these differences may be related to differing levels of self-efficacy reported for men and women. In addition, there have been no conclusive studies identifying any possible gender differences in student responses to conceptest questions. In the geosciences,
handful of studies have documented the use of conceptest questions in a large, earth science course. Lastly, little research has compared student responses to conceptest questions and prior academic achievement. Those studies that have investigated prior achievement have tended to use the SAT test rather than the ACT.

Research on the American College Test (ACT)

The American College Test (ACT) is a standardized academic assessment test typically taken by high school seniors and accepted as an admissions test by colleges and universities in the United States (ACT, 2008a). It was first administered in the fall of 1959 and used in all 50 states by 1960. The ACT consists of four separate tests covering science, math, reading, and English with an optional writing test. The four subject-specific tests are scored on a scale of 1 to 36. A fifth score called the ACT-composite score is the mean of the four subtests and is intended to measure overall student achievement across the academic spectrum. For each of the tests, the national average for 2007 (ACT, 2008b) clustered near a score of 21 (math = 21.0, reading = 21.5, science = 21, English = 20.1, composite = 21.2). The publishers of the ACT stated that the tests measured student academic achievement and were aligned with the material covered in the standard high school college preparatory curriculum. Colleges and universities typically use the test results as a general measure of a student’s ability to complete college-level work (ACT, 2008a).

This study used student ACT scores as a proxy for prior student prior achievement in reading, math, and science. This section describes prior research that used the ACT in a similar manner and discusses the claim of gender differences in student scores. Chapter 3
contains a detailed description of this instrument including an assessment of its validity and reliability.

Research Using the ACT

Few studies have been identified that used the ACT to compare concept test questions. As discussed in an earlier section, Gonzalez-Espada and Bullock (2007) compared student ACT scores with the average time needed to answer a concept test question. Of the 27 students that participated in that study, 39% of the students had ACT-composite scores below 18 and were considered in need of academic assistance by the home institution. Similarly, 71% of the students had ACT-math scores below 18.

As noted earlier, this study contained some noticeable flaws. The study drew conclusions based on a small sample size, and the authors did not describe the construct measured by the student ACT scores. They also did not make the case for why ACT scores would be related to student response times. After observing that the ACT-composite score was significantly and negatively correlated to response times ($r = -0.502, p = 0.041$) and final course grade ($r = 0.718, p = 0.002$), the authors suggested that response time might be related to student processing speed or test taking ability even though they presented no additional evidence to support this claim.

A review of the extant literature revealed that most studies which used standardized testing data analyzed data from the rival Scholastic Achievement Test (SAT) published by the Educational Testing Service (2009). Three studies are reviewed to illustrate how past researchers have used ACT scores to define participant groupings. Knight (1994) considered the role of dictionary use in the acquiring new vocabulary. The participants consisted of 105 native English speaking students taking an undergraduate Spanish. For
the study, high- and low-verbal ability groups were defined using student scores on the ACT. The results suggested that student scores on four measures of vocabulary acquisition were significantly different for the two ability groups (p-values ranged from \( p = 0.049 \) to \( p < 0.001 \)).

In another study, Martino and Hoffman (2002) used student ACT-reading scores in a manner similar to this study. They compared the reading ability of 26 African-American college freshmen against a suite of reading comprehension strategies. The authors determined the ability groups using student ACT-reading scores. The high reading ability group had ACT-reading scores between 26 and 32, whereas the low reading ability group had ACT-reading scores between 14 and 20. They used these categories to correlate reading ability against eight measures of language ability. Five of the eight measures significantly correlated with reading ability as measured by the ACT-reading test. Martino and Hoffman concluded that an integrated approach to reading best serves students with low reading abilities.

Lastly, Cole and Osterlind (2008) used student ACT scores as a measure of prior academic achievement. In that study, the authors compared the results from both high- and low-stakes administration of a general education test (the College base test). The results indicated a significant difference in student scores on this measure and student ACT scores (\( F = 655.34, \ p < .001 \)). The authors also used student ACT scores as a covariate in subsequent analyses.

In all three studies, student ACT scores were used to define student achievement or ability groupings. In the case of Cole and Osterlind (2008), ACT scores were also used as
a covariate. This study similarly used student ACT scores to define student prior achievement in science, math, and reading.

**Gender Differences on the ACT**

The literature contained several references to a known gender bias in the ACT tests, yet when examined more closely, the literature seems to be incomplete on the topic. A search using Google-Scholar (Google, 2008) with keywords “gender differences” and “ACT” yielded over 29,000 responses, and a similar search with the terms “American College Test” and “gender differences” yielded 56 responses. Yet nearly all of these studies did not evaluate any gender biases in ACT scores but rather just commented that such a bias exists. Even papers on the subject of gender differences on standardized tests or gender differences in cognitive ability tended to report data from SAT (ETS, 2009). For example, Benbow, Lubinski, Shea, and Eftekari-Sanjani (2000) discussed data from a longitudinal study on gender differences in young adults. As part of their analysis, they commented that gender differences have been observed on the ACT-math test and they cited an earlier paper (Benbow & Stanley, 1996) as the data source. Yet a review of that study yielded only references to the SAT test. A similar study by Casey, Nuttal, and Benbow (1995) compared student spatial ability to mentally rotate images with gender and SAT scores. Of the 274 participants, only 57 reported ACT scores, so the authors converted the ACT data to SAT equivalent scores. Following such a procedure eliminated any comparison of ACT scores to gender or spatial ability.

Two studies showed an observed gender bias in the ACT. Mau and Lynn (2001) analyzed standardized test and grade point average data from a nation-wide, representative sampling of college graduates (n = 1572). They found that, on average,
men scored a significant ($F = 47.39$, $p < 0.001$, $d = 0.23$) 1.16 points higher on the ACT-composite test than their female counterparts. Similar results were obtained for the SAT-math and SAT-verbal tests. The authors also noted that women earned significantly higher GPAs throughout their undergraduate career (men = 3.12, women = 3.25; $F = 235.52$, $p < 0.001$, $d = 0.30$).

Rodriguez (1997) analyzed student ACT scores from 1991-1995 for any trends or comparisons between groups. His analysis revealed that females continued to score below the national mean. Scores for Asian and European women (two racial populations that tend to score high on the ACT) were at the national mean. Rodriguez echoed the observations by Halpern et al. (2007) that scores for men were more widely distributed than scores for women, which leads to a larger percentage of men in both the upper and lower ranges of the test.

Recent analyses of ACT data suggested that the observed gender gap may be determined by the number of students who take the test rather than an innate difference between the genders. One report (ACT, 2005) evaluated trends in the observed gender gap for the ACT test results. The 2005 report by ACT on gender fairness reports mean ACT-composite scores for men and women from 1997 through 2004. The observed gender gap was small and appeared to have narrowed from a high of 0.3 in 1997 to a low of 0.1 in 2004. The authors of this report also noted that the students who took the ACT were self-selected and planned on attending college, which suggested that biases in the student sample population may contribute to the observed gender gap. In 2002, two states (Colorado and Illinois) required all high school graduates to take the ACT. Data from these two states indicated that before 2002, a gender gap of 0.2 points was observed on
the ACT-composite scores, but starting in 2002 the differences in ACT scores dropped to 0.2 and -0.1 points. The negative gender difference indicated that women scored higher than men. In addition, high school GPAs from these two states yielded similar results observed by Mau and Lynn (2001) with women earning higher GPAs than men (Colorado = 3.15 for women versus 2.91 for men; Illinois = 3.05 for women versus 2.86 for men). These results suggested that the observed gender gap may be due to cultural or societal pressures to take the exam rather than deficiencies in the test. Further research is needed to identify those factors that influence men and women to attain a high level of academic achievement.

Differences in Verbal and Spatial Ability

Like data from the ACT, gender differences have been reported for research on student cognitive and spatial abilities. In particular, most studies suggested that women scored higher on measures of verbal ability and men scored higher on measures of spatial and quantitative ability. Several recent studies have highlighted these differences and provided possible explanations for these observations.

*Gender Differences In Verbal And Quantitative Reasoning*

Strand, Deary, and Smith (2006) reported cognitive data from a large, nationally representative, sample (n = 320,000) of 11 and 12-year old students. This study used the UK version of the Cognitive Abilities Test which is a standardized instrument that assessed student verbal reasoning, quantitative reasoning, and non-verbal reasoning. Their results indicated that girls (mean = 100.6) scored significantly higher than boys (mean = 98.4) on verbal reasoning (p<0.0001, d=0.15) but boys scored significantly higher than girls on measures of quantitative reasoning (100.2 and 99.4 respectively, p <
Their data also showed that boys comprised a larger percentage of students at both ends of the distribution for quantitative and non-verbal reasoning but girls comprise a larger percentage of students in the upper 10% of the scores on verbal reasoning. The very low effect sizes (d-values below 0.2) indicated that even though the mean differences were statistically significant, 97% of the data overlapped between the two genders.

These results were consistent with the findings of Halpern et al.'s (2007) literature review of gender differences in math and science achievement. After reviewing studies relevant to this topic, the authors concluded that gender differences do exist with women (on average) being more verbal than men. They also reviewed the literature on gender differences in spatial ability and noted that men outperformed women on measures of mental rotation with moderately large effect sizes ranging from $d = 0.63$ to $d = 0.77$.

**Gender Differences In Spatial Ability**

Several studies have investigated gender differences and spatial ability. Linn and Peterson (1985) conducted a meta-analysis of the existing literature on gender differences in spatial ability. They reviewed 172 effect sizes from studies on participants of all ages. They found that the largest gender difference for spatial perception was between adult men and adult women (effect size = 0.64), but they were unable to determine whether the observed gender difference was due to cognitive changes in adults, a cohort effect, or sampling bias. For mental rotations, Linn and Peterson also found a moderately large gender difference favoring men (effect size = 0.73) for participants of all ages (child to adult). In contrast, for spatial visualization, Linn and Peterson consistently found no-significant differences between men and women of all ages (effect size = 0.13). Geary,
Gilger, and Elliot-Miller (1992) replicated the findings of Lin and Peterson by studying 347 undergraduate students and found that men outperformed women on three measures of mental rotation. On average, men outscored women on all three measures with the largest difference on the Vandenberg Mental Rotations Test \( (F_{(1,345)} = 104.38, p < 0.001) \). Their results confirmed Linn and Peterson’s conclusions that a gender difference favoring men on the ability to mentally rotate an image.

More recently, Saucier, McCreary, and Saxberg (2002) investigated whether student socialization of gender roles and gender-specific behavior significantly mediated their ability to mentally rotate an image. The 97 undergraduate students completed the Vandenberg Mental Rotations Test and as well as a survey designed to assess their views towards gender-related personality traits and gender-specific behaviors. The results found a strong gender difference in the mental rotations test \( (F_{(1,93)} = 31.25, p < 0.0001) \) in favor of the men, but the overall results found only limited support for the claim that gender-role socialization or gender-typed behaviors explained or predicted student spatial ability. Gender socialization only explained 6.3% of the total variability in spatial ability and gender-specific behaviors only explained 1.6% of the total variability. The authors tentatively concluded that gender socialization and views of gender-specific behaviors did not greatly influence student spatial ability and suggested that some unidentified trait or construct explains the strong gender difference observed on the mental rotations test.

Massa, Mayer, and Bohon (2005) tested whether the wording of the instructions affected undergraduate student performance on a spatial ability test. The 237 women and 99 men completed a spatial ability test that assessed the students’ ability to discern a geometric shape embedded within a complex design. The students also completed an
inventory designed to measure whether they viewed their gender role as masculine or feminine. The instructions for the spatial test were altered to indicate that the test measured spatial ability (a masculine characteristic) or empathy (a feminine characteristic). The results indicated that the women, but not the men, were influenced by the instructions. For the women, the spatial test scores revealed that the type of instruction and gender role significantly interacted ($F_{(3,229)} = 19.413, p < 0.001$) That is, women who were identified as being feminine scored highest on the test with empathy instructions but women who were identified as masculine scored highest on the test with spatial instructions. A similar pattern was not observed for the men with no significant interaction between instruction type and gender role ($F_{(3,91)} = 0.183, p = 0.908$). That is, men produced nearly the same test scores regardless of how the instructions were worded. This study suggested student beliefs about gender roles may impact their performance on any assessment of cognitive ability and should be considered when explaining observed gender differences on measures of spatial ability.

Nordvik and Amponsah (1998) studied 161 technology students and 293 social science students from a Norwegian technical university. Students completed four tests assessing three different types of spatial ability (mental rotation, spatial visualization, and spatial perception). In addition, the students used a seven-point Likert scale to respond whether they had participated in 38 spatially oriented activities. The results indicated that men scored higher than women on all measures of spatial ability with effect sizes ranging from 1.06 for mental rotation by social science majors to 0.33 for spatial perception.

Hamilton (1998) studied gender differences in high school science achievement by comparing student answers to multiple choice and constructed response questions taken
from the 1988 National Educational Longitudinal Study. She also interviewed 25 students who completed 20 questions taken from the larger battery of questions to identify the cognitive processes sued to answer each question. The largest gender difference from the test results came on questions that measured spatial-mechanical reasoning. For these questions, males outscored the females by more than one-third of a standard deviation even after controlling for prior achievement and courses taken. In addition, physics was the only class to significantly correlate with spatial-mechanical reasoning. Hamilton suggested that this observed gender difference in spatial reasoning was due to differences in student experiences outside of school. The interview data confirmed these observations. When discussing the acquisition of skills pertaining to quantitative science, none of the students listed experiences outside of the classroom, but half of the students reported experiences away from school when describing their acquisition of spatial-mechanical reasoning skills. Similarly, when asked to solve spatially-oriented questions, 15 students used gestures to explain their solution and 13 students reported that playing with equipment such as see-saws or cameras contributed to their spatial-mechanical understanding. Hamilton concluded that success on spatially-oriented tasks heavily depends on student visualization skills learned during childhood play and suggested that hands-on activities might close the achievement gap.

Lastly, Terlicki and Newcombe (2005) considered the role of prior experiences with computers and videogames on student spatial ability. They developed a survey that asked students to self report their experiences in using a computer and playing video games. From the 1278 students who completed the questionnaire, 180 were identified as having high or low spatial experiences when using a computer. These students were given a
mental rotations test. The results indicated a significant differences between gender and spatial experience (F = 169.52, p < 0.01) as well as performance on the spatial test (F = 28.14, p < 0.01). In addition, scores of spatial ability also varied by student computer experience (F = 44.66, p < 0.01). In all cases, the significant differences yielded small effect sizes (η² = 0.12 to 0.20). An additional structural equation model suggested computer experience partially mediated student performance on the mental rotations test. These findings suggest that student prior experiences may be influential in the development of spatial ability. In particular, students who experienced spatially-oriented tasks while using a computer or playing videogames may have enhanced their spatial ability. The authors suggest that men and women may choose to use computers for different tasks with men experiencing a larger number of tasks that foster spatial ability.

*Spatial Ability in Geoscience Education*

Given the visual nature of the geosciences (Clary, 2003; Rudwick, 1976), several researchers have investigated the role of spatial ability in learning geoscience concepts. Ishikawa and Kastens (2005) summarized the extant literature on student spatial ability and described six different mental skills that a student must master in order to become an expert geoscientist including recognizing patterns, mentally rotating an object and mentally manipulating a surface or volume. They then related the research to basic process skills such as reading a topographic map or visualizing in three dimensions and gave 14 implications for how educators could incorporate spatial understanding into their curricula. Kastens and Ishikawa (2006) expanded on this theme and described five general categories of skills used by geoscientists (such as using and making maps or identifying or classifying shapes). Both studies concluded that spatial ability is a primary
component in understanding Earth’s processes and these skills should be fostered in the larger geoscience curriculum.

Orion, Ben-Chaim, and Kali (1997) tested 32 first-year geology undergraduate students to see if student spatial ability changed after completing their first year courses. Their results suggested that completing the introductory curriculum significantly improved the spatial abilities of all the students. Both men (d = 0.55) and women (d = 0.72) improved their scores on a general spatial aptitude test but results from another test that included mental rotations indicated that men significantly improved their scores (d = 1.29), but women showed no significant gain. The results from this study suggested that geoscience courses can improve student’s spatial ability, but also noted that men show larger gains than women. The results from this study further suggested that student spatial ability could influence their performance on earth science concept test questions, but the small sample size limited the generalizability of the results.

Another study described some of the spatial process skills used by geologists. Kastens and Ishikawa (2009) asked undergraduate non-science major (n = 13), science majors (n = 14), and professional scientists (n = 6) to examine 14 wooden outcrops that simulated tilted rock strata that were arranged around the researcher’s campus and describe the overall geologic structure. The results suggested that even in a simplified situation, the students often struggled to understand the underlying orientation of the simulated strata. The non-science majors tended to advance claims without providing supporting details, even if they later demonstrated that they understood the evidence to support their claim. It was also observed that students appeared to reach conclusions based on their
generalized spatial thinking skills rather than their knowledge of geology, which suggests that courses should develop student spatial ability as part of the scientific training.

Lastly, Black (2005) demonstrated the connection between spatial ability and comprehension of geoscience concepts. In her study, 97 non-science majors were given a 16-question earth science concepts test as well as a battery of tests that assessed ability in mental rotation, spatial perception, and spatial visualization. The results indicated that all three forms of spatial ability significantly correlated with comprehension of earth science concepts and mental rotation was the best predictor of student understanding of the same geoscience material. In all, spatial ability explained one-fourth to one-third of the total variation on the earth science knowledge test. This study found no significant differences between gender, earth science knowledge, and spatial ability, though the author noted that the women had taken more science courses and were older than the men which suggested that age and prior science experiences could have confounded the results.

The research described above suggests that comprehending geoscience concepts requires a well-developed sense of spatial ability. In addition, spatial thinking can be developed by earth science courses that purposefully include exercises and experiences that enhance student spatial abilities. The results were mixed concerning gender differences in using spatial ability to solve geologic problems, but the ability to mentally rotate an object may be one factor that leads to any observed gender differences in the comprehension of geoscience concepts.

Summary

The literature supported the claim that students learned best when they experienced an active-learning environment rather than a traditional didactic lecture. Mazur’s (1997a)
peer instruction methodology is one model that has shown to boost student interest a
course and may improve student learning.

Conceptest questions are an integral part of peer instruction, yet few studies have
investigated student answers to these questions. In particular, most studies aggregated
their data to create a single percentage of correct responses and used the results to
augment other learning assessments such as a course exam, course grades, or the Force
Concept Inventory. Potential gender differences in student responses have also been
poorly studied. Preliminary results from a few studies suggested that significant
differences may be present.

This study used student scores from the ACT test to measure prior student
achievement. Literature on standardized tests often reported gender differences for the
ACT, yet few studies have used this test as a measure of student achievement. Instead,
most studies use the SAT. Recent data from ACT suggested that the observed gender gap
may be narrowing and may also be a function of student self-selection rather than an
inherent deficiency in the test.

Research from cognitive psychology suggested that, on average, men outperformed
women on measures of spatial ability but women outperformed men on measures of
verbal ability. Scores by men also exhibited a wider variance than scores for women,
meaning that men are over-represented at the upper and lower tails of the population
distribution. This does not mean that men always outperformed women, rather the
average scores for men exceeded those for women on measures of spatial ability and vice
versa for, measures of verbal ability. Additional research further suggests that
performance on measures of spatial ability may be influenced by student beliefs towards gender roles as well as the way in which the instructions are worded.
CHAPTER III
RESEARCH METHODOLOGY OF THE STUDY

Introduction

This study used an *ex post facto*, causal-comparative research design to investigate differences in the percentage of correct responses to text-only and illustrated concepttest questions when compared to gender or prior achievement. This chapter introduces the research questions and hypotheses and discusses the rationale for using an *ex post facto*, causal-comparative design and also elaborates on how personal response systems were used to collect the data. This study used data from student responses to concepttest questions and compared the results to student gender and student ACT scores. This chapter also discusses the validity and reliability of both concepttest questions and ACT scores, describes the repeated measures, mixed analysis of variance (ANOVA) analytical method, and explains why a repeated measures design was an appropriate choice for this study.

Research Questions and Hypotheses

Prior research on concepttest questions have focused on evaluating student performance in classes that used interactive pedagogies such as peer instruction (Mazur, 1997a). Most of the studies that mention concepttest questions did not include student answers to these questions, rather they used an independent measure such as the Force Concept Inventory (Hestenes et al., 1992) or in course evaluations (Lasry, 2008; Mayer et
al., 2009) to assess student learning over the entire course. Furthermore, most of these studies did not consider possible differences in student responses to different types of questions. The purpose of this study was to assess whether men and women responded similarly to text-only or illustrated concept test questions. In addition, this study also considered whether student responses to both types of concept test questions were influenced by student prior achievement in reading, math, and science. The produced two research questions with five research hypotheses.

Research Question #1: In an undergraduate earth science course, do men and women provide different responses to text-only and illustrated concept test questions?

This question led to the following research hypotheses:

Research Hypothesis #1a. $H_0$: There was no statistically significant difference between the percentage of correct student responses to text-only and illustrated concept test questions that were asked during an undergraduate earth science course.

Research Hypothesis #1b. $H_0$: There was no statistically significant difference between gender and the percentage of correct responses correct student responses to text-only concept test questions and illustrated concept test questions that were asked during an undergraduate earth science course.

Prior studies have also documented that students who performed well on course exams also performed well on concept test questions (King & Joshi, 2007). In addition, gender differences have been observed in assessments of student verbal and spatial abilities with women outscoring men (on average) on measures of verbal ability but men outscoring women on measures of spatial ability (Halpern et al., 2007). It is not known
whether such differences are present in student responses to conceptest questions.

Research Question #2 addresses this issue.

Research Question #2: Do student responses to text-only and illustrated conceptest questions vary as a function of prior achievement as measured by student ACT scores?

Three hypotheses answer this research question:

Research Hypothesis #2a. \( H_0 \): There is no statistically significant difference between the percentage of correct responses to text-only or illustrated conceptest questions asked during an undergraduate earth science course and student scores on the ACT-Reading test.

Research Hypothesis #2b. \( H_0 \): There is no statistically significant difference between the percentage of correct responses to text-only or illustrated conceptest questions asked during an undergraduate earth science course and student scores on the ACT-Math test.

Research Hypothesis #2c. \( H_0 \): There is no statistically significant difference between the percentage of correct responses to text-only or illustrated conceptest questions asked during an undergraduate earth science course and student scores on the ACT-Science test.

Research Design

This study used a causal-comparative, *ex post facto* research design to investigate differences in student responses to conceptest questions. Such a design allowed the use of a large set of data that had been previously collected, and is appropriate in studies that
identify naturally-occurring differences between groups. The study also addressed issues of reliability and validity to ensure that the results would be trustworthy.

**Characteristics of Causal-Comparative Research**

This study compared the relationship between student responses to text-only and illustrated concepttest questions with the independent variables of gender and prior reading, math, and science achievement. To analyze these differences, this study employed an *ex post facto*, causal-comparative research design using responses by undergraduate students enrolled in nine, large, undergraduate, earth science classes. According to Gay and Airasian (2000), causal-comparative studies are designed to “determine the cause, or reason, for preexisting differences in groups of individuals” (p. 349). Gall, Gall, and Borg (2007) also emphasized the use of group differences in defining causal-comparative research. They defined causal-comparative studies as…

*a type of nonexperimental investigation in which researchers seek to identify cause-and-effect relationships by forming groups of individuals in whom the independent variable is present or absent – or present at several levels – and then determining whether the groups differ on the dependent variable” (p. 306).

*Ex post facto* (Latin for ‘after the fact’) designs are a common category of causal-comparative studies in which both the outcome and the alleged cause have already occurred (Gay & Airasian, 2000).

Causal-comparative research also investigates variables that naturally occur in the research setting (Gay & Airasian, 2000), whereas experimental designs rely on random selection of participants to the different independent variables (Gall et al., 2007).

Variables that are commonly used in causal-comparative research include organismic variables (age, gender, ethnicity), ability variables (intelligence, scholastic aptitude), personality variables (self-esteem, learning style), family-related variables (income,
marital status, birth order), and school-related variables (school size, curriculum) (Gay & Airasian, 2000). In addition, causal-comparative research uses categorical independent variables to assess the dependent variable (Gall et al., 2007; Johnson, 2001).

This study met the definitions of both *ex post facto* and causal-comparative research. The students enrolled in the nine earth science classes included in this study self-selected which class they took and the classes were not separated into treatment and control groups. In addition, both gender and student prior achievement were naturally-occurring categorical variables that are not suited for random assignment.

**Threats to Validity**

*Ex post facto*, causal-comparative designs contain some inherent disadvantages that were addressed for this study. Gay and Airasian (2000) noted that causal-comparative research established a relationship between two variables but could not firmly establish a causal link. Similarly, Lee (1985) noted that the four most common threats to the validity of an *ex post facto* design were history, testing, instrumentation, and regression. This study was designed to minimize the risks posed by these threats to validity. For each of the nine classes included in this study, variations in the data due to student differences over time were minimized by analyzing questions asked over a three to five-week period rather than over an entire semester. This time frame began four weeks into the semester and was late enough for students to become comfortable with using the personal response system and answering conceptest questions but sufficiently early to include students who would later drop the course. In addition, all of the responses included in this study came from the first time each question was asked. Limiting the data to questions that students
had not seen before ensured that responses were not due the student recall of a previously asked question.

Threats due to testing and instrumentation are typically concerned with differences in the testing environment as well as irregularities in student use of the testing instrument. This study addressed both of these concerns. Both instructors used the same textbook (*The Good Earth* by McConnell et al., 2008) and covered the same topics in the same sequence. They also had more than eight years experience implementing peer instruction and personal response systems into this course, and used similar assignments both in and out of class. By collaborating together, the instructors created similar learning environments and reduced irregularities due to differences in teaching style. Similarly, threats from testing irregularities were minimized. The instructors asked similar sets of conceptest questions and included the same or similar in-class exercises for the students to complete. To ensure that the responses reflected individual comprehension, this study only investigated student responses to questions that were answered individually before discussing the question with their peers. Problems in collecting the student response data were also minimized. The students had four weeks to purchase and become accustomed to using the electronic transmitters. This early transition period also provided an opportunity for students to become familiar with each instructor’s teaching style including their use of in-class learning activities, group work, and conceptest questions. It is not known whether student responses, over the course of the study, regressed towards a mean.
Research Setting

This study took place within an introductory earth science course offered through the geology department of a large (n = 24,000 students), urban, four-year, public university in the Midwestern United States. The course consisted of a sixteen-week semester with three 50-minute classes per week and no laboratory component. Classes typically began the year with 150 - 160 students, and throughout the semester the students explored concepts pertaining to astronomy, geology, hydrology, oceanography, and meteorology.

Visual Nature of the Class Setting

The course’s physical setting reflected the visual nature of earth science (Clary, 2003; Gobert, 2005; Manduca et al., 2004; Rudwick, 1976) and was designed to educate anyone who traveled through the building. The two stairwells were decorated with full-color posters that formed scale models of the geologic time scale with Earth’s formation (the Hadean Era) in the basement and the Quaternary Epoch at the top of the stairs on the third floor. Each poster illustrated the flora, fauna, and geography of selected geologic eras, periods, or epochs. The main hallway leading to the lecture hall contained built-in cases displaying labeled minerals from local and international locales. Other cases displayed fossils, maps, and text that described life during different geologic time periods. Another large display case contained the fossil remains of a European cave bear mounted in life position. The hallway also contained several maps and posters. One large map highlighted the state’s bedrock geology, and another case contained posters displaying recent publications by the faculty. Nearby, a large computer monitor showed photographs from recent geologic fieldtrips and encouraged potential majors to register for additional geology courses. The visual displays arranged throughout the building
reinforced many of the concepts covered in the earth science or geology courses offered by the department.

The textbook also reflected the visual nature of the discipline. Of the 132 pages devoted to the four chapters included in this study, approximately 56% of the total page area contained a photograph, diagram, graph, or map. The title page for each chapter had a full-page photograph of a feature germane to that chapter. In addition, each chapter contained a two-page painting that illustrates the fundamental concept for that chapter. Only nine pages (<7%) in these four chapters did not contain any illustrations.

The two classrooms in which study took place were designed for passive listening to a lecture and did not encourage the use of active-learning pedagogies. The permanently-mounted seats were configured in fixed rows. In one of the class rooms, the seats were arranged in tiers whereas the seats in the other room are all mounted on the same level. A large screen hung from the front of both rooms, on which the instructors projected their lecture slides. During the first week of each course, the students were assigned a seat location and placed within a three or four-person work group. The students remained in these groups for the entire semester. Once or twice throughout the semester, the instructors rearranged the seating chart so students in the back of the room were given an opportunity to sit closer to the front.

Outline of a Typical Class Session

In a typical class, the students arrived and found their assigned seats. A teaching assistant would stand at the door and hand out any worksheets that the students would use during that class session as well as a handout for that day’s lecture. The handouts contained a printout of the slides from the lecture that described the basic concepts under
review but did not contain any conceptest questions. The first slide from the instructor’s PowerPoint presentation would already be displayed on the screen at the front of the room as well as a prompt for each student to sign in to the class by registering their personal response system transmitters. At the designated time the instructor began talking.

A typical lecture contained approximately three segments separated by student responses to conceptest questions or completion of an activity within their assigned groups. The students were informed at the beginning of the semester that they were expected to read the assigned chapters before coming to class and that class time was devoted to elaboration on the primary concepts from the reading. Periodically throughout the semester, the instructor reminded the students that attending class and reading the book before coming to class were two important study skills that affected the student’s final grade. Vocabulary was rarely discussed during a lecture.

The instructor began the lecture by reconnecting to the concepts from the previous lesson and moved on to new material. This review might include one or two pretest conceptest questions to assess student prior knowledge of the material. The PowerPoint slides rarely contained bulleted lists but instead displayed photographs, diagrams, or graphs that described the underlying concept under review accompanied by explanatory text. These illustrations were often taken from the book. The explanations provided by the instructor might include opportunities for the students to interact with data displayed on the screen, observe a demonstration, or physically interact with a student-manipulated model. After lecturing for approximately 15 minutes, the instructor asked a conceptest question and instructed the students to respond without talking to anyone. The results
were then displayed as a histogram showing the percentage of students who chose each possible answer. If more the approximately 75% of the students responded correctly, the instructor shifted to the next phase in the lecture. If less than approximately 75% of the students correctly responded, the instructor asked the students to discuss the question among their groups. At this point the students would discuss the question, and the instructor circulated around the room to hear what students were discussing. After approximately one or two minutes of discussion, the instructor had the students answer the same conceptest question a second time. The results were similarly displayed as a histogram and the instructor might discuss the reasons for the correct answer before moving on to the next phase in the lecture.

An example from a class session on earthquakes illustrates this dynamic, interactive teaching style. At the start of class the instructor reminded the students of an upcoming exam and began his lecture by asking how many students had experienced an earthquake. The students responded using their transmitters. After discussing the student responses to the question, the instructor discussed the three major types of faults (normal, reverse, strike-slip) by showing three photographs and explained how the earth moved along each fault. Approximately ten minutes into the class, the instructor asked the first conceptest question which required students to determine the location of an epicenter based on a written description of an earthquake. A majority of the students answered correctly, so the instructor did not ask the students to discuss the question.

Over the next ten minutes the instructor showed three conceptual drawings and four additional photographs illustrating the three types of faults and observations needed to
determine the direction of offset. He also asked two conceptest questions that assessed student ability to determine the type of fault shown in an illustration.

The instructor then demonstrated many of the physical features found in earthquakes using a machine that pulled a brick across a table. This demonstration took the remainder of the class session. Throughout the demonstration, the instructor asked the students to make predictions on what they would observe, and related what happened to the concepts covered earlier that morning. Three conceptest questions were asked during the demonstration and each question focused on student understanding of the forces that produce earthquakes. The instructor ended the lecture by introducing earthquake prediction and asking a final conceptest question.

The frequent use of conceptest questions and student discussions served to break the lecture into small chunks of time that were focused on a single concept. The use of an interactive demonstration as an analogue to an earthquake provided a shared experience that the students could refer back to when studying for the exam. Using conceptest questions reinforced the key learning objectives for the lecture. In addition, the frequent use of illustrations fostered student learning by providing a second channel through which the students could process the information (see the multimedia principle of Mayer, 2001).

In addition to using active-learning pedagogies, both instructors conducted research on their classes and their teaching. Data for this project were collected as part of a study funded by the National Science Foundation to investigate whether the use of student-manipulated, physical models in the class would enhance student learning. As part of that project, students signed a letter of consent approved by the university’s Institutional
Review Board (Appendix A). Their consent included the archiving of personal response data and retrieval of student demographic data through the university’s office for institutional research.

Research Participants

The participants in this study were undergraduate students taking one of nine introductory earth science classes taught between the fall of 2006 and the spring of 2008. For many students, this earth science class represented the first undergraduate science course that they had taken. In these classes, a majority of the students are European-American, freshmen, males, under the age of 26, who have not yet declared a major but plan on earning a bachelor’s of science degree (Table 2 and Appendix B). Besides a traditional undergraduate curriculum, the university also contains a technical college (Summit College) that offers associate and baccalaureate certificate programs that are more commonly found in a community college. Students from this school take general education courses, including earth science, as part of the larger university community.

Data Collection

The data for this study were collected from nine earth science classes taught between the fall of 2006 and spring of 2008. The instructors of these classes created an active-learning environment using a modified version of Mazur’s (1997a) peer instruction as a basis for creating a student-centered, active-learning environment. The classes were taught by two instructors with over eight year’s experience incorporating active-learning pedagogies into large, introductory earth science courses. At the beginning of every semester, each instructor assigned the students to three or four-person work groups and placed each group in a permanent seating location. During a typical class session,
students individually answered one to six conceptest questions. Unless a sufficient percentage (>75%) of students answered the question correctly, the instructor asked the groups to discuss the question before answering a second time.

Table 2. Student demographics for the nine earth science classes included in this study

<table>
<thead>
<tr>
<th>Gender</th>
<th>n</th>
<th>%</th>
<th>Race</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>303</td>
<td>54.0</td>
<td>White</td>
<td>478</td>
<td>85.2</td>
</tr>
<tr>
<td>Women</td>
<td>258</td>
<td>46.0</td>
<td>African American</td>
<td>58</td>
<td>10.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Asian/Pacific Islander</td>
<td>4</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hispanic</td>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Unknown/NR Alien</td>
<td>18</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>All Minority</td>
<td>65</td>
<td>11.6</td>
</tr>
<tr>
<td>Current College</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University College</td>
<td>354</td>
<td>63.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summit College</td>
<td>55</td>
<td>9.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Business</td>
<td>48</td>
<td>8.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fine &amp; Applied Arts</td>
<td>42</td>
<td>7.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arts &amp; Science</td>
<td>42</td>
<td>7.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>11</td>
<td>2.0</td>
<td>BS</td>
<td>368</td>
<td>65.6</td>
</tr>
<tr>
<td>Engineering</td>
<td>8</td>
<td>1.4</td>
<td>BA</td>
<td>149</td>
<td>26.6</td>
</tr>
<tr>
<td>Nursing</td>
<td>1</td>
<td>0.2</td>
<td>BFA (Fine Arts)</td>
<td>24</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AA</td>
<td>11</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BM (Music)</td>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Professional</td>
<td>6</td>
<td>1.1</td>
</tr>
</tbody>
</table>

For the purposes of this study, pseudonyms were used when referring to the instructors. At the start of this study, Dr. Anderson had taught at this university for 10 years and had published research on geoscience education for seven years. Dr. Baker had taught at this institution for 16 years and had published research pertaining to geoscience education for 10 years. Both instructors collaborated on developing the materials used in this earth science course including the conceptest questions, in-class exercises, homework
assignments, and assessments. Both men conducted educational research in their courses and used the results to improve their teaching.

Given that this study investigated differences in how individuals responded to conceptest questions, only responses to conceptest questions provided by individual students were included in the final database. Responses given after group discussions were not included in this study. This determination was made by observing the times and sequences of the questions asked during a given class session. If a question appeared only one time it was classified as a question answered individually and included in the final database. If a question was asked a second time later that class session or in any subsequent class session, it was not included in the final database. If the instructor asked the same question at the beginning and end of a class session (as a pretest and post-test), only the initial, pretest question was included in the database for further analysis.

Data on student gender and prior achievement were gathered from demographic information obtained through the university’s Office of Institutional Research. For the purposes of this study, student prior achievement was measured using results from three of the five scores (reading, math, and science) from the American College Test (ACT, 2008a). The ACT-Reading test asked the students to read a short passage and answer comprehension questions concerning the passage. This assessment process mirrored the answering of text-only conceptest questions because both involved the reading, decoding, and interpretation of text as it pertained to a given, multiple choice test question. The ACT-Math and ACT-Science scores were included in this study because the skills included in these tests (interpreting graphs, reading diagrams and/or maps, or assessing scientific knowledge) were hypothesized to be vital in correctly answering illustrated
conceptest questions. One must keep in mind that the figures in illustrated conceptest
questions required student interpretation before the correct answer could be determined.
For example, if a conceptest question asked for the number of tectonic plates displayed
on a map, the student had to properly interpret the map and map symbols before
answering the question.

Following the example set by ACT (2008b), this study divided the student scores into
four quartiles using the cut-off values from ACT’s national report on student scores from
2007 and included results from 1,300,599 students. Table 3 lists the cut-off values for
each quartile and test. The lowest quartile is operationally defined as the low achievement
group, and subsequent quartiles are named the low-average, high-average, and high
achievement groups.

Table 3. Ranges of ACT Scores for each Quartile Defined in this
Study

<table>
<thead>
<tr>
<th>Achievement Level</th>
<th>Science</th>
<th>Math</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Achievement^a</td>
<td>24 - 36</td>
<td>25 - 36</td>
<td>26 - 36</td>
</tr>
<tr>
<td>High-Average Achievement^a</td>
<td>21 - 23</td>
<td>20 - 24</td>
<td>21 - 25</td>
</tr>
<tr>
<td>Low-Average Achievement^a</td>
<td>18 - 20</td>
<td>17 - 19</td>
<td>17 - 20</td>
</tr>
<tr>
<td>Low Achievement^a</td>
<td>1 - 17</td>
<td>1 - 16</td>
<td>1 - 16</td>
</tr>
</tbody>
</table>

^a Based on data from ACT (2008c) on students who graduated in 2007

Measurement

For this study, both conceptest questions and ACT scores were used as instruments to
measure student performance. This section describes each instrument and discusses why
each was appropriate for this study.

Conceptest Questions

Conceptest questions are conceptually-based, multiple-choice questions asked as part
of the peer instruction pedagogy (Mazur, 1997a). After lecturing for a short time period
(approximately 10-20 minutes), the instructor uses a conceptest question to assess student comprehension of the material. Students typically answer individually before discussing the questions with their peers. Answers are typically collected using an electronic personal response system.

Conceptest questions were originally designed for an introductory physics class at Harvard University (Mazur, 1997a). The questions investigated in this study addressed concepts from earth science, and many of the questions were written by the two instructors from this study and described in McConnell et al. (2006). Each question assessed a single concept and the incorrect answers were often based on common incorrect conceptions.

Conceptest questions were originally designed to formatively assess student learning during the lecture and provide feedback for both the students and the instructor (Crouch & Mazur, 2001). Yet, little research has been conducted on the validity and reliability of conceptest questions. Validity refers to the degree to which an instrument measures what it is purported to measure (Ardovino, Hollingsworth, & Ybarra, 2000), and Gall et al. (2007) related validity to the use of theory and research results to interpret test scores. The conceptest questions from this study were developed and reviewed by expert earth scientists, therefore the conceptest questions used in this study contained accurate content information. As such, it is assumed that the questions used in this study met the criteria for content validity.

In contrast, other ways of measuring validity have yet to be assessed for any set of conceptest questions. Criterion validity assesses the degree to which the questions reflect a group of abilities (Salkind, 2006) and construct validity describes the ability of an
instrument to measure what it is purported to measure. Previously published studies on conceptest questions ignored the validity issue and apparently assumed that all conceptest questions were equally valid. One current study has begun to assess the validity of these earth science conceptest questions; however the results are not yet available (David Steer, personal communication).

Gall et al. (2007) defined reliability as the absence of measurement error from a set of test scores. Reliability could also be considered as the consistency of a test when administered to multiple people (Ardovino et al., 2000). One measure of reliability is item consistency which refers to an assessment of the degree to which these conceptest questions measure the appropriate content knowledge. Item consistency for those questions could be assumed if those questions were designed and reviewed by content experts and are only asked when discussing these concepts in class. This questioning process was followed in the nine classes included in this study, so item consistency is assumed to have been met for the purposes of this study. Like validity, little is known about the reliability of these conceptest questions; however research is currently underway to determine the reliability of the conceptest questions used in this study (David Steer, personal communication).

Several studies that have reported student response data aggregated the answers over the entire academic term (Freeman et al., 2007; Gray et al., in press; Stowell & Nelson, 2007). This study limited its focus to conceptest questions asked while presenting material on the solid earth (the geosphere) including plate tectonics, earthquakes, volcanoes, as well as rock and mineral formation. These topics were chosen because they form a coherent body of knowledge pertaining to the structure, composition, and
kinematics of the solid earth and all contributed to student comprehension of plate
tectonics. In addition, plate tectonics is a widely researched topic that students often
struggle to understand in part because of its abstract nature and operation at scales
beyond typical human experience (Gobert & Clement, 1999). In addition, these topics
were not discussed until the students had spent three weeks using the personal response
system. Consequently, any novelty from using the transmitters and answering the
conceptest questions should have subsided by the start of the study.

This study operationally defined two types of conceptest questions. Text-only
questions (Table 4) were conceptest questions that only contain text; whereas illustrated
conceptest questions (Table 5) contained a graph, map, photograph, diagram, or other
graphic that must be interpreted in order to determine the correct answer. Each question
was assigned to the appropriate category based on the presence or absence of an
illustrative graphic. If the question contained only text, it was designated as a text-only
question. Conversely, if a question contained any type of graphic, it was designated as an
illustrated conceptest question. These categories were mutually exclusive and covered the
entire set of questions, so finding a question that qualified for neither or both categories
did not occur.

During a typical lecture, students answered conceptest questions from two different
contexts. First the questions were individually answered by the students. If less than 75%
of the class failed to select the correct answer, the instructor had the option of asking the
students to discuss the question with their peers before voting a second time. This study
investigated individual student responses to different types of conceptest questions,
consequently the data set only included questions that were individually answered.
Table 4. Defining Characteristics and Examples of Text-Only Concepttest Questions

<table>
<thead>
<tr>
<th>Defining Characteristics</th>
<th>Examples</th>
</tr>
</thead>
</table>
| Questions contain only text and do not include a graph, map, diagram, or picture. Text-only questions also do not refer to any physical object or handout in the student's possession (such as a rock specimen). | **Definition/Knowledge**

**Atoms Concepttest**

Which are present in an atom's nucleus?
- A. Electrons
- B. Electrons and protons
- C. Neutrons
- D. Neutrons and protons

**Igneous Rocks Concepttest**

Imagine that all minerals found in rocks were the same color. What information would you no longer be able to infer?
- A. Texture
- B. Cooling rate
- C. Composition

<table>
<thead>
<tr>
<th>Scenario/Application</th>
<th>Multiple Criteria</th>
</tr>
</thead>
</table>

**Earthquake Concepttest**

An earthquake occurred on the Erie Fault 5 km beneath Ashtabula, Ohio. Damage from the earthquake was greatest in nearby Chardon. The fastest report of shaking was recorded in Akron. Where was the earthquake's epicenter?
- A. The Erie Fault
- B. Ashtabula
- C. Chardon
- D. Akron

**Viscosity Concepttest**

All other factors being equal, which magma would most easily release gases?
- A. High viscosity magma
- B. Low viscosity magma
- C. Neither, magma does not trap gases
Table 5. Defining Characteristics and Examples of Illustrated Concepttest Questions

<table>
<thead>
<tr>
<th>Defining Characteristics</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Graph</strong></td>
<td><strong>Map</strong></td>
</tr>
<tr>
<td>Questions contain a graph, map, diagram, or picture that required interpretation before the student can accurately answer the question. Also includes questions that contain only text but refer to a physical object or handout in the student's possession (such as a rock specimen).</td>
<td></td>
</tr>
<tr>
<td><strong>Igneous Rocks Concepttest</strong></td>
<td><strong>Plate Tectonics Concepttest</strong></td>
</tr>
<tr>
<td>The lines shown on the idealized graph below depict temperature time cooling histories for magma. What rock is best represented by cooling history A?</td>
<td></td>
</tr>
<tr>
<td>A. Granite  B. Basalt</td>
<td></td>
</tr>
<tr>
<td><strong>Physical Object</strong></td>
<td><strong>Diagram</strong></td>
</tr>
<tr>
<td><strong>Igneous Rocks Concepttest</strong></td>
<td><strong>Earth's Structure Concepttest</strong></td>
</tr>
</tbody>
</table>
| What type of igneous rock is sample #1?  
A. Gabro  B. Basalt  
C. Granite  D. Rhyolite |
| Which of the images below best approximates the relative distribution of Earth’s core, mantle, and crust?  
a)  
Big core, thin crust  
b)  
Big core, thick crust  
c)  
Small core, thick crust  
d)  
Small core, thin crust |
| If the San Andreas fault moves 500 cm per big earthquake, and fault movement is equivalent to plate motion (2.5 cm/yr): How many years of plate motions must accumulate to produce one big earthquake?  
A. 2 years  
B. 20 years  
C. 200 years  
D. 2000 years |
Answers collected after student discussions tend to produce a more homogenous data set with a larger percentage of students selecting the correct answer (Mazur, 1997b). Eliminating the answers collected after peer discussion removed the effect of peer discussion as a potentially confounding variable and more accurately assessed the responses given by individuals.

After each conceptest question was asked, the instructor’s computer recorded and tabulated the student responses to that question. A spreadsheet was generated for all of the questions asked during a particular class session including a copy of the question as well as the student responses. For this study, student responses were categorized based on the type of question asked (text-only versus illustrated). For each student, the percentage of correct responses was calculated for each type of question. These percentages were then analyzed as the dependent variable.

**ACT**

The American College Test (now known just as the ACT) is a standardized, multiple-choice test that is typically given to high school seniors (American College Test, 2008a). The test has four required subtests covering reading, English, math, and science with an optional writing test. Each test is scored on a 1-36 scale with a mean score of approximately 20. Quartile rankings for the three tests used in this study are given in Table 3. Besides the four subject tests, ACT reports a composite score using the average score from the four subject tests. An optional writing test was introduced in 2004. The makers of the ACT state that the tests are aligned with the material covered in the standard high school curriculum and are a general measure of a student’s ability to complete college-level work (ACT, 2008a).
The publishers of the ACT tests assessed the reliability and validity of the tests to ensure that they can be properly used to assess student achievement (ACT, 2007). Table 6 lists the calculated reliabilities and standard errors for the three tests used in this study. These values were calculated from a representative sample of 12,000 students taken across the six national administrations of the test during the 2005-2006 academic year. Reliability was calculated using the formula:

\[
\text{reliability} = 1 - \left( \frac{\text{standard error}}{\text{scale score variance}} \right)
\]

The report does not provide criteria for interpreting these values, but Gall et al. (2007) note that instruments with reliabilities greater than 0.80 are considered to be reliable and standardized tests (such as the ACT) typically have reliabilities near 0.90. The median reliabilities for all three tests meet Gall et al.’s criterion, and were considered to be reliable for the purposes of this study.

<table>
<thead>
<tr>
<th></th>
<th>Science</th>
<th>Math</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scale Score Reliability</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>0.80</td>
<td>0.91</td>
<td>0.85</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.74</td>
<td>0.89</td>
<td>0.85</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.83</td>
<td>0.92</td>
<td>0.87</td>
</tr>
<tr>
<td><strong>Standard Error</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>2.00</td>
<td>1.47</td>
<td>2.18</td>
</tr>
<tr>
<td>Minimum</td>
<td>1.90</td>
<td>1.43</td>
<td>2.11</td>
</tr>
<tr>
<td>Maximum</td>
<td>2.12</td>
<td>1.56</td>
<td>2.26</td>
</tr>
</tbody>
</table>

\(^a\) Based on data from ACT (2007).

ACT also assessed the validity of their tests by comparing test results to student grade point average (GPA) and level of coursework (ACT, 2007). Their results indicated that students who took more rigorous courses (such as calculus) earned higher scores on the corresponding ACT test. In addition, a multiple regression analysis was performed and
found that regression coefficients were significantly related to high school GPA and level of high school courses completed.

The ACT claimed that it measures student achievement rather than student aptitude (ACT, 2008a), and several recent studies have used the ACT as a measurement of student academic achievement (Ayers & Qualls, 2001; Cole & Osterlind, 2008; Koenig, Frey, & Detterman, 2008; Rodriguez, 1997). Though ACT clearly claimed their tests measure student achievement (ACT, 2008a), several recently published studies have used the test to measure a variety of constructs including: reading ability (Martino & Hoffman, 2002), student aptitude (Coyle, 2006; Fry, 2004; Whitmire, 2001), “math background” (Spencer, Steele, & Quinn, 1999), and “under-performing students” (Ting, 1997).

Even though the ACT was designed to measure a student’s achievement regardless of gender or race, differences in ACT scores have been observed. Rodriguez (1997) illustrated that women and students of color score lower on the ACT-Science test than males or Caucasian students. Similar differences have been observed for the ACT-Mathematics test (Halpern et al., 2007; Zwick, 2002). In general, past research has suggested that the ACT tests underestimate college grade point averages for women and minorities and overestimate grade point averages for men (Duckworth & Seligman, 2006).

In contrast, recent tests results have suggested that the observed gender gap may have been determined by the number of students who took the test rather than an innate difference between the genders. American College Test (2005) and Hyde, Lindberg, Linn, Ellis, and Williams (2008) used the same data to describe how gender differences in the ACT-Mathematics test have diminished to the point where the two scores are no
longer significant, and their effect sizes are close to zero. They also noted that data from some states indicated that women now score higher than men on this test. The observed gender gap was also not present in test scores from states (Colorado and Illinois) that required all students to take the ACT tests. Other research conducted by ACT found that differences in ACT scores due to differences in gender or race/ethnicity were significantly related to academic preparedness rather than biases in the test itself (Noble, Davenport, Schiel, & Pommerich, 1999; Noble, Roberts & Sawyer, 2006). Based on these reports, the ACT may not be biased towards women, and perhaps can be reliably used to measure student academic achievement.

This study compared student ACT scores (divided into quartile rankings) to differences in student responses to text-only and illustrated conceptest questions. These questions required students to be proficient in reading the questions (reading), interpreting graphs and diagrams (math), and assimilating scientific knowledge (science). Successful achievement in each of these areas may have influenced student success on one or both of the types of conceptest questions. Like Cole and Osterlind (2008), this study used results from the ACT as a measure of prior achievement. The tests used in this study included the ACT-Reading, ACT-Math, and ACT-Science tests. The ACT-Reading test was included instead of the ACT-English test because the ACT-Reading test required students to read a passage and answer comprehension and critical analysis questions based on the reading; whereas the ACT-English test required students to answer questions, such as verbal analogies, that assessed student knowledge of the English language. For the purposes of this study, the ACT tests were used to measure prior reading achievement and assessed student reading ability. The ACT-Reading test more
closely aligned with the goals of this study and therefore the ACT-English test is not used. Similarly, the ACT-Composite test was not included in this study. A goal of this study was to determine if preparation in a given content area (such as science) prepared students to answer a specific type of concepttest question. The ACT-Composite score was an overall score that measured general student achievement and academic preparedness rather than any achievement in a content-specific area such as math, and therefore was not applicable to this study.

The ACT scores were integer values ranging from zero to 36, but *ex post facto*, causal-comparative research required the use of a categorical independent variable (Gall et al., 2007). To accommodate this data requirement, the student ACT scores were converted into quartile values following the national quartile rankings (Table 3) provided by ACT (2008b). Quartile rankings divide the national scores into four categories so that each category contained an approximately equal number of scores. For this study, the top quartile was referred to as the high-achievement group, followed by the low-high achievement, high-low achievement, and low achievement groups.

**Data Analysis**

This study sought to identify any differences in student responses for the two different types of concepttest questions (text-only versus illustrated) asked in an introductory, earth science course. Specifically, did student responses vary by gender and/or prior achievement? Answering these questions required multiple variables and multiple analyses. All analyses were conducted using SPSS (version 15.0).
Variables

For the purposes of this study, the following variables were analyzed and used to answer each of the five research hypotheses (Table 7).

Gender – In this study, the term gender was used to denote the differences between males and females or men and women. There is no clear distinction between the use of gender or sex to refer to differences between men and women. Halpern et al. (2007) discussed the advantages of using either term and concluded that the choice is somewhat arbitrary and depends on the intentions of the author. This study focused on responses provided by men and women. Answers provided by individual students may be a function of both social/cultural environmental forces as well as biological forces. To connote this more general viewpoint, the term gender was chosen over sex; however this study did not measure or determine any of the possible social/cultural forces that may have shaped individual student gender identities. Gender was determined from student records maintained by the university’s Office of Institutional Research. For students where this field was left blank, a gender determination was not made and the students were removed from the study.

Prior Achievement – Prior achievement referred to the demonstrated success of an individual in a given academic domain. For the purposes of this study, prior achievement was divided into three categories (reading, math, and science achievement) and measured using the three corresponding scores from the ACT (ACT-Reading, ACT-Math, ACT-Science).
Table 7. Variables and Statistical Tests Associated With Each Research Hypothesis

<table>
<thead>
<tr>
<th>Research Hypotheses</th>
<th>IV</th>
<th>DV</th>
<th>Statistical Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>n/a</td>
<td>Question Type - The percentage of correct responses to text-only and illustrated CTQ&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Repeated Measures analysis</td>
</tr>
<tr>
<td>1b</td>
<td>Gender</td>
<td>The percentage of correct responses to text-only and illustrated CTQ&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Repeated Measures analysis</td>
</tr>
<tr>
<td>2a</td>
<td>Gender</td>
<td>Question Type - The percentage of correct responses to text-only and illustrated CTQ&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Repeated Measures analysis</td>
</tr>
<tr>
<td>2b</td>
<td>ACT-Reading, Math, &amp; Science</td>
<td>The percentage of correct responses to text-only and illustrated CTQ&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Repeated Measures analysis</td>
</tr>
<tr>
<td>2c</td>
<td>Gender, ACT-Reading, Math, &amp; Science</td>
<td>The percentage of correct responses to text-only and illustrated CTQ&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Repeated Measures analysis</td>
</tr>
</tbody>
</table>

<sup>a</sup> CTQ = Conceptest Questions
Scores were then converted into quartile rankings using the definitions listed in Table 3. These scores were obtained through the university’s Office of Institutional Research. Those students who did not report results from the ACT were not included in this study.

Question Type – The two categories of conceptest questions that were operationally defined in this study. *Text-only* conceptest questions were conceptually-based, multiple-choice questions that only contain text and were asked in a class as part of the peer instruction (Mazur, 1997a) pedagogy. In contrast, *illustrated* conceptest questions were conceptually-based, multiple choice questions that contained text plus a photograph, diagram, map, or graph that required interpretation in order to correctly answer the question. Student answers to these two types of questions were collected using e-Instruction’s CPS (eInstruction, 2009) personal response system. Data files for each class session were provided by each instructor. The percentage of correct responses for each type of question was calculated using the number of correct responses aggregated over the entire three- to five-week study period.

*Error Factors*

For this study, the type I and type II errors were set. Type I error is the probability that the analysis would indicate a significant difference when in fact there was no difference. Type I error is typically represented by the Greek letter alpha (α). Conversely, type II error was the probability that the analysis would indicate that a significant difference did not exist when in fact it does. Type II error is typically represented by the Greek letter beta (β). In addition, power is defined as the probability that the analysis indicates a significant difference when in fact a difference did exist (Stevens, 2007). Power was calculated as Power = 1 − β, where β is the type II error level. Cohen (1992)
recommended setting $\beta = 0.20$ and power $= 0.80$. For this study, power was set at 0.80, meaning that an 80% probability that the analyses are correct was considered acceptable. For Type I error, the $\alpha$ level must be selected by the researcher ahead of time but is also a function of sample size. Larger values of $n$ allow for smaller values of $\alpha$. Cohen (1988) published several sets of tables relating $\alpha$ and sample size to power and effect size. The exploratory nature of this study left the possibility that observed differences in student responses to conceptest questions might yield small effect sizes ($d = 0.2$), so it was determined to set $\alpha$ at a level to detect the smallest possible effect sizes even at the risk of increasing Type I error. Cohen’s power tables for t-tests indicated that an $\alpha = 0.01$ would need a sample size of more than 600 to provide sufficient power and detect a small effect size. The same tables indicated that raising the $\alpha$-level to $\alpha = 0.05$ decreased the smallest detectable effect size to $d = 0.20$. This level of effect size is defined by Cohen (1988, 1992) as the lower limit of a small effect size and values of $d$ smaller than 0.20 represent no appreciable difference in the two means and would be difficult to detect with the naked eye. Given the possibility that any significant differences in student responses could yield small effect sizes, it was determined that the Type I error would be set at $\alpha = 0.05$.

Stevens (2002) provided data suggesting that a higher $\alpha$-level could be used. He noted that Cohen (1988) did not provide a power table for repeated measures analyses. By repeatedly measuring the same variable, a smaller sample size can achieve an acceptable level of power. Stevens’ calculations required setting the correlation between the measurements and knowing the number of measurements used in the analysis. This study used the percentage of correct responses for (1) text-only and (2) illustrated conceptest
questions as the repeated measures. Assuming a low correlation between these two variables (the most conservative estimate) yields samples sizes of $n = 268$ for $\alpha = 0.05$ and $n = 404$ for $\alpha = 0.01$. Even though these results suggest that the Type I error level for this study could be reduced to $\alpha = 0.01$ and still maintain an acceptable level of power, it was decided that the Type I error would be set at $\alpha = 0.05$. Doing so assured that the sample size used in this study would be sufficiently large to detect effect sizes even smaller than $d = 0.20$ at the risk of committing a Type I error.

Data Manipulation

The raw student data were retrieved from the databases generated by the electronic personal response system (eInstruction’s CPS software, eInstruction, 2009) and placed into an Excel spreadsheet. The system created a separate spreadsheet for each class session with student responses paired with their transmitter identification numbers and student names. These data were cross-referenced with a list of students who signed consent forms and a unique project identification number was assigned to each student and transmitter number. Student demographic data were added by matching student names to information provided by the university. Once the project numbers were assigned, all student names were removed from the database and the student response data were added from each class session.

As the data were retrieved from the personal response system software, a list of conceptest questions was assembled and questions were categorized as being text-only or illustrated questions. (See Appendices D and E for a list of these questions.) Besides the student responses to each question, the total number of text-only and illustrated questions asked during the study was determined. For each participant the number of questions
answered, the number of correct responses, and the percentage of correct responses for each types of question were then calculated. Percentage values were entered into SPSS (version 15.0) as values ranging between zero and one hundred rather than as decimals less than one.

**Data Reporting**

Unlike the physical sciences, educational researchers are given latitude in selecting the precision with which they report results from inferential statistics. The fifth edition of the publication manual from the American Psychological Association (APA, 2001) states that authors can choose whether to report significance levels as equalities (e.g. \( p = 0.02 \)) or inequalities (e.g. \( p < 0.01 \)). For this study, significance levels and results from multivariate statistical analyses (e.g. \( F \) and \( t \) values or Box’s \( M \)) are reported as equalities to the second decimal places unless the significance was smaller than \( p = 0.01 \). In these cases, \( p \)-values are reported to three decimal places (e.g. \( p = 0.001 \)) or listed as inequalities if the significance is smaller than 0.001. Effect sizes are reported to two decimal places. For other values reported in this study, the following conventions were used. Exact values such as the number of women students in the study were reported as whole numbers. Any arithmetically calculated values such as means were reported to one decimal place. These conventions were used to report findings from this study only. Data from other studies are reported as written in the original sources.

**Data Pre-Screening and Descriptive Statistics**

Before conducting the inferential statistics associated with the research hypotheses, the data were analyzed for outliers, descriptive statistics were calculated, and internal consistencies were determined. Tests for outliers included constructing histograms and
calculating the z-scores for each participant’s response scores. Mertler and Vannatta (2005) recommended removing participants with z-scores greater than ± 4.0. The robustness of the study was further determined using the Kolmogrov-Smirnov test for normality. Mertler and Vannatta (2005) recommend that significance tests be conducted at the $\alpha = 0.01$ or 0.001 level to determine if the data are unduly skewed. Descriptive statistics included means and standard deviations for all of the study variables. Internal consistencies were calculated using Cronbach’s alpha (Stevens, 2007).

Repeatead Measures

Research questions one and two were examples of a “within-subjects” design (Stevens, 2007) that were best answered using a repeated measures analysis of variance. In a “within-subjects” design all students provide data for each of the dependent variables. For this study all of the students provided responses to text-only and illustrated questions and a control group that did not answer any of the questions was not present.

Advantages and disadvantages.

In a repeated measures test, the same participants experience more than one treatment or are evaluated more than one time. Repeatedly measuring the dependent variable(s) for the same participants produced several advantages when compared to a typical between-subjects ANOVA. First, the number of participants needed to yield significant results is reduced (Davis, 2002; Stevens, 2007). For each individual participant, the repeated measures analysis evaluated the effect of the independent variable(s) on the dependent variable(s). Second, this process eliminated variability between the participants due to individual differences (Stevens, 2007) which improved the power of the test. Stevens (2007) argued that a repeated measures analysis was more powerful than a randomized
experimental design. Recall that power is the percent probability that the researcher uses to make a correct decision when evaluating the results from an inferential statistical analysis (Stevens, 2007). In addition, a repeated measures design allowed participants to serve as their own control because the same subjects are evaluated under the control and treatment conditions (Davis, 2002).

Research Hypothesis 1a was answered using a repeated measures design that did not include an independent variable. For this test, the two dependent variables (the two types of questions) served as the two repeated measures and any significant differences would indicate students in general did not provide the same percentage of correct responses to both types of questions.

A repeated measures analysis that includes one or more independent variables is also called a “mixed-ANOVA” design. This type of analysis was used to answer Research Hypotheses 1b through 2c with gender (Question #1) and data from each of the three ACT tests (Question #2) serving as the independent variable.

Repeated measures analyses have two primary disadvantages (Davis, 2002). First, as the name implies, measurements of the same variable must be repeatedly collected or all participants have to be assessed across multiple variables. This study addressed this issue by only selecting students who answered both text-only and illustrated concept test questions. Those that answered one type but not the other were not included in the final analysis. Second, the researcher often cannot control the circumstances under which the data are collected. Participants may not provide data for all measurements. This disadvantage is not an issue for this study because the dependent variables contain data
Aggregated across multiple class sessions, thus ensuring that all of the participants were assessed multiple times.

Assumptions for repeated measures analysis.

Repeated measures, mixed-ANOVA analyses must meet three assumptions in order to yield valid results. Those assumptions are:

1. Independent observations
2. Multivariate normality
3. Homogeneity of covariances (sphericity)

The final assumption is often referred to as sphericity, and for mixed ANOVA analyses like the ones used in this study, sphericity must hold true for all levels of the independent and repeated dependent variables (Leech et al., 2005). In this study it was assumed that students accessed their content knowledge when answering each question. By limiting the study to the first time students answered a given question, any influence between questions was minimized and assured that the independence assumption was met.

Multivariate normality is difficult to quantify, so most techniques rely on a visual inspection of various graphs. Mertler and Vannatta (2005) recommended investigating the bivariate scatter plot matrix for all variables. Normally distributed combinations of variables produce elliptical patterns. Rencher (2002) described how Q-Q plots can also be used to assess multivariate normality. These graphs compare the sample data for one variable against a normal distribution. Data that are normally distributed produce a diagonal line. Stevens (2007) warned that such graphical tests are highly subjective and cautioned on relying on them for a final normality determination. Stevens (2007) also
stated that repeated measures ANOVAs are robust towards violations of normality; however he recommended that the data should approximate a normal distribution for each category within the independent variable. To verify the results from these visual tests, the Kolmogrov-Smirnov test for normality was used to assess normality for the response data for both types of questions within each category of independent variable. For example, Kolmogrov-Smirnov normality values for men and women were determined for responses to both text-only and illustrated questions.

To account for unequal variances, Leech et al. (2005) recommended using Mauchly’s or Box’s tests for sphericity. Stevens (2007) noted that many research designs in the social and behavioral sciences violate the sphericity assumption and argued that Mauchly’s test is sensitive to departures from normality. He recommended using the Greenhouse-Geisser or Huynh-Feldt tests. This study used all three tests to determine whether the data satisfied the sphericity assumption and violations were accounted for using the Greenhouse-Geisser test and the Huynh-Feldt epsilon corrections (Leech et al., 2005; Mertler & Vannatta, 2005; Stevens, 2007).

This project did not have a true control class, yet by using a repeated measures design, the student responses served as their own controls when comparing responses to the text-only and the illustrated concept test questions. Furthermore, this analysis, combined with the large sample size (n > 500), greatly increased the power of the study which in turn improved the confidence in the results.

*Types of results.*

The repeated measures, mixed-ANOVA design provided several tools for analyzing the data and answering each of the research hypotheses including within-subjects,
between-subjects, and post hoc analyses. A within-subjects analysis tested whether the interaction between the student responses the associated independent variable was significant. This analysis tested whether gender or prior achievement resulted in differences in responses to text-only and illustrated questions and whether the trend across the multiple measures was significantly linear, quadratic or cubic in shape. For example, to answer Research Hypothesis 1b, the within-subjects analysis tested whether men and women correctly answered a significantly different percentage of text-only questions than illustrated questions. This analysis did not compare results between the genders. Besides testing the interaction between the variables, the within-subjects analysis also tested the main effect of whether there were any differences in student responses to the two types of questions without considering the independent variable. Results from this test answered Research Hypothesis 1a.

Besides testing for differences within each subject, SPSS also calculated any significant differences between subjects. The between-subjects analysis determined whether significant differences existed between the categories of the independent variable. For example, the between subjects analysis tested whether men and women correctly answered a different percentage of questions.

The post hoc analyses determined the magnitude of any observed differences in student responses and provided details on which situations resulted in significant differences. For this study, three types of post hoc analyses were conducted. The first post hoc test, pairwise comparisons, tested whether significant differences were present for all possible combinations of the independent variable. For example, when comparing ACT scores to student responses, the pairwise comparisons tested whether each possible

123
pair of categories (high and, high-average, high and low-average, high and low, etc) yielded significantly different responses. These results characterized the nature of any significant relationships found in the within-subjects analysis. The second type of post hoc test, a univariate ANCOVA was used to determine the magnitude of any difference in student responses between the genders. This analysis provided data on the magnitude of any differences in the percentage of correct responses given by men and women. The third type of post hoc test consisted of paired t-tests to test whether students within each ACT achievement category correctly answered a significantly different percentage of text-only and illustrated questions.

For this study, the effect size of any significant difference was calculated to determine the magnitude of the observed difference. Two effect sizes were calculated based on the nature of the statistical test that was used. For the post hoc pairwise combinations, paired t-tests, and univariate ANCOVA, effect size was calculated using Cohen’s d value (Cohen, 1992). This metric is the ratio of the mean difference divided by the average standard deviation for the two populations. Values for d (Table 8) are expressed in units of standard deviations, so an effect size of d = 0.5 means that the two mean values are half of a standard deviation apart which represents approximately 19%. Viewed another way, an effect size of d = 0.50 means that only 33% of the two populations overlap. Larger effect sizes denote smaller regions of overlap and greater differences between the two populations. There is not theoretical upper limit to d, but it is rare to find d values larger than one (Leech, 2007). Cohen set these levels to correspond with the ease with which the difference could be observed. Small effect sizes are present but most likely are
difficult to observe, whereas medium effect sizes should be readily observable. Large
effect sizes represent gross differences between the two variables in question.

Table 8. Corresponding Values of $d$ and Partial Eta Squared for Different Levels of Effect Sizes

<table>
<thead>
<tr>
<th>Effect Magnitude</th>
<th>$d$</th>
<th>Partial Eta Squared</th>
<th>% Overlap$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>0.8</td>
<td>0.51</td>
<td>14.7</td>
</tr>
<tr>
<td>Medium</td>
<td>0.5</td>
<td>0.36</td>
<td>33.0</td>
</tr>
<tr>
<td>Small</td>
<td>0.2</td>
<td>0.14</td>
<td>47.4</td>
</tr>
</tbody>
</table>

$^a$ Between sample populations

The second effect size metric reported in this study is the partial eta squared ($\eta^2_p$) value which is the ratio of the sum of the squares ($SS_{between}$) associated with the independent variable with the sum of those squares plus the sum of the squares ($SS_{error}$) for the error term (Levine & Hullet, 2002). Numerically, the formula can be expressed as $\eta^2_p = \frac{SS_{between}}{SS_{between} + SS_{error}}$. This measure is automatically calculated by SPSS for univariate and multivariate analyses of variance and Levine and Hullet noted that partial eta squared was incorrectly reported in earlier versions of SPSS. For this study, partial eta squared is reported for the univariate ANCOVA post hoc analysis used to answer Research Hypothesis #1b and was included in places where a value for $d$ could not be determined.

Summary

This study used an *ex post facto*, causal-comparative design to investigate difference in correct student responses to two types of concept test questions. The independent
variables included gender and prior achievement (as measured by the Act-Reading, ACT-Math, and ACT – Science tests). The dependent variable was question type (text-only versus illustrated) and was measured by the percentage of correct responses answered by each student. To ensure that a sufficient level of power was incorporated into this study, the results were interpreted using a type I error of $\alpha = 0.05$ and a power of 0.80.

The data were analyzed using a repeated measures design for which the two independent variables were assessed. Possible covariate factors were considered and included when a significant interaction was observed. The repeated measures analysis assessed both the within-subjects and between subjects differences and three types of post hoc analyses (pairwise combinations, paired t-tests, univariate ANCOVA) further described the relationships between these variables. Effect sizes (Cohen’s (1992, 1988) $d$-value and the partial eta squared term) were calculated to determine the magnitude of any significant difference.
CHAPTER IV
RESEARCH RESULTS

Introduction

This chapter presents the statistical results pertaining to the study. The descriptive
statistics and steps undertaken to screen the data and satisfy the assumptions associated
with the repeated measures inferential statistical method are presented followed by the
results for each of the research hypotheses.

Data Description and Screening

This section describes the criteria used to select both the student participants and
concept test questions used in this study and documents how the resulting database was
analyzed to ensure all statistical assumptions were met.

Selection of Student Participants

Official enrollment records provided by the university indicated that 1236 students
had enrolled in the nine classes included in this study. Thirty-one of these students
repeated the course and were listed twice on the enrollment sheets. In all cases, the
participants from this study were only included the first semester they took the course
which left 1205 students whose responses were available for further analysis (Table 9).
Students who added the class after the generation of the enrollment lists were not
included in this total and were excluded from this study. In addition, the lists did not
identify those students who initially enrolled in the course but dropped the class prior to
the start of this study. Students who answered zero conceptest questions during the study were assumed to have dropped the course and were not included.

Table 9. Student Demographics Used to Determine Study Sample Size

<table>
<thead>
<tr>
<th>Data Categories</th>
<th>All Students</th>
<th>Remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
</tr>
<tr>
<td>Total Number of Enrolled Students</td>
<td>1236</td>
<td>100</td>
</tr>
<tr>
<td>Repeat Students</td>
<td>31</td>
<td>2.5</td>
</tr>
<tr>
<td>Post-Secondary Students</td>
<td>39</td>
<td>3.2</td>
</tr>
<tr>
<td>Students with Signed IRBs</td>
<td>822</td>
<td>66.5</td>
</tr>
<tr>
<td>Students with ACT Scores</td>
<td>984</td>
<td>79.6</td>
</tr>
<tr>
<td>Students Answered Five or More Text-Only Questions</td>
<td>1004</td>
<td>81.2</td>
</tr>
<tr>
<td>Students Answered Five or More Illustrated Questions</td>
<td>1124</td>
<td>90.9</td>
</tr>
</tbody>
</table>

Number of Students Included in This Study = 561

The number of students included in this study was further reduced by the availability of student information regarding their demographic and standardized test results (Table 9). Thirty-nine students (3.2%) were high school students taking the course as part of an accelerated post-secondary program and were removed from the study because they were under 18 years of age and could not legally sign the consent form to participate in the study. More than half of the students (822 students – 66.5%) signed permission forms as
required by the university’s Institutional Review Board to participate in another study that included the collection and analysis of student personal response data. (Appendix A contains a copy of this letter.) Results from the ACT test were available for 984 (79.6%) of the students.

The number of participants available for analysis was further reduced to account for two potential sources of error. A review of the data identified seven students who only answered illustrated questions and one student who only answered text-only questions. The analyses in this study compared student responses to text-only and illustrated concepttest questions and required at least one response to each type of question, so all eight students were excluded from the study. In addition, 73 students answered fewer than five text-only questions and an additional two students answered fewer than five illustrated questions. For example, one student answered eleven illustrated questions but only one text-only question. By answering only one text-only question, the student could only score a zero or 100% percent correct for this type of question. Answering such a small number of questions would only generate an extremely large or small percentage of correct responses that would add to the furthest extremes of the sample distribution. Such results could only adversely bias the study’s analysis by generating artificially high or low percentages of correct responses, so all 75 students were removed from consideration, which left a total of 561 students available for the study.

Participant Demographics

The student participants included in this study were selected from nine classes of an undergraduate earth science course at a large, urban university on the Midwestern United States (Table 10). Demographic data were then obtained from the university’s admission
records. Slightly more men than women participated in the study (303 men compared to 258 women). Only 65 students (11.6%) identified themselves as a racial minority (African American, Native American, Pacific Islander, or Hispanic), and 18 students (3.2%) did not provide any information concerning their racial identity. A majority of the students (63.1%) had not yet declared a major and were listed as part of the University College, yet 65.6% of the students anticipated pursuing a Bachelors of Science degree.

(Additional student demographic information is presented in Appendix B.)

Table 10. Numbers of Students From Each Class Included in This Study

<table>
<thead>
<tr>
<th>Class #</th>
<th>Semester</th>
<th>Instructor</th>
<th>Class Time</th>
<th>Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fall 06</td>
<td>Baker</td>
<td>8:50</td>
<td>51</td>
</tr>
<tr>
<td>2</td>
<td>Fall 06</td>
<td>Baker</td>
<td>11:00</td>
<td>76</td>
</tr>
<tr>
<td>3</td>
<td>Spring 07</td>
<td>Baker</td>
<td>12:05</td>
<td>84</td>
</tr>
<tr>
<td>4</td>
<td>Fall 07</td>
<td>Baker</td>
<td>8:50</td>
<td>55</td>
</tr>
<tr>
<td>5</td>
<td>Fall 07</td>
<td>Baker</td>
<td>11:00</td>
<td>79</td>
</tr>
<tr>
<td>6</td>
<td>Fall 07</td>
<td>Anderson</td>
<td>1:10</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>Spring 08</td>
<td>Baker</td>
<td>8:50</td>
<td>49</td>
</tr>
<tr>
<td>8</td>
<td>Spring 08</td>
<td>Anderson</td>
<td>11:00</td>
<td>60</td>
</tr>
<tr>
<td>9</td>
<td>Spring 08</td>
<td>Baker</td>
<td>12:05</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>Totals</td>
<td></td>
<td></td>
<td>561</td>
</tr>
</tbody>
</table>

The nine classes were taught by two instructors (Dr. Anderson and Dr. Baker) at four different class times (Table 11). The number of class sessions devoted to the topics included in this study varied from 15 to nine sessions. During the Spring of 08, Dr. Baker
was gone from the class for one week due to travel and illness. To keep his course on
pace with Dr. Anderson’s class, Dr. Baker did not ask any conceptest questions
associated with volcanoes and mountain building. The names of the instructors are
pseudonyms.

Selection and Classification of Conceptest Questions

Following the peer instruction model of Mazur (1997a), the students in this study
used their transmitters to individually answer conceptest questions without consulting
with their peers. For approximately 10% of the questions (Table 12), the students
discussed the question with their peers before answering the question a second time. In
the 105 class sessions selected for this study (Table 11), students were asked to answer
534 questions including 413 individual questions and 92 group questions (Table 12). To
ensure that the data reflected individual student interactions with the conceptest questions
only the 413 individually-answered conceptest questions were included in this study.
Answers to each question were electronically recorded using e-instruction’s CPS personal
response system and downloaded as an Excel spreadsheet.

The individually-answered conceptest questions were assigned to one of two
categories based on how the question was worded (Tables 4 and 5) and all of the
questions used in this study are listed in Appendices C and D. Text-only questions
contained text and did not include any graphs, maps, diagrams, pictures, or physical
specimens that students could refer to while answering the question. These questions
typically required the students to apply their content knowledge to arrive at the answer.
Table 11. Number Of Class Sessions Devoted To Each Content Topic

<table>
<thead>
<tr>
<th>Topic</th>
<th>Fall 2006</th>
<th>Fall 2007</th>
<th>Fall 2007</th>
<th>Spring 2007</th>
<th>Spring 2007</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class #</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Continental Drift &amp; Plate Tectonics</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Earthquakes</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Volcanoes &amp; Mountain Building</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Rocks &amp; Minerals</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>15</strong></td>
<td>15</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

Note. The classes with zero days devoted to volcanoes and mountain building reflect days off due to travel and illness for Dr. Baker. Topics were discussed on other days but conceptest questions for this topic were not asked.
<table>
<thead>
<tr>
<th>Types of Questions</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>60</td>
<td>50</td>
<td>45</td>
<td>45</td>
<td>41</td>
<td>54</td>
<td>38</td>
<td>41</td>
<td>39</td>
<td>413</td>
</tr>
<tr>
<td>Group</td>
<td>18</td>
<td>15</td>
<td>12</td>
<td>4</td>
<td>3</td>
<td>24</td>
<td>5</td>
<td>10</td>
<td>1</td>
<td>92</td>
</tr>
<tr>
<td>Repeat</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Survey/Non-Conceptest</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>Unknown Answer/Other</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td><strong>Total # of Questions</strong></td>
<td>84</td>
<td>74</td>
<td>61</td>
<td>51</td>
<td>47</td>
<td>78</td>
<td>45</td>
<td>53</td>
<td>41</td>
<td>534</td>
</tr>
</tbody>
</table>
Some questions required the understanding of related vocabulary, and other text-only questions required students to evaluate one or more criteria in order to arrive at an answer. For example, question T-33 (Figure 1) assessed student understanding of the term epicenter and question T-13 (Figure 1) required students to evaluate the likelihood that each answer would trigger a volcanic eruption.

<table>
<thead>
<tr>
<th>Conceptest Question T-33</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earthquake intensity measures _____</td>
</tr>
<tr>
<td>A. Energy released from earthquake</td>
</tr>
<tr>
<td>B. Amplitude of seismic waves on seismogram</td>
</tr>
<tr>
<td>C. Damage resulting from an earthquake</td>
</tr>
<tr>
<td>D. Displacement on faults</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conceptest Question T-13</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the most important factor in controlling the viscosity of magma?</td>
</tr>
<tr>
<td>A. Energy released from earthquake</td>
</tr>
<tr>
<td>B. Amplitude of seismic waves on seismogram</td>
</tr>
<tr>
<td>C. Damage resulting from an earthquake</td>
</tr>
<tr>
<td>D. Displacement on faults</td>
</tr>
</tbody>
</table>

Figure 1. Examples of text-only conceptest questions that required students to access knowledge of vocabulary (Question T-33) or evaluate criteria (Question T-13).

Illustrated questions included a graph, map, diagram, or picture as well as text. In each case, the question required interpretation of the image before the student could accurately answer the question (Table 5). For example, question I-9 (Figure 2) contained four maps of the world with different sets of bold lines running across the map. The
question asked students to identify the map that represented divergent plate boundaries.

In order to correctly answer this question, students had to accurately read the maps before they could answer the questions. Similarly, question I-18 (Figure 2) showed an idealized seismogram and asked students to identify the point at which the S-waves had reached the location of the seismograph. For this question, students had to correctly interpret the graph and relate that information to the concepts pertaining to the different types of earthquake waves.

![Figure 2. Examples of illustrated concepttest questions that required the interpretation of a map and diagram or referred to a physical specimen that the student could manipulate.](image-url)
Nineteen questions were classified as illustrated questions even though they did not contain an image embedded with the question. In each case, the question referred to a physical specimen or graph that the students had at their desks. For example, after the students examined two vials containing liquids of different viscosities, question I-23 asked which vial contained a fluid with the higher viscosity. Similarly, when studying igneous rocks, the students had four rocks at their desk and question I-31 asked students to identify the rock type of sample number 1. In every case, the students had a graph or physical object at their desk at the time the concept test question was asked. In addition, the question required the students to interpret their observations of the specimen before they could arrive at the correct answer. This cognitive process was more closely related to the illustrated questions than the text-only questions, so they were classified as illustrated concept test questions.

Data Calculations

Individual student responses to the text-only and illustrated concept test questions were collected for each class session within the scope of this study. From these data, the number of questions asked, the number of questions answered, and the number of correct responses were determined. For both the text-only and illustrated questions, the percentage of correct responses were calculated by taking the number of correct responses divided by the number of questions answered multiplied by one hundred. All of the statistical analyses reported in this study used these percentages as the dependent variables.
Identifying Outliers

The demographic and student response data collected for this study were placed in a spreadsheet (Microsoft Excel 2003) and uploaded into SPSS (v. 15.0) for further analysis. Descriptive statistics were calculated to identify outliers and to ensure that the data conformed with the assumptions for the repeated measures analysis. The data for all of the study variables indicated that significant outliers were not present in this data set. Calculated z-score values (Tables 13 – 17) were all within the range specified by Mertler and Vannatta (2005) (-4.0 < z < +4.0), suggesting that outliers were not present in these samples. As such, the student response data did not require any transformations.

Testing Assumptions of Repeated Measures Design

The repeated measures analysis assumed that the data are independently observed, normally distributed, and spherical. Each assumption was evaluated before the research hypotheses were tested. Sample independence is related to the research design, and for this study it was assumed that individual student responses were given independent of the other responses and independent from responses provided by other students.

Multivariate Normality.

Multivariate normality was assessed using visual and numerical methods and indicated that the data satisfied this assumption. The bivariate scatterplot matrix (Appendix E) produced an elliptical pattern for the interaction between the two dependent variables (responses to text-only and illustrated questions) which suggested that the data satisfied the multivariate normality assumption. The plots (Appendix E) were not elliptical for the other combinations because the independent variables contained a small number of categories.
Table 13. Normality Data For Text-Only and Illustrated Conceptest Questions Analyzed in This Study

<table>
<thead>
<tr>
<th></th>
<th>mean(^a)</th>
<th>st. dev</th>
<th>min</th>
<th>max</th>
<th>Kolmogorov-Smirnov Z</th>
<th>sig. (^b)</th>
<th>min z-value</th>
<th>max z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text-Only Questions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Asked</td>
<td>14.3</td>
<td>3.6</td>
<td>9.0</td>
<td>21.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Answered</td>
<td>11.8</td>
<td>4.1</td>
<td>5.0</td>
<td>21.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Correct</td>
<td>7.6</td>
<td>3.4</td>
<td>0.0</td>
<td>18.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Correct(^c)</td>
<td>63.8</td>
<td>17.3</td>
<td>0.0</td>
<td>100.0</td>
<td>1.23</td>
<td>0.10</td>
<td>-3.6</td>
<td>2.1</td>
</tr>
<tr>
<td><strong>Illustrated Questions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Asked</td>
<td>31.0</td>
<td>5.3</td>
<td>23.0</td>
<td>42.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Answered</td>
<td>24.7</td>
<td>7.1</td>
<td>6.0</td>
<td>42.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Correct</td>
<td>16.2</td>
<td>5.8</td>
<td>2.0</td>
<td>32.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Correct(^c)</td>
<td>65.3</td>
<td>13.6</td>
<td>18.2</td>
<td>100.0</td>
<td>0.95</td>
<td>0.33</td>
<td>-3.3</td>
<td>2.3</td>
</tr>
</tbody>
</table>

Note. \(N = 561\) students. Cronbach's Alpha = 0.544 for percentage of correct text-only and illustrated conceptest questions. \(^a\) Results are mean values per student. \(^b\) Significance set at \(p \leq 0.05\) (two-tailed) \(^c\) % Correct = [((# Correct) / (#Answered)) * 100]
Table 14. Normality Data Categorized by Gender For the Percentage of Correct Responses to Text-Only and Illustrated Conceptest Questions

<table>
<thead>
<tr>
<th></th>
<th>number of students</th>
<th>mean&lt;sup&gt;a&lt;/sup&gt;</th>
<th>st. dev.</th>
<th>min</th>
<th>max</th>
<th>Kolmogorov-Smirnov Z</th>
<th>sig. &lt;sup&gt;b&lt;/sup&gt;</th>
<th>min z-value</th>
<th>max z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text-Only Questions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>303</td>
<td>65.3</td>
<td>17.4</td>
<td>20</td>
<td>100</td>
<td>1.40</td>
<td>0.04*</td>
<td>-2.5</td>
<td>2.1</td>
</tr>
<tr>
<td>Women</td>
<td>258</td>
<td>62.0</td>
<td>17.1</td>
<td>0</td>
<td>100</td>
<td>0.90</td>
<td>0.39</td>
<td>-3.6</td>
<td>2.1</td>
</tr>
<tr>
<td><strong>Illustrated Questions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>303</td>
<td>67.9</td>
<td>13.1</td>
<td>24</td>
<td>100</td>
<td>1.22</td>
<td>0.10</td>
<td>-3.0</td>
<td>2.3</td>
</tr>
<tr>
<td>Women</td>
<td>258</td>
<td>62.3</td>
<td>13.6</td>
<td>18</td>
<td>94</td>
<td>0.69</td>
<td>0.73</td>
<td>-3.3</td>
<td>1.9</td>
</tr>
</tbody>
</table>

<sup>a</sup> Results are mean values per student.  
<sup>b</sup> Significance set at p ≤ 0.05 (two-tailed)  
* Significant at p<0.05
Table 15. Normality Data Categorized by Student ACT-Reading Categories For the Percentage of Correct Responses to Text-Only and Illustrated Concepttest Questions

<table>
<thead>
<tr>
<th>Number of</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Kolmogorov-Smirnov Z</th>
<th>Significance</th>
<th>Minimum Z-Value</th>
<th>Maximum Z-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text-Only Questions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>123</td>
<td>71.0</td>
<td>15.6</td>
<td>28.6</td>
<td>100.0</td>
<td>0.83</td>
<td>0.50</td>
<td>-2.0</td>
</tr>
<tr>
<td>High-Average</td>
<td>174</td>
<td>65.0</td>
<td>16.8</td>
<td>0.0</td>
<td>100.0</td>
<td>0.69</td>
<td>0.72</td>
<td>-3.6</td>
</tr>
<tr>
<td>Low-Average</td>
<td>147</td>
<td>61.7</td>
<td>16.3</td>
<td>14.3</td>
<td>92.9</td>
<td>0.77</td>
<td>0.59</td>
<td>-2.8</td>
</tr>
<tr>
<td>Low</td>
<td>117</td>
<td>56.8</td>
<td>18.1</td>
<td>0.0</td>
<td>100.0</td>
<td>0.90</td>
<td>0.40</td>
<td>-3.6</td>
</tr>
</tbody>
</table>

| **Illustrated Questions** |       |                    |         |         |                      |             |                |                |
| High      | 123  | 71.3               | 11.4    | 39.1    | 95.0                 | 0.66        | 0.78           | -1.9           | 1.9            |
| High-Average | 174  | 66.8               | 12.8    | 25.0    | 100.0                | 0.79        | 0.57           | -2.9           | 2.3            |
| Low-Average | 147  | 63.1               | 13.4    | 20.0    | 90.9                 | 0.78        | 0.58           | -3.3           | 1.9            |
| Low       | 117  | 59.8               | 14.4    | 18.2    | 91.3                 | 0.77        | 0.60           | -3.0           | 1.9            |

\(^a\) Results are mean values per student. \(^b\) Significance set at \(p \leq 0.05\) (two-tailed)
Table 16. Normality Data Categorized by Student ACT-Math Categories For the Percentage of Correct Responses to Text-Only and Illustrated Concepttest Questions

<table>
<thead>
<tr>
<th></th>
<th>number of students</th>
<th>mean(^a)</th>
<th>st. dev.</th>
<th>min</th>
<th>max</th>
<th>Kolmogorov-Smirnov Z</th>
<th>sig. (^b)</th>
<th>min z-value</th>
<th>max z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text-Only Questions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>91</td>
<td>72.0</td>
<td>15.3</td>
<td>33.3</td>
<td>100.0</td>
<td>0.75</td>
<td>0.63</td>
<td>-1.7</td>
<td>2.1</td>
</tr>
<tr>
<td>High-Average</td>
<td>177</td>
<td>64.4</td>
<td>16.8</td>
<td>0.0</td>
<td>100.0</td>
<td>0.78</td>
<td>0.57</td>
<td>-3.6</td>
<td>2.1</td>
</tr>
<tr>
<td>Low-Average</td>
<td>169</td>
<td>64.3</td>
<td>16.7</td>
<td>20.0</td>
<td>100.0</td>
<td>0.90</td>
<td>0.40</td>
<td>-2.2</td>
<td>2.1</td>
</tr>
<tr>
<td>Low</td>
<td>124</td>
<td>56.1</td>
<td>17.4</td>
<td>0.0</td>
<td>92.9</td>
<td>0.62</td>
<td>0.83</td>
<td>-3.6</td>
<td>1.7</td>
</tr>
<tr>
<td><strong>Illustrated Questions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>91</td>
<td>75.1</td>
<td>11.2</td>
<td>33.3</td>
<td>96.3</td>
<td>0.72</td>
<td>0.68</td>
<td>-1.4</td>
<td>2.3</td>
</tr>
<tr>
<td>High-Average</td>
<td>177</td>
<td>65.9</td>
<td>12.1</td>
<td>23.8</td>
<td>92.0</td>
<td>0.94</td>
<td>0.34</td>
<td>-3.0</td>
<td>1.9</td>
</tr>
<tr>
<td>Low-Average</td>
<td>169</td>
<td>64.0</td>
<td>12.9</td>
<td>18.2</td>
<td>100.0</td>
<td>0.64</td>
<td>0.81</td>
<td>-2.7</td>
<td>1.9</td>
</tr>
<tr>
<td>Low</td>
<td>124</td>
<td>59.1</td>
<td>14.2</td>
<td>20.0</td>
<td>91.3</td>
<td>0.61</td>
<td>0.85</td>
<td>-3.3</td>
<td>1.9</td>
</tr>
</tbody>
</table>

\(^a\) Results are mean values per student. \(^b\) Significance set at p ≤ 0.05 (two-tailed)
Table 17. Normality Data Categorized by Student ACT-Science Categories For the Percentage of Correct Responses to Text-Only and Illustrated Conceptest Questions

<table>
<thead>
<tr>
<th></th>
<th>number of students</th>
<th>mean&lt;sup&gt;a&lt;/sup&gt;</th>
<th>st. dev.</th>
<th>min</th>
<th>max</th>
<th>Kolmogorov-Smirnov Z</th>
<th>sig. &lt;sup&gt;b&lt;/sup&gt;</th>
<th>min z-value</th>
<th>max z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text-Only Questions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>136</td>
<td>71.9</td>
<td>15.1</td>
<td>28.6</td>
<td>100.0</td>
<td>0.87</td>
<td>0.43</td>
<td>-2.0</td>
<td>2.1</td>
</tr>
<tr>
<td>High-Average</td>
<td>156</td>
<td>62.6</td>
<td>17.6</td>
<td>0.0</td>
<td>100.0</td>
<td>0.66</td>
<td>0.78</td>
<td>-3.6</td>
<td>1.7</td>
</tr>
<tr>
<td>Low-Average</td>
<td>173</td>
<td>62.4</td>
<td>16.6</td>
<td>14.3</td>
<td>100.0</td>
<td>0.81</td>
<td>0.53</td>
<td>-2.8</td>
<td>2.1</td>
</tr>
<tr>
<td>Low</td>
<td>96</td>
<td>56.5</td>
<td>17.2</td>
<td>0.0</td>
<td>100.0</td>
<td>0.64</td>
<td>0.80</td>
<td>-3.6</td>
<td>2.1</td>
</tr>
<tr>
<td><strong>Illustrated Questions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>136</td>
<td>72.3</td>
<td>11.6</td>
<td>33.3</td>
<td>95.0</td>
<td>0.82</td>
<td>0.51</td>
<td>-1.6</td>
<td>1.9</td>
</tr>
<tr>
<td>High-Average</td>
<td>156</td>
<td>64.9</td>
<td>13.3</td>
<td>23.8</td>
<td>100.0</td>
<td>0.50</td>
<td>0.96</td>
<td>-3.0</td>
<td>2.3</td>
</tr>
<tr>
<td>Low-Average</td>
<td>173</td>
<td>63.9</td>
<td>12.3</td>
<td>20.0</td>
<td>91.3</td>
<td>0.55</td>
<td>0.92</td>
<td>-3.3</td>
<td>1.9</td>
</tr>
<tr>
<td>Low</td>
<td>96</td>
<td>58.8</td>
<td>14.8</td>
<td>18.2</td>
<td>88.9</td>
<td>0.85</td>
<td>0.46</td>
<td>-2.9</td>
<td>1.5</td>
</tr>
</tbody>
</table>

<sup>a</sup> Results are mean values per student.  
<sup>b</sup> Significance set at p ≤ 0.05 (two-tailed)
The Q-Q plots (Appendix F) all plotted close to the normality line and suggested that the data were normally distributed. Lastly, the numerous Kolmogrov-Smirnov tests verified that, when grouped by question type (Table 13), gender (Table 14), or ACT test (Tables 15 – 17), the data yielded Kolmogrov-Smirnov values that were below the critical significance level of $p = 0.05$. The only exception was for the responses to the text-only questions by the men. The responses from this group produced a significant value ($Z = 1.4, p = 0.04$). Mertler and Vannatta (2005) as well as Stevens (2002) suggested that normal distributions have skewness and kurtosis values between -1.2 and 1.2. The data for the male responses to text-only questions yielded skewness and kurtosis values (-0.3 and -0.5 respectively) well within this range, which suggested that these responses were normally distributed. Given that the repeated measures method is robust to violations of normality, this possible minor violation by one sub-category of data was regarded as insignificant; therefore the data were assumed to have met the assumption of multivariate normality.

**Sphericity.**

For each of the repeated measures analyses, Mauchly’s and Box’s tests for sphericity were conducted along with the Greenhouse-Geisser and Huynh-Feldt epsilon tests (Table 18). For all four research hypotheses, Box’s test yielded nonsignificant $p$-values. The results from the Mauchly’s, Greenhouse-Geisser, and Huynh-Feldt tests all yielded values of 1.0, so a $p$-value for the Mauchly’s test could not be calculated. These results indicated that the data did not violate the sphericity assumption.
Summary of Assumptions Testing.

The data analyzed in this study passed all three of the assumptions in the repeated measures method. The data were collected independently of one another, were normally distributed, and maintain sphericity of the error covariances. The only violation to these assumptions was the normality of responses to text-only questions by men, but the repeated measures method is robust to violations of normality, so this minor deviation was not a threat to the overall study.

Table 18. Tests for Sphericity for Each Repeated Measures Analysis

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>IV</th>
<th>Box's Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a &amp; b</td>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>ACT-Reading</td>
<td>9.38</td>
<td>1.04</td>
</tr>
<tr>
<td>ACT-Math</td>
<td>13.81</td>
<td>1.52</td>
</tr>
<tr>
<td>ACT-Science</td>
<td>13.35</td>
<td>1.47</td>
</tr>
</tbody>
</table>

Notes. Results from Mauchly's W, Greenhouse-Geisser, and Huynh-Feldt tests all equal 1.00. For all analyses, the dependent variables were the percentage of correct responses to text-only and illustrated questions.

Identifying Instructor as a Covariate Variable

This study included student response data collected from classes taught by two different instructors. The mean percentage of correct responses by students in Dr. Baker’s classes was higher for both the text-only and illustrated questions. Dr. Anderson’s students (n = 110) correctly answered 57.8% of the text-only questions compared to 65.2% for the students (n = 451) in Dr. Baker’s classes. Similarly, students in Dr.
Anderson’s classes correctly answered a smaller percentage (58.4%) of the illustrated questions than students from Dr. Baker’s class (66.58%). Independent-samples t-tests indicated that there was no significant difference between the percentage of correct responses to the text-only questions ($t_{375} = -2.35$, $p = 0.02$), but the responses to the illustrated questions yielded a significant difference ($t_{375} = -4.84$, $p < 0.001$). The presence of a significant difference in student responses from classes taught by different instructors suggested that factors pertaining to how each instructor incorporated the concept test questions into his courses might affect the percentage of correct responses. To account for this possibility, the variable of instructor was included in all analyses as a covariate.

Mertler and Vannatta (2005) stated that in order for a variable to qualify as a covariate, a variable must satisfy three assumptions.

1. The inclusion of the covariate is relevant to the design of the study.
2. There is a linear relationship between the covariate and the dependent variables.
3. The regression slopes for the covariate and all of the subgroups are equal (homogeneity of regression).

The first assumption was satisfied with the argument that the implementation of concept test questions by different instructors could bias the outcomes. The second assumption was tested by comparing the Pearson correlation coefficients for instructor versus the two dependent variables (the two types of questions). A significant correlation indicated a linear relationship between the two variables (Table 19). A significant correlation (at the $p < 0.05$ level) was observed between the two instructors and student
responses for both the text-only and illustrated questions. In addition, the covariate variable did not significantly correlate with the independent variable which confirmed that these two variables did not measure the same construct. The third assumption was tested by examining the multivariate interaction between all of the variables. If the regression slopes were parallel, then the interaction between all of the variables would not be significant. For Research Question #1, the independent variable was gender, so the covariate could only be used if the interaction between gender, instructor, and student responses to both the text-only and illustrated questions was not significant. The resulting multivariate analysis of variance produced a non significant interaction (Wilk’s $\Lambda = 1.0$, $F_{(1,373)} = 0.01$, $p = 0.95$). For Research Question #2, the independent variables were ACT-Reading, ACT-Math, and ACT-Science categories. Like Research Question number 1, a multivariate analysis indicated that the regression slopes were parallel for the independent variables of ACT-Reading (Wilk’s $\Lambda = 0.99$, $F_{(6, 1104)} = 1.07$, $p = 0.38$), ACT-Math (Wilk’s $\Lambda = 0.99$, $F_{(6, 1104)} = 0.66$, $p = 0.68$), and ACT-Science (Wilk’s $\Lambda =

<table>
<thead>
<tr>
<th>Instructor</th>
<th>Text-Only Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>sd</td>
</tr>
<tr>
<td>Instructor</td>
<td>1.8</td>
</tr>
<tr>
<td>Gender</td>
<td>1.5</td>
</tr>
<tr>
<td>Text-Only Questions</td>
<td>63.8</td>
</tr>
<tr>
<td>Illustrated Questions</td>
<td>65.3</td>
</tr>
</tbody>
</table>

Note. N = 561 students  
*a Results are mean values per student.  
* Significant at $p < 0.05$
1.00, F(6, 1104) = 0.17, p = 0.98). Thus the variable Instructor qualified as a covariate for both research questions and was included in all relevant analyses.

Analysis of Research Questions

This section provides the numeric results from the statistical tests used to evaluate each of the research questions.

*Research Question #1*

Research Question #1 asked whether there was a significant difference in the percentage of correct responses to text-only conceptest questions and illustrated conceptest questions answered by men and women in an undergraduate earth science course. Two research hypotheses were included in this question. Hypothesis 1a stated that there was no statistically significant difference between the percentage of correct student responses to text-only and illustrated conceptest questions asked during an undergraduate earth science course. Hypothesis 1b stated that there was no statistically significant difference between gender and the percentage of correct student responses to text-only and illustrated conceptest questions asked during an undergraduate earth science course. The repeated measures – mixed ANOVA analysis compared both the multivariate interaction between the dependent variables and the independent variables as well as any univariate differences between the two dependent variables. With these results, both research questions were answered using a single statistical test. This section reports the results for each hypothesis.

*Results for research question #1.*

A repeated measures mixed-ANOVA assessed whether there was a significant different in the percentage of correct responses to both types of conceptest questions
(Hypothesis 1a) and whether men and women gave significantly different answers to each question type (Hypothesis 1b). The within-subject effects for question type were not significant when Instructor was included as a covariate ($F_{(1,559)} = 0.02$, $p = 0.89$), but yielded a significant difference if the covariate was not included ($F_{(1,559)} = 4.05$, $p = 0.05$). Similarly, the within-subjects interaction between gender and question type produced no significant difference ($F_{(1,559)} = 2.24$, $p = 0.14$). In contrast, the between-subjects effect for gender indicated a significant difference in the percentage of correct responses provided by men and women ($F_{(1,559)} = 15.18$, $p < 0.001$).

Figure 3. Differences by gender of the percentage of correct responses for each type of question.
A review of the adjusted and unadjusted means (Table 20, Figure 3) showed that men correctly answered a higher percentage of text-only and illustrated questions. In addition, the means are closer together for the text-only questions, which suggested that the observed significant difference might not hold for both question types. Two post hoc one-way ANCOVA analyses were performed to determine whether men and women provided significantly different percentages of correct responses to each type of question. The results (Table 20) indicated that there was a significant difference between responses to both types of questions. Using the adjusted means, men correctly answered 65.1% of the text-only questions compared to 62.1% for the women. The small effect size for this difference (Cohen’s $d = 0.17$) indicated that the average male student correctly answered more questions than 56.8% of the women. For the illustrated questions, men correctly answered 67.7% of the questions compared to 62.5% for the women. The effect size for this difference was larger than for the text-only questions ($d = 0.39$) which indicated that the average male student correctly answered more questions than 65.2% of the women.

Table 20. Adjusted and Unadjusted Means For the Percentage of Correct Responses For Text-Only and Illustrated Questions

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Adjusted$^a$</th>
<th>Unadjusted$^b$</th>
<th>Post Hoc One-Way ANCOVAs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Men</td>
<td>Women</td>
<td>Men</td>
</tr>
<tr>
<td>Text-Only</td>
<td>65.1</td>
<td>62.1</td>
<td>65.3</td>
</tr>
<tr>
<td>Illustrated</td>
<td>67.7</td>
<td>62.5</td>
<td>67.9</td>
</tr>
</tbody>
</table>

Note. Covariate (Instructor) = 1.80

$^a$ Adjusted mean values after taking the covariate instructor into consideration. $^b$ Unadjusted means for each question type * $p<0.01$
**Research Question #2**

Research question number two compared student responses to text-only and illustrated concepttest questions with scores in the ACT-Reading, ACT-Math, and ACT-Science tests. For each student, the ACT scores were placed into one of four quartiles based on national test data. Three repeated measures mixed-ANOVAs assessed any differences based in student prior achievement. Research hypothesis 2a compared student ACT-Reading scores the percentage of correct student responses to text-only and illustrated questions with. Similarly, hypothesis 2b compared the percentage of correct student responses with ACT-Math scores and hypothesis 2c compared the percentage of correct student responses with ACT-Science scores. The results from these analyses were used to assess whether demonstrated student achievement in each area might influence student responses on the two types of questions.

Two variables (instructor and gender) were considered for inclusion as covariate variables. Instructor as a covariate for Research Question #1 significantly influenced the outcome of that analysis, so instructor was considered for Research Question #2. Prior research had suggested that ACT results could vary by gender (Mau and Lynn, 2001). Independent samples t-tests comparing gender to student ACT scores (not their quartile rankings) for each test found no significant difference in ACT-Reading scores ($t_{(559)} = 0.27, p = 0.79$) but significant differences for ACT-Math ($t_{(559)} = 4.34, p < 0.001$) and ACT–Science ($t_{(559)} = 3.48, p = 0.001$) scores. These significant differences suggested that gender should be included as a covariate variable; however adding a second covariate would decrease the power of the study and increase the Type I error. Given that
the variable Instructor was found to significantly impact the analysis for Research Question #1, it was decided to include Instructor but not gender as a covariate.

Results for research hypothesis 2a.

Research hypothesis 2a compared student rankings from the ACT-Reading test to the percentage of correct responses on the text-only and illustrated concept test questions. To test the hypothesis, a repeated measures mixed-ANOVA was performed using student quartile rankings on the ACT-Reading test as the independent variable and the percentage of correct responses on the text-only and illustrated questions as the two dependent variables.

The within-subjects effects determined whether students at each ACT achievement category correctly answered a significantly different percentage of text-only questions than illustrated questions. The repeated measures mixed-ANOVA yielded a non-significant interaction between ACT-Reading achievement categories and the two types of questions \( (F_{(3,556)} = 0.52, p = 0.67) \). The main effects indicated that there was no significant difference between the responses to the two types of questions \( (F_{(1,556)} = 0.08, p = 0.78) \) but the between-subjects effects indicated a significant difference by ACT category \( (F_{(1,556)} = 24.41, p < 0.001, \text{ partial } \eta^2 = 0.11) \).

A review of the adjusted and unadjusted means (Table 21, Figure 4) showed that students who scored in the high ACT-Reading achievement category answered the highest percentage of text-only and illustrated concept test questions followed by students in the high-average and low-average categories. Students who scored the lowest on the ACT-Reading test correctly answered the lowest percentages of questions. Including the variable Instructor as a covariate tended to reduce the scores in the top two categories by
less than half of a percent and raised the scores from the bottom two categories by a similar amount.

Table 21. Adjusted and Unadjusted Means For the Percentage of Correct Responses To Text-Only and Illustrated Questions Sorted by ACT Achievement

<table>
<thead>
<tr>
<th>ACT Achievement Category</th>
<th>n</th>
<th>Adjusted Means</th>
<th></th>
<th>Unadjusted Means</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Text-Only</td>
<td>Illustrated</td>
<td>Text-Only</td>
<td>Illustrated</td>
</tr>
<tr>
<td>ACT-Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>123</td>
<td>70.5</td>
<td>70.7</td>
<td>71.0</td>
<td>71.3</td>
</tr>
<tr>
<td>High-Average</td>
<td>174</td>
<td>65.0</td>
<td>66.7</td>
<td>65.0</td>
<td>66.8</td>
</tr>
<tr>
<td>Low-Average</td>
<td>147</td>
<td>62.0</td>
<td>63.4</td>
<td>61.7</td>
<td>63.1</td>
</tr>
<tr>
<td>Low-Average</td>
<td>117</td>
<td>57.0</td>
<td>60.0</td>
<td>56.8</td>
<td>59.8</td>
</tr>
<tr>
<td>ACT-Math</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>91</td>
<td>71.7</td>
<td>74.7</td>
<td>72.0</td>
<td>75.1</td>
</tr>
<tr>
<td>High-Average</td>
<td>177</td>
<td>64.2</td>
<td>65.7</td>
<td>64.4</td>
<td>65.9</td>
</tr>
<tr>
<td>Low-Average</td>
<td>169</td>
<td>64.4</td>
<td>64.2</td>
<td>64.3</td>
<td>64.0</td>
</tr>
<tr>
<td>Low-Average</td>
<td>124</td>
<td>56.4</td>
<td>59.4</td>
<td>56.1</td>
<td>59.1</td>
</tr>
<tr>
<td>ACT-Science</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>136</td>
<td>71.8</td>
<td>72.2</td>
<td>71.9</td>
<td>72.3</td>
</tr>
<tr>
<td>High-Average</td>
<td>156</td>
<td>62.6</td>
<td>64.9</td>
<td>62.6</td>
<td>64.9</td>
</tr>
<tr>
<td>Low-Average</td>
<td>173</td>
<td>62.3</td>
<td>63.7</td>
<td>62.4</td>
<td>63.9</td>
</tr>
<tr>
<td>Low-Average</td>
<td>96</td>
<td>56.9</td>
<td>59.2</td>
<td>56.5</td>
<td>58.8</td>
</tr>
</tbody>
</table>

Note. Covariate (Instructor) = 1.80

To further investigate these differences, a pairwise Bonferroni post hoc analysis compared student responses between the different ACT-Reading achievement categories (Table 22). Results suggested that student responses were significantly different for all combinations of ACT achievement categories except between the high-average and low-average categories. Large effect sizes were observed for the mean differences between the students from the high ACT-Reading achievement category and all other categories.
The effect size of $d = 0.82$ for the difference between the high and low reading achievement categories indicated that the average student in the high category answered 29.4% more questions than the average student from the low category. The remaining significant differences yielded small effect sizes. These differences indicated that students who scored high on the ACT-Reading test correctly answered a significantly different percentage of questions than students in the other three categories. The smallest significant difference was between students from the low and low-average reading achievement categories. The small effect size of $d = 0.27$ indicated that students in the low-average category correctly answered, on average, 10.6% more questions than students in the low category.

Figure 4. Percentage of correct student responses for each type of question compared to their ACT-Reading achievement categories adjusted for instructor as a covariate. P-values indicate whether the responses by question type are significantly different.

The preceding Bonferroni analysis compared ACT-Reading achievement categories to a combined percentage for both question types. A series of paired t-tests assessed
whether students from any one ACT-Reading achievement category correctly answered a significantly different percentage of text-only questions than illustrated questions (Table 23). The results indicated that students from all four categories of ACT-Reading scores correctly answered a similar, non-significant, percentage of text-only and illustrated questions.

Table 22. Bonferroni Post Hoc Pairwise Comparisons of Correct Student Responses Categorized by ACT Achievement Category

<table>
<thead>
<tr>
<th>ACT Achievement Category</th>
<th>High</th>
<th>High-Average</th>
<th>Low-Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sig.</td>
<td>d</td>
<td>sig.</td>
</tr>
<tr>
<td>ACT - Reading</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-Average</td>
<td>0.004*</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Low-Average</td>
<td>&lt; 0.001*</td>
<td>0.56</td>
<td>0.11</td>
</tr>
<tr>
<td>Low</td>
<td>&lt; 0.001*</td>
<td>0.82</td>
<td>&lt; 0.001*</td>
</tr>
<tr>
<td>ACT-Math</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-Average</td>
<td>&lt; 0.001*</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>Low-Average</td>
<td>&lt; 0.001*</td>
<td>0.63</td>
<td>1.00</td>
</tr>
<tr>
<td>Low</td>
<td>&lt; 0.001*</td>
<td>1.05</td>
<td>&lt; 0.001*</td>
</tr>
<tr>
<td>ACT-Science</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-Average</td>
<td>&lt; 0.001*</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>Low-Average</td>
<td>&lt; 0.001*</td>
<td>0.65</td>
<td>1.00</td>
</tr>
<tr>
<td>Low</td>
<td>&lt; 0.001*</td>
<td>0.95</td>
<td>0.001*</td>
</tr>
</tbody>
</table>

Note. Effect sizes (d) were not calculated for non-significant comparisons. *p < 0.05
Table 23. Paired T-Test Post Hoc Analysis Comparing Student Responses to Different Question Types For Each Achievement Category

<table>
<thead>
<tr>
<th>ACT Achievement Category</th>
<th>Mean Difference</th>
<th>sd</th>
<th>t</th>
<th>df</th>
<th>sig</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACT-Reading</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>-0.30</td>
<td>16.37</td>
<td>-0.20</td>
<td>122</td>
<td>0.84</td>
<td>-</td>
</tr>
<tr>
<td>High-Average</td>
<td>-1.78</td>
<td>17.43</td>
<td>-1.34</td>
<td>173</td>
<td>0.18</td>
<td>-</td>
</tr>
<tr>
<td>Low-Average</td>
<td>-1.32</td>
<td>17.11</td>
<td>-0.94</td>
<td>146</td>
<td>0.35</td>
<td>-</td>
</tr>
<tr>
<td>Low-Average</td>
<td>-2.95</td>
<td>19.04</td>
<td>-1.67</td>
<td>116</td>
<td>0.10</td>
<td>-</td>
</tr>
<tr>
<td><strong>ACT-Math</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>-3.09</td>
<td>14.56</td>
<td>-2.02</td>
<td>90</td>
<td>0.05*</td>
<td>0.21</td>
</tr>
<tr>
<td>High-Average</td>
<td>-1.57</td>
<td>18.43</td>
<td>-1.13</td>
<td>176</td>
<td>0.26</td>
<td>-</td>
</tr>
<tr>
<td>Low-Average</td>
<td>0.27</td>
<td>17.91</td>
<td>0.19</td>
<td>168</td>
<td>0.85</td>
<td>-</td>
</tr>
<tr>
<td>Low-Average</td>
<td>-2.99</td>
<td>17.28</td>
<td>-1.93</td>
<td>123</td>
<td>0.06</td>
<td>-</td>
</tr>
<tr>
<td><strong>ACT-Science</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>-0.46</td>
<td>15.49</td>
<td>-0.35</td>
<td>135</td>
<td>0.73</td>
<td>-</td>
</tr>
<tr>
<td>High-Average</td>
<td>-2.25</td>
<td>19.09</td>
<td>-1.47</td>
<td>155</td>
<td>0.14</td>
<td>-</td>
</tr>
<tr>
<td>Low-Average</td>
<td>-1.48</td>
<td>17.20</td>
<td>-1.13</td>
<td>172</td>
<td>0.26</td>
<td>-</td>
</tr>
<tr>
<td>Low-Average</td>
<td>-2.24</td>
<td>17.90</td>
<td>-1.23</td>
<td>95</td>
<td>0.22</td>
<td>-</td>
</tr>
</tbody>
</table>

Note. Paired t-tests do not include the variable Instructor as a covariate
* p < 0.05

**Results for research hypothesis 2b.**

Research hypothesis 2b compared student rankings from the ACT-Math test to the percentage of correct responses on the text-only and illustrated conceptest questions. To test the hypothesis, a repeated measures mixed-ANOVA was performed using student quartile rankings on the ACT-Math test as the independent variable, the percentage of
correct responses on the text-only and illustrated questions as the two dependent variables, and Instructor as a covariate.

The within-subjects effects determined whether students at each ACT achievement category correctly answered a significantly different percentage of text-only questions than illustrated questions. The repeated measures mixed-ANOVA yielded a non-significant interaction between ACT-Math achievement categories and the two types of questions ($F_{(3,556)} = 1.11, p = 0.34$). The main effects indicated that there was not a significant difference between the responses to the two types of questions ($F_{(1,556)} = 0.01, p = 0.93$) but the between-subjects effects indicated a significant difference by ACT category ($F_{(1,556)} = 30.10, p < 0.001, \text{partial } \eta^2 = 0.14$).

![Figure 5. Percentage of correct student responses for each type of question compared to their ACT-Math achievement categories. P-values indicate whether the responses by question type are significantly different.](image-url)
A review of the adjusted and unadjusted means (Table 21, Figure 5) showed that students who scored in the high ACT-Math achievement category correctly answered the highest percentage of text-only and illustrated conceptest questions followed by students in the high-average and low-average categories. Students who scored the lowest on the ACT-Math test correctly answered the lowest percentages of questions. The students in the two middle categories appeared to answer a similar percentage of text-only and illustrated questions. Students in all of the categories correctly answered a higher percentage of illustrated questions than text-only questions, except for the students in the low-average category. Their scores on text-only questions were slightly higher than their responses to the illustrated questions. Including the variable Instructor as a covariate tended to reduce the scores in the top two categories by less than half of a percent and raised the scores from the bottom two categories by a similar amount.

To further investigate whether the differences between the ACT achievement categories were significant, a pairwise Bonferroni post hoc analysis compared student responses between the different ACT-Math achievement categories (Table 22). The results indicated that student responses were significantly different for all combinations of ACT achievement categories except between the high-average and low-average categories. Largest effect size (d = 1.05) was observed for the mean differences between the students from the high and low ACT-Math achievement categories. This difference indicated that students in the high math achievement category correctly answered, on average, 35.3% more questions than students in the low achievement category. The smallest significant difference was between students in the low and low-average math achievement categories. The effect size (d = 0.42) indicated that students in the low-
average category correctly answered, on average, 16.3% more questions than students in the low achievement category. These differences indicated that students who scored high on the ACT-Math test correctly answered a significantly different percentage of questions than students in the other three categories. A review of the mean values (Table 16) and graph (Figure 5) indicated that students in the high category were more often correct than students in the high-average category and the low-average category. Students who scored the lowest on the ACT-Math test also correctly answered the lowest percentage of questions. It was also noted that, contrary to the other categories, students in the low-average category correctly answered a slightly higher percentage of text-only questions than illustrated questions.

The Bonferroni analysis compared ACT-Math achievement categories to a combined percentage for both question types. A series of paired t-tests assessed whether students from any one ACT-Math achievement category correctly answered a significantly different percentage of text-only questions than illustrated questions (Table 21). The results indicated that students from the high ACT-Math achievement category answered a significantly higher percentage of illustrated questions than text-only questions with a small effect size of $d = 0.21$. This difference indicated that the mean score for the illustrated questions was 8.3% higher than the average score for the text-only questions. Students from the remaining three categories of ACT-Math scores correctly answered a similar, non-significant, percentage of text-only and illustrated questions. It was also noted that the significance level for students from the low ACT-Math achievement category was close to the threshold value of $p = 0.05$. 
Results for research hypothesis 2c.

Research hypothesis 2c compared student rankings from the ACT-Science test to the percentage of correct responses on the text-only and illustrated concepttest questions. To test the hypothesis, a repeated measures mixed-ANOVA was performed using student quartile rankings on the ACT-Science test as the independent variable, the percentage of correct responses on the text-only and illustrated questions as the two dependent variables, and the variable Instructor as a covariate.

The within-subjects effects determined whether students at each ACT category correctly answered a significantly different percentage of text-only questions than illustrated questions. The repeated measures mixed-ANOVA yielded a non-significant interaction between ACT-Science categories and the two types of questions ($F_{(3,556)} = 0.32, p = 0.81$). The main effects indicated that there was not a significant difference between the responses to the two types of questions ($F_{(1,556)} = 0.05, p = 0.85$) but the between-subjects effects indicated a significant difference by ACT category ($F_{(1,556)} = 29.67, p < 0.001, \text{partial } \eta^2 = 0.14$).

A review of the adjusted and unadjusted means (Table 21, Figure 6) showed that students who scored in the high ACT-Science achievement category correctly answered the highest percentage of text-only and illustrated concepttest questions followed by students in the high-average and low-average categories. Students who scored the lowest on the ACT-Science test correctly answered the lowest percentages of questions. Students in all of the categories answered a higher percentage of illustrated questions than text-only questions. Including the variable Instructor as a covariate tended to reduce the scores in the top two categories by less than half of a percent and raised the scores from the
bottom two categories by a similar amount. To further investigate whether the differences between the ACT achievement categories were significant, a pairwise Bonferroni post hoc analysis compared student responses between the different ACT-Science categories (Table 22). The results indicated that student responses were significantly different for all combinations of ACT achievement categories except between the high-average and low-average categories. The largest difference was between students in the high and low science achievement categories. The large effect size (d = 0.95) indicated that students in the high achievement category answered an average of 32.9% more questions than students in the low achievement category.

![Figure 6](image.png)

Figure 6. Percentage of correct student responses for each type of question compared to their ACT-Science achievement categories. P-values indicate whether the responses by question type are significantly different.

The smallest difference was between students in the low-average and low science achievement categories. The effect size (d = 0.32) indicated that students in the low-
average category answered an average of 12.6% more questions than students in the low achievement category. These differences indicated that students who scored high on the ACT-Science test correctly answered a significantly different percentage of questions than students in the other three categories.

A review of the mean values (Table 17) and graph (Figure 6) indicated that students in the high category answered the highest percentage of questions followed by the students in the high-average category and the low-average category. Students who scored the lowest on the ACT-Science test also correctly answered the lowest percentage of questions.

The preceding Bonferroni post hoc analysis compared ACT-Science achievement categories to a combined percentage for both question types. A series of paired t-tests assessed whether students from any one ACT-Science achievement category correctly answered a significantly different percentage of text-only questions than illustrated questions (Table 21). The results indicated that students from all four categories of ACT-Science scores correctly answered a similar, non-significant, percentage of text-only and illustrated questions.

*Summary of results for research question #2.*

The three repeated measures mixed-ANOVA analyses yielded similar results. For all three ACT tests there were no significant differences in the percentage of correct student responses to text-only and illustrated concept test questions. The post hoc analyses of possible differences within a given ACT achievement category found that students from only one category (high ACT-Math) correctly answered a significantly different percentage of questions. When analyzed for between-subject differences, all three tests
yielded significant differences between the four ACT achievement categories with the largest effect sizes for differences between the high ACT achievement category and the remaining three categories.

Summary of Results

The two research questions and five research hypotheses presented herein were tested using inferential statistics. The results (summarized in Table 24) indicated that men typically answered a larger percentage of questions than women. This difference was significant but yielded a small effect size ($d = 0.21$). The analyses of ACT scores demonstrated that for ACT-Reading, ACT-Science and ACT-Math scores, the percentage of correct student responses to illustrated questions were not significantly larger than responses to text-only questions. Pairwise comparisons revealed that responses from each ACT category were significantly different from all other categories except for comparisons between responses from the low-average and low categories. The largest effect sizes were for the differences between the high and low ACT categories. Paired t-tests indicated that students from each ACT category did not correctly answer a significantly larger percentage of text-only questions than illustrated questions. The lone exception was students who scored high on the ACT-Math test. These students correctly answered a significantly higher percentage of illustrated questions than text-only questions.
Table 24. Summary of Analytical Results For Each Research Hypothesis

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Interaction on Combined Dependent Variable</th>
<th>Dependent Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Text-Only</td>
<td>Illustrated</td>
</tr>
<tr>
<td>Question Type</td>
<td>No significant difference</td>
<td>n/a</td>
</tr>
<tr>
<td>Gender</td>
<td>No significant difference</td>
<td>Significant differences (d = 0.17)</td>
</tr>
<tr>
<td>ACT-Reading</td>
<td>No significant difference</td>
<td>Significant differences between ACT achievement categories</td>
</tr>
<tr>
<td>ACT-Math</td>
<td>No significant difference except for high ACT responses</td>
<td>Significant differences between ACT achievement categories</td>
</tr>
<tr>
<td>ACT-Science</td>
<td>No significant difference</td>
<td>Significant differences between ACT achievement categories</td>
</tr>
</tbody>
</table>

Note. n/a = not applicable. Significance set at p < 0.05
CHAPTER V

DISCUSSION OF RESEARCH RESULTS

Introduction

This chapter interprets the analytical results described in Chapter IV and is divided into three sections. The first section describes how the results answer the research questions and hypotheses. The second section discusses how these results compare with the previous research findings presented in Chapter II, and the third section discusses ways in which the findings from this study can be applied to geoscience education. Topics for further research are suggested throughout this discussion.

Conclusions Drawn From Research Hypotheses

This section interprets the research results for both research question as well as the five research hypotheses. It also discusses the implications of including the variable Instructor as a covariate.

*Instructor as a Covariate*

Including instructor as a covariate changes the significance of the within-subjects analysis) for the responses to the two types of questions. Without the variable, the analysis yields a significant difference ($p = 0.05$), but when Instructor is added as a covariate the difference is eliminated ($p = 0.89$). This change in significance suggests that differences in how the two instructors taught their classes may have influenced the answers provided by the students. Several factors might explain the observed differences...
in percentage of correct responses generated for the classes taught by each instructor. First, the instructors taught a different number of classes and contributed an unequal number of students to the project. Dr. Baker taught seven classes and contributed approximately 80% of the students (n = 451) whereas Dr. Anderson taught just two classes and contributed approximately 20% of the students (n = 110). Other possible explanations for the observed difference in student responses between the two instructors include differences in how the conceptest questions were administered, differences in the questions asked, or differences in teaching style. Differences in student demographics or prior academic achievement can be ruled out because the student characteristics did not significantly vary between the instructors. There was also no difference in student responses by time of day when the classes were taught. The two instructors used a common set of questions and both were proficient in implementing the peer instruction pedagogy in their classes, therefore it seems most likely that the observed differences by instructor are due to the unequal size in the number of students taught by each instructor.

Further research on how instructors implement peer instruction and conceptest questions may yield new insight on this observed difference. For example, students may be influenced by the level to which the instructor has adopted a constructivist, inquiry-driven pedagogy. The Reformed Teaching Observation Protocol (RTOP: Piburn & Sawada, 2000) is one possible instrument may assess whether differing pedagogies influence student responses to conceptest questions. This valid and reliable instrument contains 25 items that a trained observer evaluates to assess how well the instructor incorporates constructivist and inquiry-driven practices. The protocol yields a numeric
score that could be compared with student responses to the associated concept test questions.

Research Question #1

Research Question #1 - Is there a significant difference in the percentage of correct responses by men and women to text-only concept test questions and illustrated concept test questions asked during an undergraduate earth science course?

Research Question #1 assesses whether men and women answer a different percentage of correct responses to text-only and illustrated concept test questions. Two research hypotheses are associated with this question. Hypothesis #1a considers whether the students gave similar percentages of correct responses for text-only and illustrated questions. Hypothesis 1b includes gender as a between-subjects dependent variable and considers whether men and women correctly answer a similar percentage of text-only and illustrated questions.

Research hypothesis #1a.

Research Hypothesis #1a. H₀: There is no statistically significant difference between the percentage of correct student responses to text-only concept test questions versus illustrated concept test questions asked during an undergraduate earth science course.

Research Hypothesis 1a considers whether there was any difference in the percentage of correct responses between the two types of concept test questions (text-only versus illustrated questions). The within-subjects analysis of the main effect for question type provides a non-significant result (p = 0.89) which fails to reject the null hypothesis and suggests that the way in which the question is worded (with or without an illustration) does not affect the percentage of students who answer correctly.
**Research hypothesis #1b.**

Research Hypothesis #1b. H₀: There is no statistically significant difference between gender and the percentage of correct responses correct student responses to text-only conceptest questions versus illustrated conceptest questions asked during an undergraduate earth science course.

Research Hypothesis 1b considers whether men and women correctly answer a similar percentage of questions. The non-significant within-subjects interaction between gender and question type (p = 0.14) suggests that men and women similarly respond to each type of question. That is, one gender does not correctly answer a larger percentage of text-only questions but the other gender answers a larger percentage of illustrated questions. This is illustrated in Figure 3 by the non-crossing lines. Yet the significant between-subjects effect of gender and question type (p < 0.001) suggests that men correctly answer a higher percentage of both text-only and illustrated conceptest questions. This is shown in Figure 3 by the gap between the two lines. The post hoc ANCOVA analyses indicate that the observed gender gap is largest for the correct responses to the illustrated questions (d = 0.37) than responses to the text-only questions (d = 0.17). Cohen (1988) qualitatively defined effect sizes of these magnitudes as small effect sizes and suggested that such differences would be difficult to observe in the classroom. Yet a difference of d = 0.37 also indicates that the average man answers more questions than 65.2% of the women. This is not a trivial difference and should be observable during the course of a typical class.

**Research Question #2**

Research Question #2: Are there significant differences in the percentage of correct responses for the two types of conceptest questions when compared with different levels of prior achievement as measured by student ACT scores?
Research Question #2 compares student responses to text-only and illustrated conceptest questions to student prior achievement in reading, math, and science as measured by student ACT scores. This question is divided into three related research hypotheses. Hypothesis #2a compares student prior reading achievement (as measured by scores on the ACT-Reading test) to the percentage of correct responses to both the text-only and illustrated questions. Research Hypothesis #2b compares student prior achievement in mathematics (as measured by scores on the ACT-Math test) to the percentage of correct responses on the text-only and illustrated questions. Research Hypothesis #2c compares student prior achievement in science (as measured by scores on the ACT-Science test) to the percentage of correct responses on the text-only and illustrated questions. The results for all three hypotheses were nearly identical and as such are discussed together rather than separately.

*Research hypotheses #2a, 2b, and 2c.*

Research Hypothesis #2a-c. $H_0$: There is no statistically significant difference between the percentage of correct responses to text-only or illustrated conceptest questions asked during an undergraduate earth science course and student scores on the ACT-Reading test (Hypothesis #2a), ACT-Math test (Hypothesis #2b), or ACT-Science test (Hypothesis #2c).

For all three hypotheses, the non-significant results for reading ($p = 0.67$), math ($p = 0.34$), and science ($p = 0.81$) indicate that student responses to text-only and illustrated conceptest questions are indistinguishable even when considering prior achievement. The paired t-test post hoc analyses (Table 23) further support this finding. Only students with high demonstrated math achievement correctly answered a significantly different percentage of illustrated and text-only questions, yet the small effect size suggests that this difference, though significant, would be difficult to observe during peer instruction.
These findings are consistent with the results from Research Hypothesis #1a which also finds no significant difference in student responses to the two types of questions.

Even though student achievement in reading, math, and science fails to discriminate student responses to text-only and illustrated conceptest questions, the data do suggest that student responses and prior achievement are related. The significant between-subjects effects for reading (p < 0.001), math (p < 0.001), and science (p < 0.001) coupled with the significant Bonferroni post hoc analyses (Table 22) indicate that student who have demonstrated high achievement in any of these three areas also correctly answer a larger percentage of questions than students who have demonstrated a low level of achievement. Such a relationship may also be related to student test-taking skills. Students may use a similar set of psychological or academic skills to answer both standardized tests and conceptest questions. Additional research is needed to determine those psychological skills needed to correctly answer conceptest questions.

Summary of Findings

The data from this study suggest the following findings:

1. There is no statistically significant difference between student responses to text-only and illustrated conceptest questions (Research Hypothesis #1a).
2. When answers to conceptest questions are compared within each gender, men and women correctly answer a similar percentage of text-only and illustrated questions (Research Hypothesis #1b).
3. Yet, when answers are compared between the genders, men did correctly answer a significantly higher percentage of text-only and illustrated conceptest questions (Research Hypothesis #1b).
4. Demonstrated prior achievement in reading, math, and science fails to distinguish student responses to text-only and illustrated questions (Research Hypotheses #2a, 2b, and 2c). That is, regardless of the level of achievement, students correctly answer a similar percentage of questions.

5. Yet when student responses provided by students from each of the three prior achievement categories (reading, math, and science) are compared, significant differences emerge. Students who score highly on any of the three tests also scored significantly higher than students from the other achievement categories (Research Hypotheses #2a, 2b, and 2c). That is, students who do well on these tests also do well on both types of conceptest questions. Conversely, students who perform poorly on any of the tests also perform poorly on both types of questions.

Connections With Prior Research

The results from this study add to the growing body of knowledge on differences in student responses to active learning pedagogies. Specifically, this study provides information on the possible role of gender and prior achievement in the answering of conceptest questions.

Gender

This study investigates the student responses to conceptest questions. In contrast, the small number of studies that have previously addressed the issue of gender and student responses using personal response systems have focused on investigating differences in student attitudes (Kay, 2009; et al., 2008) or student performance in the class (King & Joshi, 2008; Lorenzo, et al., 2006). Kay et al. (2008) argues that gender differences in
attitudes towards personal response systems can be explained by differences in self-efficacy towards using a computer. In addition, the observed gender differences in responses to conceptest questions may reflect differences in student spatial abilities or differences in the skills used by each gender to correctly answer these questions.

Differences in attitudes towards computers.

The results from this study suggest that men and women provide similar responses to both text-only and illustrated questions, however the results also suggest that gender is a factor in determining responses to conceptest questions. The lack of a significant difference between question type and gender suggests that there is little difference in how men and women respond to each type of question.

The significant difference within each type of question suggests that an unknown factor may influence how men and women select the correct answer. The effect sizes, especially for the illustrated questions, indicate that this difference is not trivial. There are several possible explanations for why men correctly answered a higher percentage of questions than women including differences in attitudes towards using a personal response system, differences in spatial ability, or differences in how each gender attacks each problem. In all cases, one must keep in mind that social norms may mediate any of these factors (Massa et al., 2005; Saucier et al., 2002).

Differing attitudes towards using a personal response system.

Men may be more comfortable with using a technology which might bolster their self-efficacy and encourage them to provide the correct response. Previous studies report a small but persistent trend in which men are more comfortable than women at using computers (Kay, 2008; Terlicki & Newcombe, 2005). Kay (2009) suggests that tangential
affective or cognitive attitudes towards computers may bias students in their use of personal response systems, but the limited research on gender differences in attitudes towards personal response systems also reports contradictory results. MacGeorge (2008a) report that women gave more positive evaluations of the technology than men but Kay (2009) found that men gave higher ratings than women. The small effect sizes reported in this study suggest that any observed differences by gender are very small. It is possible that any variations in the attitudes are contextually driven (Kay, 2008) and may not indicate an innate difference in how men and women respond to conceptest questions. Also, self-reported attitude surveys do not correlate with actual student use and attitudes towards computers (Kay, 2008).

It is difficult to explain how attitudes towards computers would influence student attitudes towards personal response systems. The transmitter is based on the technology used in a television remote control (Abrahamson, 2006) and requires little or no knowledge of computers to use. As Kay (2009) describes, the system is easy to use and students do not complain of struggles using the devices. In addition, today’s university students are using hand-held electronic communication devices such as wireless text-messaging at an increasing rate (Faulkner & Culwin, 2004). Many of the earlier studies on student attitudes towards computers investigated learning and attitude effects when students sat at a computer to learn a prescribed set of content material. Personal response systems and personal electronic communication devices such as wireless text-messaging require minimal understanding of a computer to operate. It seems prudent to proceed cautiously when comparing past results on student attitudes towards computers with student usage of personal response systems. Today’s undergraduate students (the sample
of this study) may be more accepting of the new forms of electronic communications
technologies than previous subjects, and so gender differences observed in the past may
not be present in today’s university student population. For example, Price (2006) found
women taking an online course performed significantly better than their male
counterparts and reported a higher level of self-confidence. In addition, Prinsen, Volman,
& Terwel (2007) reviewed the literature in computer-mediated communication and noted
that the research suggests that men and women use computers in different ways.

More research is needed on the relationship between gender and student attitudes
towards using personal response systems and answering concept test questions. For
example, many of the studies cited by Prinsen et al. (2007) investigated school-age
students rather than undergraduate students. A study that assessed possible gender
differences in attitudes towards using personal electronic communication devices among
undergraduate students might provide the data needed to resolve this question. Such a
study could include an attitude survey as well as direct observation of student behaviors
in a class that uses personal response systems. Interviews of students might also reveal
and differences in attitudes in how men and women perceive personal response systems
and concept test questions.

*Differences in spatial ability.*

Differences in spatial ability may provide a second explanation for the observed
gender differences. Men are known to score higher than women on tests of spatial ability
and reasoning (Geary et al., 1992; Linn & Peterson, 1985; Nordvik & Amponsah, 1998)
and women tend to score higher than men on measures of verbal reasoning (Halpern et
al., 2007). Spatial ability has also been shown to correlate with success in understanding
geoscience concepts (Black, 2005). The illustrated conceptest questions may require
students to analyze and interpret a visual construct (graph, diagram, photograph, object,
or map) before they can accurately select the correct answer. As Kastens and Ishikawa
(2006) point out, geoscientists use an array of spatial abilities to understand the world
around them and to display geoscience information. It seems reasonable to assume that
the information provided on the illustrated questions could require a similar cognitive
skill set. In this context, it is not surprising that men correctly answered a significantly
higher percentage of questions than women. Prior findings also suggest that differences
in spatial ability may be most closely linked to the ability to mentally rotate an object
(Black, 2005; Orion et al., 1997), so it is possible that a similar connection may exist with
answers to illustrated conceptest questions.

Women typically score higher than men on measures of verbal reasoning, so it was
expected that the women would have correctly answered a higher percentage of text-only
questions than the men. The students in this study exhibited the opposite behavior with
the men correctly answering a significantly higher percentage of questions than the
women. Possible differences in student spatial ability may explain why the men were
significantly better than the women at correctly answering the text-only questions. Many
of the text-only questions included in this study may have contained a spatial component.
For example, question T-4 (Appendix C) refers to a fictional fault and three cities located
at different distances from the fault. One way of answering this question was to mentally
construct a map showing the fault and three cities. Another question (T-19) asks what
information would be lost if all the minerals in an igneous rock were the same color. One
way to ascertain the correct answer is to mentally visualize a rock with white minerals
and check to see whether one can still determine the rock’s texture, cooling rate, or composition. Kastens and Ishikawa (2009) found that students used their spatial ability to solve a geologic field problem. If students access their spatial abilities to solve text-only questions, then their answers could be influenced by how well they can visualize the data presented within the question. Under such conditions, the documented tendency for men to possess better spatial abilities might explain why the men in this study correctly answered a larger percentage of text-only questions than the women.

Hamilton (1998) provides a possible explanation for the observed differences in responses between men and women. She reports that for high school students, the spatial-mechanical reasoning displayed the largest gender effect including items that require student visualization or knowledge acquired outside of school. During interviews, students used visual gestures and words to describe spatial reasoning. These findings are consistent with the viewpoint that earth science is a highly visual discipline that contains numerous concepts that transcend the typical student’s experience. Thus it appears that students may tap into their ability to mentally manipulate spatial data when answering content-related questions. Given the visual nature of the geosciences, it seems likely that a similar process may occur when students answer text-only conceptest questions.

Further research connecting student spatial ability and answers to conceptest questions is needed. A series of baseline studies could ascertain whether the gender differences observed in this study are significantly related to one or more measures of spatial ability such as mental rotation, spatial perception, and spatial visualization. A research program along this line could use the instruments reported by Black (2005) to ascertain which (if any) of these constructs relates to correct student responses to
conceptest questions. Any significant differences could then be tested across a variety of institutional and class settings to determine whether the differences observed in this study hold true at other schools or other disciplines.

*The possible influence of other psychological and social variables.*

It is also possible that cognitive or affective variables other than spatial ability influence men and women when they respond to conceptest questions. Cognitive load is one possible variable. Cognitive load is the ability of students to process new and difficult information (Sweller, Chandler, Tierney, & Cooper, 1998; van Merriënboer & Sweller, 2005). If a subject is difficult to learn, a student’s working memory can only process a limited amount of data. The student will struggle to learn a new concept if the amount of data is larger than the capacity of the student’s working memory. It is possible that students who correctly answer a large percentage of questions also possess a greater capacity to work with complex information whereas students who answer few questions quickly become frustrated and struggle with issues pertaining to cognitive load.

Student prior experiences may also contribute to their ability to answer both types of conceptest questions. Research suggests that student experiences affect the development of high-order thinking skills such as spatial ability. For example, Terlicki and Newcombe (2005) found that student experiences relate to spatial ability. Their results suggest that men are more likely to play computer games that develop spatial ability whereas women do not. Such social norms may lead to differential development of the process skills needed to correctly answer both text-only and illustrated conceptest questions.

Another possible factor is student response time. Gonzalez-Espada and Bullock (2007) found that students who provide correct responses to conceptest questions are
quicker to select an answer than students who provide an incorrect response. Lasry, Mazur, Watkins, & Van Wieren (2009) also found that students respond faster if they know the answer. They also report that response times are inversely related to student confidence in their answers. That is, students who are confident answer faster than students who are not confident. It is possible that the male students in this study were more confident than their female peers and quickly answered each question but the women were not given sufficient time to arrive at the correct response. Instead the women may have felt overly pressured to choose an answer and just decided to randomly select a response to receive credit.

The preceding list of possible explanations is not exhaustive but does illustrate the need for additional research in this area. Cognitive load is currently assessed by interviewing students or by analyzing test questions to determine the difficulty in answering each question. Such an analysis has not been done for conceptest questions of any type but could determine whether cognitive load is a contributing factor in answering conceptest questions. In addition, interviewing students and asking them to describe their thought processes while answering text-only and illustrated questions would be a valuable means of identifying any other cognitive or affective variables that may influence student responses. This process is called a “think aloud” (Cotton & Gresty, 2006) and could also identify the psychological process skills (such as abstract thinking) required to solve these questions. The results might identify possible differences in how men and women solve text-only and illustrated questions. It would also provide data needed to evaluate the possible spatial nature of the text-only questions.
Prior Achievement

Results from this study also suggest that student prior achievement in reading, math, and science does not influence responses to text-only and illustrated questions. The results further suggest that conceptest questions are a valid tool of formatively assessing student comprehension.

Prior achievement and response differences between text-only and illustrated questions.

Student responses to text-only and illustrated conceptest questions do not significantly vary when compared to student rankings based on scores from their ACT-Reading, ACT-Math, and ACT-Science tests. It had been expected that students with high prior achievement in reading comprehension (the ACT-Reading test) would correctly answer a significantly higher percentage of text-only questions than illustrated questions. Similarly, it was originally expected that high prior achievement in math (ACT-Math test) or science (ACT-Science test) would correctly answer a higher percentage of illustrated questions than text-only questions. The finding that students at all achievement levels correctly answered a similar percentage of text-only and illustrated questions for all three tests suggests that student proficiency in reading, math, or science may not influence or affect the ability to answer text-only and illustrated conceptest questions. The non-significant post hoc analyses comparing student achievement level on each of the three tests with student responses to text-only and illustrated questions further support the finding that the inclusion of an illustration does not appear to benefit student answers.
These results appear to contradict findings from multimedia learning that find a significant difference in student comprehension between students who read text-only material and those who read text accompanied by an illustration. For example, students who view a text with an associated illustration experience higher learning gains than students who just read a series of text (Butcher, 2006; Hegarty & Just, 1993) and when the illustration is placed adjacent to the accompanying text (Mayer, 2001). This connection between relative proximity of an illustration to its associated text is called the Spatial Contiguity Principle (Mayer, 2001). Similarly, Uesaka, Manalo, and Ishikawa (2006) found that math students who used diagrams to solve problems achieved higher scores than students who relied solely on computations. Mayer (1989) also found that labeled illustrations fostered higher student recall and problem-solving when reading text on a scientific topic suggesting that illustrations aid in student learning. Yet, one must keep in mind that this study considers student responses to test questions where the illustration is an integral part of the question whereas Mayer’s (1989) students used the illustrations as aids in understanding the accompanying text. In other words, conceptest questions do not use illustrations to augment and clarify information provided in the text but rather include vital data needed to answer the question. For example, Question I-7 (Appendix D) provides a map of tectonic plate boundaries and asks the student to identify the number of plates on the map. If students use the illustration in this question in the same manner that Mayer’s students used their illustrations, the text would include all of the data necessary to correctly answer the question and the map would visual reinforce the information provided in the text. The fact that illustrations in conceptest questions serve a different function than illustrations in a textbook or online suggests that additional
research is needed to explore the dynamics of how students make use of illustrations in conceptest questions.

Only one student group (high ACT-Math scores) correctly answered a significantly higher percentage of illustrated questions than text-only questions. Yet the results suggest that student ability in mathematics may influence their responses, and is consistent with the findings Curcio (1987) who found that math achievement of fourth and seventh grade students is significantly correlated to levels of student prior achievement. It is not known whether a similar relationship exists for undergraduate students.

In the present study, the observed relationship between high ACT-Math scores and significantly a higher percentage of correct responses to illustrated questions may reflect one or more explanatory factors. Students with high math achievement may possess a more sophisticated ability to solve a given problem, they may have been more familiar in reading graphs, or they may be poor readers. The latter seems unlikely given that reading comprehension does not significantly differentiate responses to illustrated and text-only questions. It is also unlikely that familiarity with graphs is responsible for the observed difference. Very few of the questions required students to analyze a graph and draw conclusions so it appears that the skill of reading and interpreting a graph is not the sole determiner. More likely the students who score the highest on the ACT-Math test already possess the analytical and psychological skills needed to interpret data presented in a visual form such as a map, graph, or photograph and may have drawn the necessary inferences or conclusions from those data. A think aloud study would be one method of determining which skills contributed to the higher scores for the illustrated questions.
Response differences within different achievement groups.

For all three tests (ACT-Reading, ACT-Math, ACT-Science) students from within each achievement category (high, high-average, low-average, low) correctly answered a significantly different percentage of questions. For example, students from the high achievement category consistently answered a higher percentage of questions than students in the other three categories. The results are consistent across all categorical combinations except the responses between students in the high-average and low-average categories. There are several possible explanations for these observations. First, it is tempting to conclude that conceptest questions are an accurate measure of student understanding and can be used as a formative assessment tool. Yet this study does not contain an independent measure (such as grades) for student understanding, so it is not known if answers to these questions reflect student understanding of the course material. Another possible explanation is that student responses to conceptest questions are solely based on prior levels of achievement or prior content knowledge. Under this scenario, the students respond to what they already know rather than what they have learned in class. Yet this explanation does not account for the fact that this earth science course covers content that is typically unfamiliar to the student (Libarkin, Anderson, Dahl, Beilfuss, & Boone, 2005).

The most likely explanation is that the student responses to conceptest questions reflect the general student academic preparedness to succeed in an undergraduate earth science course. The stated intent of the ACT tests is to measure student academic preparedness for success at the university level (ACT, 2008a). The data from this study support this assertion made by ACT. Furthermore, O’Reilly and McNamara (2007) found
that science knowledge correlates with achievement in science. In the present study, students with higher scores in reading, math, and science answered a higher percentage of conceptest questions. However one must keep in mind that only 25% of the participants in this study (141 students) fell into the same achievement category for all three tests which means that three quarters of the students fell into different achievement categories, so making sweeping conclusions concerning ACT category and student responses must be viewed as tentative at best.

Additional studies are needed to determine what factors are present in students who score high on the ACT but lacking in students who score low on the same measures. Once these factors have been identified, a comparison with student responses to conceptest questions could provide meaningful data on the factors that contribute to student responses.

Applications to Geoscience Education

The results from this study are relevant to the use of conceptest questions and personal response systems in a large earth science course. Such relevant conclusions include the observation that wording of a question does not affect student response rates, student ACT scores can provide data on student success at conceptest questions, and the possible impacts of spatial efficacy on students to answer conceptest questions on men and women. In addition, factors pertaining to how instructors implement and utilize conceptest questions may also influence student responses.

Had Mayer’s (2001) Spatial Contiguity Principle applied to conceptest questions, then students would have correctly answered a higher percentage of illustrated questions than text-only questions. The observation that the type of question does not change the
percentage of correct responses suggests that instructors need not worry about the format of the question (text-only versus illustrated) and rather should focus on whether the question is well written and relevant to the classroom discussion. In addition, later research on student responses to conceptest questions need not separate out questions according to whether they include an illustration.

Additional research is needed to determine whether answering text-only questions causes students to access different problem-solving processes than answering illustrated questions. It is possible that the uniform percentages observed for the two types of questions were confounded by an inherent spatial component within the text-only questions. If this is the case, then the two types of questions access the same skill set, but if text-only questions can be identified that require a separate set of skill to answer than for illustrated questions, then this recommendation would not be valid. Additional research is needed to resolve this issue.

The uniform results across the three ACT tests suggest that student success in any of these areas may not influence their scores on conceptest questions. The uniformity of data across the three tests also suggests that instructors and researchers could substitute a more general measure of student academic achievement (such as the ACT-Composite score) in place of the content-specific measures. It is tempting to suggest that instructors could use student ACT scores to predict student success at answering conceptest questions and provide a metric for identifying possible at-risk students before they fail. Yet the analyses reported in this study only suggest that such a connection exists and does not directly address this question. A more rigorous analysis using a multiple regression model to could ascertain which variables best predict student success at conceptest questions, and
inclusion of data on student exams or grades could connect responses to conceptest questions to course achievement.

The results from this study also suggest that providing opportunities in class for students to develop their spatial abilities may influence their responses on conceptest questions. The connection between spatial ability and achievement in the geosciences is well documented (Black, 2005; Kastens & Ishikawa, 2006). Furthermore, Hamilton (1998) found that spatial ability resulted in gender differences similar to the ones observed in this study. She also found that student experiences outside of the classroom may affect achievement in the classroom. These studies suggest that spatial ability may be a contributing factor in how students respond to conceptest questions. Orion, Ben-Chaim, and Kali’s (1997) observation that introductory geoscience courses can improve student spatial ability further suggests that instructors should include spatial activities as a means of fostering student comprehension of the overall course content. Further research is needed to determine whether spatial ability plays a crucial role in student responses to conceptest questions.

Lastly, the findings from this study illustrate the importance of the class environment developed by the instructor. The finding that the variable Instructor is a significant covariate suggests that the two instructors varied in their implementation of conceptest questions. The determining factor is not known at this time but could be related to differences in the questions asked, lecture style, or the types of in-class activities. Differences in the implementation of peer instruction seems unlikely given that Dr. Anderson and Dr. Baker both used asked a similar number of questions in a similar manner. They also used the same textbook and personal response system. Student
demographics can also be ruled out because they do not systematically vary between the two instructors. Additional research on variations in how instructors adopt the peer instruction methodology might isolate one or more contributing factors. One possible study would be to observe several classes that use a personal response system. By observing the instructors, one could record how the instructors interact with the students, introduce the questions, as well as the time provided for student responses. Observations of students could uncover student behaviors and attitudes towards answering the questions. If low-achieving students require longer periods of time to respond (Lasry et al., 2009) then observing how the students respond to each question might reveal whether instructors unwittingly frustrate the poor-achieving students by not providing sufficient time to answer the questions.
The data from this study suggest that men and women respond differently to both text-only and illustrated concepttest questions. Differences in student spatial ability are one possible explanatory mechanism but further research is needed to establish this link. Furthermore, the observation that the text-only questions could contain a significant spatial component further explains why men scored higher than women on both types of questions. Additional research is needed to determine whether student spatial ability is linked to responses to concepttest questions.

This study also found that student prior achievement in reading, math, and science does not differentiate responses to either type of question. It is likely that both the ACT tests and student responses measure student academic achievement which supports the assertion that concepttest questions are a valid formative assessment tool. It is possible that the more generalized ACT-Composite test could be a reliable predictor of student responses on concepttest questions and serve to identify possible at-risk students before they fail the course. More research is needed to support this claim.

Peer instruction and concepttest questions have proven to be a popular method for introducing active-learning pedagogies into the undergraduate classroom. Yet the findings from this study must be tempered by the words of Clark (2001) and Mayer et al. (2009) that technology does not teach students and does not cause them to learn.
Learning occurs when students interact with new information and connect their new knowledge with what they have previously learned. Technologies such as personal response systems provide a tool for fostering student learning, but the emphasis should remain on improving student learning rather than trusting that concept test questions are a quick panacea for learning.
REFERENCES


189


*Dissertation Abstracts International*, 66 (10), 141.


APPENDIX A

INSTITUTIONAL REVIEW BOARD CONSENT LETTER

Title of Study: Collaborative Research: Evaluating student learning in geoscience curricula that employ concept tests using electronic student response systems

Dear Student,  1/6/2009

You are invited to participate in a research project being conducted by Dr. David Steer, a faculty member in the Department of Geology and Environmental Science at The University of Akron.

The purpose of this study is to determine the extent to which in-class conceptual questions answered using electronic student response systems aid in your learning of earth science concepts. Over 10,000 students are expected to participate in this study over the next four years.

During this course you will answer conceptual multiple choice questions as part of normal classroom activities. Your responses will be collected and compared with students at six other colleges across the country. You may also be observed by student observers, write responses to various questions and perhaps be interviewed. We may ask to photocopy and archive some work done for the class. Some class activities may be video taped with edited portions posted on the WWW as a learning tool for other faculty. Though you will be recognizable in those video clips, no identifying personal information will be posted on-line. All activities are part of the normal classroom active-learning activities that all students will complete as part of this course.

You will receive no direct benefit from your participation in this study, but your participation may help us better understand the efficacy of using electronic response systems to teach key concepts in earth science.

Your participation in this research is voluntary and you may refuse to participate, or may discontinue participation at any time, without penalty or loss of benefits to which you are otherwise entitled. Please note that all of your responses are appreciated and will remain confidential. Refusal to participate simply means that your responses will not be analyzed or included in the research database. You must still complete all classroom activities.

Confidentiality of records will be maintained by coding of data with random numbers. Your written responses to modeling questions, answers to multiple choice questions and observations will be archived for comparison to pre- and post-course assessments, exam question responses, course grades, prior course work and demographic data.
Any identifying information collected will be kept in a secure location and only the researchers will have access to the data. Participants will not be individually identified in any publication or presentation of the research results. Only aggregate data will be used. Your signed consent form will be kept separate from your data, and nobody will be able to link your responses to you. We may contact you in the future to follow up on your perceptions of learning with this methodology or to probe long-term retention of key concepts.

If you have any questions about this study, you may call Dr. David Steer at 330-972-2099. This project has been reviewed and approved by The University of Akron Institutional Review Board. If you have any questions about your rights as a research participant, you may call the IRB at (330) 972-7666 or 1-888-232-8790.

Sincerely,

David Steer
Associate Professor
Department of Geology and Environmental Science
The University of Akron

I have read the information provided above and all of my questions have been answered. I voluntarily agree to participate in this study. I will receive a copy of this consent form for my information.

☐ I consent to the use of videotape of myself in a public forum - such as on the web or in conference presentations. Please check the box if you agree.

----------------------------------  ----------------------------------
Participant Signature                Date

----------------------------------
Printed Name
## APPENDIX B

### STUDENT DEMOGRAPHIC DATA

<table>
<thead>
<tr>
<th>Gender</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>303</td>
<td>54.0</td>
</tr>
<tr>
<td>Women</td>
<td>258</td>
<td>46.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>478</td>
<td>85.2</td>
</tr>
<tr>
<td>African American</td>
<td>58</td>
<td>10.3</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
<td>4</td>
<td>0.7</td>
</tr>
<tr>
<td>Hispanic</td>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>Unknown/NR Alien</td>
<td>18</td>
<td>3.2</td>
</tr>
<tr>
<td>All Minority</td>
<td>65</td>
<td>11.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshman</td>
<td>416</td>
<td>74.2</td>
</tr>
<tr>
<td>Sophomore</td>
<td>120</td>
<td>21.4</td>
</tr>
<tr>
<td>Junior</td>
<td>17</td>
<td>3.0</td>
</tr>
<tr>
<td>Senior</td>
<td>8</td>
<td>1.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age (Years)</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>67</td>
<td>11.9</td>
</tr>
<tr>
<td>19</td>
<td>211</td>
<td>37.6</td>
</tr>
<tr>
<td>20</td>
<td>156</td>
<td>27.8</td>
</tr>
<tr>
<td>21</td>
<td>59</td>
<td>10.5</td>
</tr>
<tr>
<td>22</td>
<td>36</td>
<td>6.4</td>
</tr>
<tr>
<td>23</td>
<td>11</td>
<td>2.0</td>
</tr>
<tr>
<td>24</td>
<td>8</td>
<td>1.4</td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>1.2</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>33</td>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>35</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>37</td>
<td>1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class Time</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:50</td>
<td>155</td>
<td>27.6</td>
</tr>
<tr>
<td>11:00</td>
<td>215</td>
<td>38.3</td>
</tr>
<tr>
<td>12:05</td>
<td>141</td>
<td>25.1</td>
</tr>
<tr>
<td>1:10</td>
<td>50</td>
<td>8.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instructor</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anderson</td>
<td>110</td>
<td>19.6</td>
</tr>
<tr>
<td>Baker</td>
<td>451</td>
<td>80.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intended Degree</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>BS</td>
<td>368</td>
<td>65.6</td>
</tr>
<tr>
<td>BA</td>
<td>149</td>
<td>26.6</td>
</tr>
<tr>
<td>BFA (Fine Arts)</td>
<td>24</td>
<td>4.3</td>
</tr>
<tr>
<td>AA</td>
<td>11</td>
<td>2.0</td>
</tr>
<tr>
<td>BM (Music)</td>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>Professional</td>
<td>6</td>
<td>1.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class #</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Fall 06 Baker 850</td>
<td>51</td>
<td>9.1</td>
</tr>
<tr>
<td>2 Fall 06 Baker 1100</td>
<td>76</td>
<td>13.5</td>
</tr>
<tr>
<td>3 Spring 07 Baker 1205</td>
<td>84</td>
<td>15.0</td>
</tr>
<tr>
<td>4 Fall 07 Baker 850</td>
<td>55</td>
<td>9.8</td>
</tr>
<tr>
<td>5 Fall 07 Baker 1100</td>
<td>79</td>
<td>14.1</td>
</tr>
<tr>
<td>6 Fall 07 Anderson 110</td>
<td>50</td>
<td>8.9</td>
</tr>
<tr>
<td>7 Spring 08 Baker 850</td>
<td>49</td>
<td>8.7</td>
</tr>
<tr>
<td>8 Spring 08 Anderson 110</td>
<td>60</td>
<td>10.7</td>
</tr>
<tr>
<td>9 Spring 08 Baker 1205</td>
<td>57</td>
<td>10.2</td>
</tr>
</tbody>
</table>

*Note.* N = 561 students
APPENDIX C

TEXT-ONLY CONCEPTEST QUESTIONS USED IN THIS STUDY

Question T-1

Wegener’s Scientific Method Conceptest

What was Wegener’s hypothesis?
A. The coastlines of South America and Africa fit together like jigsaw puzzle pieces.
B. Earth’s surface can be divided up into a series of plates made up of lithosphere.
C. Earth’s landmasses used to form a single supercontinent.
D. Fossils of the same organisms are found on different continents.

Question T-2

Seafloor Age Conceptest

Which statement is TRUE about the relationship between age and topography of the ocean floor?
A. Deeper regions of the ocean floor are younger.
B. The Pacific Ocean is larger than the Atlantic Ocean because it contains older oceanic floor.
C. Oldest oceanic crust is only present near trenches.
D. Youngest oceanic crust is near the ridges.

Question T-3 (Removed to I-108, 108a, & 108b)
Question T-4 (See T-26)

**Earthquake Concepttest**

An earthquake occurred on the Erie Fault, 5 km beneath Ashtabula, Ohio. Damage from the earthquake was greatest in nearby Chardon. The farthest report of shaking was recorded in Akron. Where was the earthquake's epicenter?

A. The Erie Fault  
B. Ashtabula  
C. Chardon  
D. Akron

Question T-5

**Seismic Waves Concepttest**

The city of Guildhall experienced two earthquakes within a 1-month interval. The first earthquake was magnitude 2; the second earthquake was magnitude 7. The P-waves from the magnitude 2 earthquake traveled ______ than the P-waves from the magnitude 7 earthquake.

A. faster  
B. slower  
C. at the same speed

Question T-6

**Seismic Waves Concepttest**

A major earthquake occurred along the plate boundary between the Pacific and North American plates in southern Alaska. Scientists at a seismograph station in Anchorage, AK, determined the magnitude of the earthquake to be 7.3. What would be a reasonable estimate of the earthquake magnitude determined at a seismograph station in northern California?

A. 8.1; CA experiences more damage than Alaska  
B. 7.3; All stations should record the same value  
C. 6.1; it is farther away, it should be smaller  
D. None; it would not be felt that far away
Question T-7

Earthquake Measurement Concept

Two sites (L1, L2) record earthquake intensity and earthquake magnitude for the same earthquake. L1 is located close to the epicenter and L2 is farther away. Where is the intensity greatest and what happens to the magnitude calculated at the different sites?

A. Intensity is greatest at L1; calculated magnitude is the same at each site.
B. Intensity is greatest at L2; calculated magnitude is the same for each site.
C. Intensity is greatest at L1; calculated magnitude decreases from L1 to L2.
D. Intensity is greatest at L2; calculated magnitude decreases from L1 to L2.

Question T-7a

Earthquake Concept

Two sites (L1, L2) record earthquake magnitude and earthquake intensity data for the same earthquake. L1 is located close to the epicenter and L2 is farther away. Where is the intensity greatest, and what happens to the earthquake magnitude calculated at the different sites?

A. Intensity is greatest at L1; calculated magnitude is the same at each site.
B. Intensity is greatest at L2; calculated magnitude is the same at each site.
C. Intensity is greatest at L1; calculated magnitude decreases with distance from the focus.
D. Intensity is greatest at L2; calculated magnitude decreases with distance from the focus.

Question T-8

Viscosity Concept

All other factors being equal, which magma would most easily release gases?

A. High viscosity magma
B. Low viscosity magma
C. Neither, magma does not trap gases

Question T-9

Viscosity Concept

All other factors being equal, which magma would flow the fastest?

A. High viscosity magma
B. Low viscosity magma
C. Neither, magma does not have viscosity
Question T-10

**Viscosity Concepttest**

All other factors being equal, which magma would flow the slowest?

A. High viscosity magma  
B. Low viscosity magma  
C. Neither, magma does not have viscosity

---

Question T-11

**Viscosity Concepttest**

All other factors being equal, which magma is most likely to cause an explosive eruption?

A. High viscosity magma  
B. Low viscosity magma  
C. Neither, viscosity is not a factor

---

Question T-12

**Magma Composition Concepttest**

Which will produce more explosive eruptions?

A. Volcanoes in New Zealand that have andesite magma  
B. Volcanoes in East Africa that have basalt magma
Question T-13

Magma Viscosity Conceptest

What is the most important factor in controlling the viscosity of magma?
A. The amount of volcanic gases in the magma
B. The types of volcanic gases in the magma
C. The amount of silica in the magma
D. The types of silica in the magma

Question T-14

Viscosity Conceptest

Which type magma most likely has the lowest viscosity?
A. High silica content, high temperature
B. High silica content, low temperature
C. Low silica content, high temperature
D. Low silica content, low temperature

Question T-15

Volcanoes Conceptest

Which product of the volcanic eruption traveled farthest from Mt. St. Helens?
A. Pyroclastic flow  B. Lahar  C. Lava
Atoms Conceptest
Which are present in an atom's nucleus?
A. Electrons
B. Electrons and protons
C. Neutrons
D. Neutrons and protons

Question T-17
Atoms Conceptest
An atom that loses an electron will have a ________ charge?
A. positive
B. negative

Question T-18
Igneous Rocks Conceptest
What is the most likely cooling rate and composition of a dark-colored, large grained igneous rock?
A. Cooled rapidly, low silica
B. Cooled rapidly, high silica
C. Cooled slowly, low silica
D. Cooled slowly, high silica
Question T-19

Igneous Rocks Conceptest

Imagine that all minerals found in rocks were the same color. What information would you no longer be able to infer?

A. Texture
B. Cooling rate
C. Composition

Question T-20

Igneous Rocks Conceptest

What is the most likely cooling rate and composition of a light-colored, small grained igneous rock?

A. Cooled rapidly, low silica
B. Cooled rapidly, high silica
C. Cooled slowly, low silica
D. Cooled slowly, high silica

Question T-21

Igneous Rocks Conceptest

An igneous rock known as porphyry contains both large and small crystals. Which is the best explanation for the formation of this rock? The rock experienced a two stage cooling process with

A. Initial slow cooling at depth followed by rapid cooling at the surface
B. Initial rapid cooling at depth followed by slow cooling at the surface
C. Initial slow cooling near the surface followed by rapid cooling at depth
D. Initial rapid cooling near the surface followed by slow cooling at depth.
Question T-22

**Rock Cycle Conceptest**

Wood burned in a fire is converted to ash and cinders. This could be seen as an analog for the formation of:

A. igneous rock  
B. metamorphic rock  
C. sedimentary rock.

---

Question T-23 (Blue question)

**Concept – Magma Viscosity**

Imagine taking a straw and bubbling air through liquids in two glasses – milk and a milkshake.

Which liquid has the higher viscosity?  
A. Milk  B. Milkshake

Gas (air) would escape more readily from which liquid?  
Which “eruption” would be the most explosive (violent)?

---

Question T-24

**Continental Drift Conceptest**

Which of these lines of evidence did Wegener **NOT** use to support his continental drift hypothesis?

A. The distribution of fossils on different continents  
B. Fit of the edges of the continents  
C. Match of mountain belts between different continents  
D. Patterns of earthquake locations within the ocean basins  
E. Palaeoclimate data (especially data on glacial deposits).
Question T-25

Plate Tectonics Concept Test

Approximately how old is the oldest oceanic crust?

A. 15 million years
B. 450 million years
C. 180 million years
D. 4 billion years

Question T-26 (See T-4)

Earthquake Concept Test

An earthquake occurred on the Erke fault 5 kilometers beneath San Gabriel. Damage from the earthquake was greatest in nearby Fremont. The farthest report of shaking was recorded in Stockton. Where was the earthquake's epicenter?

A. The Erke Fault
B. San Gabriel
C. Fremont
D. Stockton

Question T-27

Earthquake Concept Test

The difference in arrival times of P and S waves at a seismograph station varies with ____________.

A. the size of the fault
B. the size of the earthquake
C. the distance from the earthquake source
D. all of the above
Question T-28

**Earthquake Conceptest**

The difference in arrival times of P and S waves _______ with distance from the earthquake source.

A. Increases  
B. Decreases  
C. Stays the same

Question T-29

**Earthquake Conceptest**

A large earthquake occurred along a fault and was recorded at a seismograph station 300 kilometers away. The next day, a smaller aftershock earthquake occurred at the exact same location on the fault. What statement is most accurate?

A. P-waves would have traveled to the seismograph station more quickly following the first earthquake  
B. P-waves would have traveled to the seismograph station more quickly following the second earthquake  
C. The P-waves would have taken the same time to reach the station after each earthquake

Question T-30

**Earthquake Conceptest**

How much would ground motion increase between magnitude 4.5 and 5.5 earthquakes?

A. About a quarter more  
B. About 5 times as much  
C. About 10 times as much  
D. About 32 times as much
Question T-31

**Earthquake Concept**

Most earthquakes occur at_______ depths.

A. Shallow (0-70 kilometers)
B. Intermediate (100-300 kilometers)
C. Deep (>300 kilometers)

Question T-32

**Earthquake Concept**

Which of the following determine the amount of damage that a city may suffer from an earthquake?

A. Earthquake magnitude
B. Distance from earthquake source
C. Earth materials underlying city
D. All of these

Question T-33

**Earthquake Concept**

Earthquake intensity measures

A. Energy released from earthquake
B. Amplitude of seismic waves on seismogram
C. Damage resulting from the earthquake
D. Displacement on faults
Question T-34

**Earthquake Concept Test**

If you lived in a city located 20 kilometers from the epicenter of a magnitude 8.7 earthquake, which of the following would you be most likely to experience?

A. ground shaking
B. landslides
C. fault rupture
D. tsunami
E. liquefaction

Question T-35

What heat transfer processes best describe the change from cold water to boiling water on a stove?

a. Mostly convection, then conduction
b. Mostly conduction, then convection
c. Just conduction
d. Just convection

Question T-36

Which statement best illustrates the reason the seafloor becomes older moving away from a mid-ocean ridge?

a. Older crust forms far away from the ridge.
b. Crust gets older as it migrates toward the ridge.
c. Crust gets older as it migrates from the ridge.
Question T-37
Which statement best illustrates the reason the seafloor gets deeper away from the ridge?

a. Deeper regions are warmer and less dense than near the ridge.
b. Deeper regions are warmer and more dense than near the ridge.
c. Deeper regions are colder and less dense than near the ridge.
d. Deeper regions are colder and more dense than near the ridge.

Question T-38
How much does ground motion change between a magnitude 5 and magnitude 3 earthquake?

a. 2 times smaller for the magnitude 3
b. 20 times smaller for the magnitude 3
c. 10 times smaller for the magnitude 3
d. 100 times smaller for the magnitude 3

Question T-39 (Removed to I-109)

Question T-40
Concept - Volcanic Eruptions
Read the passage below and identify how many scientific errors are present (compared to reading).

Prior to May 18, 1980, there were several indications of a possible eruption of Mount St. Helens. Earthquakes rattled the volcano more frequently than before. The earthquakes triggered a large landslide that formed a bulge on the north slope of the volcano. Lava flows spilled out of vents around the base of the mountain. Volcanic gases poisoned several people working in the nearby national forest. Luckily, all these signals resulted in a complete evacuation of the area.

A. 1 error        C. 3 errors
B. 2 errors        D. 4 errors
Question T-41

Volcanoes and Other Mountains Concepttest

How would the elevations of mountains differ if Earth’s crust was composed of less dense rocks? Mountains would be

A. Higher
B. Lower
C. Unchanged in elevation

Question T-42 (See question T-12)

Viscosity Concepttest

All other factors being equal, which magma is more likely to cause an explosive eruption?

A. Andesitic magma
B. Basaltic magma

Question T-43

Active Volcanoes Concepttest

Which is the best explanation for the source of most molten rock (magma) from volcanoes on the continents?

A. Magma comes from the molten rock in Earth’s outer core
B. Magma comes from a layer of molten rock located in the upper mantle
C. Magma comes from pockets of molten rock beneath Earth’s surface
Question T-44

Volcanoes Concepttest

List the following products of the eruption of Mount St. Helens in the order of distance they travel from the volcano (less distance → greater distance).

A. Pyroclastic flow, tephra, lava, lahar
B. Lahar, pyroclastic flow, tephra, lava
C. Tephra, lava, pyroclastic flow, lahar
D. Lava, pyroclastic flow, lahar, tephra

Question T-45

Volcanoes and Other Mountains Concepttest

What is the most likely explanation of why the Rocky Mountains are more than double the elevation of the Appalachian Mountains?

A. The Rocky Mountains are older and have been forming for a longer time and are therefore higher.
B. The Appalachians are older than the Rocky Mountains and have been lowered by erosion.
APPENDIX D

ILLUSTRATED CONCEPT TEST QUESTIONS USED IN THIS STUDY

Question I-1

Earth's Features Concept Test
Which is the actual map of Earth's features?

Question I-2

Earth's Structure Concept Test
Which of the images below best approximates the relative distribution of Earth's core, mantle, and crust?

a) Big core, thin crust
b) Small core, thick crust
c) Big core, thick crust
d) Small core, thin crust
Question I-3

**Continental Drift Concepttest**

Which piece of the broken china plate best fits with piece #1?

- A. Piece #3
- B. Piece #5
- C. Piece #6A
- D. Piece #6B
- E. Piece #6C

Question I-4

**Seafloor Topography Concepttest**

Which image best approximates the shape of the ocean floor in the Atlantic Ocean?

- A
- B
- C
- D

Question I-5

**Concept - Seafloor Topography**

Where in the world could you find this cross section? continent - oceans ridge - trench (Please on shape, not height)

Map of the world with labeled features and a question about finding the correct location.
Question I-6

Concept - Plate Tectonics
Plate Tectonic Cycle
Ocean ridges and subduction zones are boundaries between "plates" of lithosphere

How many plates are in this image?

Question I-7 and I-7a

Plate Tectonics Concepttest

How many plates are in this image?

A. 2  B. 3  C. 4  D. 5  E. 6

Question I-8

Plate Motions Concepttest

Review the map below and identify which pair of locations is moving closer together as a result of plate tectonics?

Question I-9

Plate Boundaries Concepttest

Which map best represents the locations of oceanic ridges?

A.  
B.  
C.  
D.  

Question I-10

Plate Tectonics Concepttest

Where would you find an example of a mature divergent plate boundary?

A.  
B.  
C.  
D.  

Question I-11

Concept - Convergent Plate Boundary

2. Oceanic lithosphere vs. Continental lithosphere
   - More dense oceanic lithosphere descends subduction zone
   - Volcanoes and mountains located on overriding plate
   - Earthquakes (x) located on descending plate
Question I-12

Plate Tectonics Concepttest

Which map best represents the locations of convergent plate boundaries?

A  
B  
C  
D

Question I-13

Plate Tectonics Concepttest

Which model best represents the plate configuration at an ocean-continent convergent boundary?

A  
B  
C

Question I-14

Fault Concepttest

Fault in an orchard, southern California

Hebgen Lake fault, MT

Approximately how much do faults move during a large earthquake?

A. 0.5 meters
B. 5 meters
C. 50 meters
D. 500 meters
Question I-15

The Earthquake Machine

What will happen when we turn the handle?
The tubing will stretch and . . .
A. the brick will move steadily forward
B. the brick will not move at all
C. the brick will initially not move, then jump forward

Question I-16

The Earthquake Machine

Predict what will happen if we use two bricks instead of one.
Will movement on the “fault” be more or less?
A. More
B. Less
C. The same

Question I-17

Fault Concepttest

If the San Andreas fault moves 500 cm per big earthquake, and fault movement is equivalent to plate motion (2.5 cm/yr): How many years of plate motions must accumulate to produce one big earthquake?
A. 2 years
B. 20 years
C. 200 years
D. 2000 years
Question I-18

Seismic Waves Concepttest

Earthquakes produce seismic waves that can be identified as body waves (P, S) and surface waves. Review the seismogram from the 2001 Ashkhabad earthquake. Which letter corresponds to the first S wave?

Question I-19

Earthquake Intensity Concepttest

Following a shallow earthquake, twenty residents located throughout the affected region were surveyed. Their observations were quantified using the Modified Mercalli scale. Which of the following figures best describes the relationship between earthquake intensity and distance from the earthquake epicenter?

Question I-20

Earthquake Distribution Concepttest

Which of the maps below is the best representation of the distribution of earthquakes of magnitude 7 or larger?

Question I-20a

Earthquake Distribution Concepttest

Which of the maps below is the best representation of the distribution of large earthquakes?
Question I-21

Earthquake Distribution Concepttest

The figures below show the location of a plate boundary (red line) and the distribution of earthquake epicenters (filled circles). The size of the filled circle indicates the earthquake magnitude.

Which figure best represents a convergent plate boundary between oceanic and continental plates?

- Figure A
- Figure B
- Figure C
- Figure D
- Figure E

Question I-22

Earthquake Distribution Concepttest

The figures below show the location of a plate boundary (dashed line) and the distribution of earthquake foci (filled circles). The color of the filled circle indicates the depth of the earthquake.

Which figure best illustrates a convergent plate boundary between oceanic and continental plates?

- Figure A
- Figure B
- Figure C
- Figure D
- Figure E
Question I-23

**Viscosity Demonstration**

Examine the contents of the two vials. Which fluid has the highest viscosity?

A. Green fluid
B. Clear fluid
C. No way to tell

---

Question I-24

**Viscosity Demonstration**

Examine the contents of the two vials. Which fluid appears to hinder gas escape the most?

A. Green fluid
B. Clear fluid
C. No way to tell

---

Question I-25

**Viscosity Demonstration**

What are the fluids?

A. Green = oil, clear = syrup
B. Green = syrup, clear = oil
C. No way to tell
Question I-26

Active Volcanoes Concepttest

Which of the maps below best represents the distribution of active volcanoes? (Why?)

Question I-27

Volcanoes and Magma Concepttest

Which image best illustrates the source of most magma from active volcanoes?

- A: Magma forms from a source in the mantle
- B: Magma forms from a source layer in the mantle
- C: Magma forms from an isolated source below volcanoes

Question I-28

Volcanoes and Plates Concepttest

Which type of volcanoes are present at convergent plate boundaries?

- A: Shield
- B: Stratovolcano
- C: Cinder Cone
Question I-29

Volcanoes Concepttest

Which type of volcano is this?
A. Shield  B. Stratovolcano  C. Cinder Cone

Question I-30

Igneous Rocks Concepttest

The lines shown on the idealized graph below depict time-temperature cooling histories for magma. What rock is best represented by cooling history A?
A. Granite  B. Basalt

Question I-31

Igneous Rocks Concepttest

What type of igneous rock is sample #1?
A. Gabbro  C. Granite
B. Basalt  D. Rhyolite
Question I-32 – Rock sample 3.

Igneous Rocks Concepttest

What type of igneous rock is sample #1?
A. Gabbro  C. Granite
B. Basalt  D. Rhyolite

Question I-33

Igneous Rocks Concepttest

The lines shown on the idealized graph below depict time-temperature cooling histories for magma. What rock is best represented by cooling history D?
A. Gabbro  B. Rhyolite

Question I-34 Which sample is the limestone? 7

Concept - Sedimentary Rocks

Identify the chemical (rock salt), biochemical (limestone), & clastic (shale, sandstone) sedimentary rocks in each sample bag

Sedimentary Rocks Concepttest

Which rock sample is limestone?
5 6 7 8
**Question I-35**

**Sedimentary Rocks Concept Test**

Weathering, transportation, and deposition can occur during steps:

1. 
2. 
3. 
4. 
5. 
6. 

**Question I-36**

**Sedimentary Rocks Concept Test**

The diagram below illustrates five potential combinations of temperature and depth that are characteristic of different rock types.

Which letter best represents the conditions necessary for the formation of a sedimentary rock?

1. 
2. 
3. 
4. 
5. 

**Question 36a**

**Review**

\[ \text{Which rock is most likely sedimentary?} \]

- A
- B
- C
- D
- E

**Question I-37**

**Concept - Metamorphic Rocks**

Two types of metamorphism:

1. **Contact metamorphism**
   - Changes due to increases in temperature only
   - Example: limestone around a magma chamber is baked by the heat of the magma to form marble

2. **Regional metamorphism**
   - Changes due to increases in temperature and pressure

- **Magma chamber**
- **Zone of contact metamorphism**
- **Country rock**
Question I-38

**Rock Cycle Concepttest**

The following diagram illustrates the Rock Cycle. Which term or phrase best represents Step g?

A. Comamination and compaction  
B. Heat and Pressure  
C. Weathering, Transportation, deposition  
D. Cooling and solidification  
E. Melting

---

Question I-39

**Rock Cycle Concepttest**

The graph below illustrates how the temperature changed with time for part of the rock cycle. Which of the following is best represented by the graph?

A. Sediment is lithified to form sedimentary rock  
B. Sedimentary rocks are metamorphosed to metamorphic rocks  
C. Metamorphic rocks are uplifted to Earth’s surface  
D. Magma cools to form plutonic igneous rock  
E. Sedimentary rock is converted to magma

---

Question I-39a

**Rock Cycle Concepttest**

The graph below illustrates how the temperature changed with time for part of the rock cycle. Which of the following is best represented by the graph?

A. Sediment is lithified to form sedimentary rock  
B. Sedimentary rocks are metamorphosed to metamorphic rocks  
C. Metamorphic rocks are uplifted to Earth’s surface  
D. Magma cools to form plutonic igneous rock  
E. Sedimentary rock is converted to magma

---

Question I-40

**Plate Motions Concepttest**

The continental crust at E is moving toward the

A. Southeast  
B. Southwest  
C. Northeast  
D. Northwest
Question I-41

Plate Motions Conceptest

The island of Bermuda is a former volcano on the floor of the western Atlantic Ocean. Approximately how far and in what direction would the island travel in 100 years?

A. 20 centimeters to the west
B. 20 centimeters to the east
C. 200 centimeters to the west
D. 200 centimeters to the east

Question I-42

Plate Tectonics Conceptest

Which cross section best represents the plate boundary configuration at location #1?

Question I-43

Plate Tectonics Conceptest

Which cross section best represents the plate boundary configuration at location #4?
Question I-44

Plate Tectonics Conceptest

Which one of these boundaries pictured in the diagrams below is least likely to have active volcanoes?

Question I-45

Plate Tectonics Conceptest

Which cross section best represents the plate boundary configuration at location #8?

Question I-46

Atoms Conceptest

The atomic model represents an element with 2 protons. Which atom represents an ion with a negative charge of 2?
Question I-47

**Sedimentary Rocks Concept Test**

Which rock sample is a clastic sedimentary rock?

A. 5  B. 6  C. 7  D. 8

Question I-48

**Rock Cycle Concept Test**

The following diagram illustrates the Rock Cycle. Which term or phrase best represents Step a?

- A. Cenomotonic and compaction
- B. Heat and Pressure
- C. Weathering, transportation, deposition
- D. Cooling and solidification
- E. Melting

Question I-49

**Plate Tectonics Concept Test**

**Graph Exercises**

The depth of the seafloor

A. Becomes shallower as you move away from the oceanic ridge system
B. Becomes deeper as you move away from the oceanic ridge system
C. Remains at the same depth as you move away from the oceanic ridge system
Question I-50

Plate Tectonics Concept Test

The graph of the age of the seafloor should look like this.

A. True
B. False

Question I-51

Plate Tectonics

Which set of arrows best shows the relative positions of crust, mantle, lithosphere, and asthenosphere?

A = Asthenosphere
C = Crust
L = Lithosphere
M = Mantle

Question I-52

Evidence from the Seafloor

Plate Tectonic Cycle

Ocean ridges and trenches (subduction zones) are boundaries between “plates” of lithosphere.

How many plates are in this image?

A. 1
B. 2
C. 3
D. 4
E. 5
Question I-53 (See I-8)

Plate Tectonics Concepttest

Review the map below and identify which part of the ocean is caused by subduction as a result of plate tectonics?

a. Togo and Taiwan  b. Galapagos Islands and Hawaii  c. Bombay and Sydney  d. Galapagos Islands and North America

![Map of plate tectonics]

Question I-54

Plate Tectonics Concepttest

Examine the diagram below and determine which statement is most accurate?

- A. a is a divergent boundary, b is a transform boundary, c is a transform boundary
- B. a & b are transform boundaries, c is a convergent boundary
- C. a is a convergent boundary, b is a transform boundary, c is a convergent boundary
- D. a is a transform boundary, b is a divergent boundary, c is a convergent boundary

![Diagram of plate tectonics]

Question I-55

Plate Tectonics Concepttest

Which of the locations on the map are examples of divergent plate boundaries?

A. 1, 5, 8  B. 3, 4, 5  C. 2, 7, 9  D. 1, 5, 6  E. 3, 7, 9

![Map of plate tectonics]
Question I-56

Plate Tectonics Concepttest

Which of the locations on the map all represent examples of convergent plate boundaries?

A. 1, 6, 8
B. 3, 4, 5
C. 2, 7, 9
D. 1, 5, 6
E. 3, 7, 9

Question I-57

Plate Tectonics Concepttest

Oceanic trenches are present at which pair of locations?

A. A, C
B. B, E
C. G, H
D. D, F

Question I-58

Plate Tectonics Concepttest

Where did seafloor spreading happen earlier?

A. Between Australia and Antarctica
B. Between South America and Africa
Question I-59

**Seafloor Spreading Model: Part 2**

5. What do the two strips of paper represent?
   A. Crust
   B. Lithosphere
   C. Asthenosphere

6. Draw arrows on the yellow sheet to illustrate the direction of movement of the plates.

7. How many plates are illustrated by this model (both paper strips and on the yellow sheet)?

---

Question I-60

**Plate Tectonics Conceptest**

Examine the diagram below and predict where the symbol ● would have been located approximately 20 million years earlier.

---

Question I-61

**Fault Conceptest**

This is the San Andreas fault, a strike-slip fault in California.

Which set of arrows correctly indicates the direction of fault slip?

A. B.
Question I-62

Earthquake Concept

How many of these are in the wrong location?

A. First arrival
B. Last arrival
C. Body wave
D. Rayleigh wave
E. 45° in crust
F. In Earth’s interior
G. On Earth’s surface
H. Produced by slip on fault

Question I-63 (See I-18)

Seismic Waves Concepttest

Earthquakes produce seismic waves that can be identified as body waves (P, S) and surface waves.

Review the seismogram from the 2001 Ashitaka earthquake. Which letter corresponds to surface waves?

Question I-64 (What is ample #4?)

Concept - Igneous Rocks

Which rock samples correspond to . . . ?

A. A low silica volcanic rock
B. A high silica volcanic rock
C. A low silica plutonic rock
D. A high silica plutonic rock
Question I-65

Which statement best summarizes the global heat flow map?

A. High in middle of continents, low in oceans
B. High in middle of oceans, low on continents
C. About the same on both

Question I-66

Which statement best summarizes the seafloor age map?

A. Old in middle, young on edges
B. Young in middle, old on edges
C. Young in middle and on edges
D. Old in middle and on edges

Question I-67

Which statement best summarizes the seafloor depth map?

A. Shallow in middle, deep toward edges
B. Deep in middle, shallow toward edges
C. About the same everywhere
Question I-68

Which statement best summarizes the volcano locations map?

A. All random
B. All in patterns
C. Some patterns, some random

Question I-69

The figures below show the location of a plate boundary (red line) and the distribution of earthquake epicenters (filled black circles). The size of the filled circle indicates the earthquake magnitude. Given the distribution and size of the earthquakes, which figure best represents a mid-ocean ridge system?

Question I-70

Plate Tectonics Concepttest

How many tectonic plates are present in this diagram?

a. 2  
b. 3  
c. 4  
d. 4  
e. 5  
f. 6
Question I-71

Plate Tectonics Concepttest

Where is the deepest seafloor?

A.  
B.  
C.  
D.  
E.  

Question I-72

Plate Tectonics Concepttest

Which letters are getting farther apart?

a. A, C  
b. B, D  
c. C, E  
d. D, E  

Question I-73

Plate Tectonics Concepttest

Which location is a volcanic island arc?

A.  
B.  
C.  
D.  
E.  
F.  

246
Question I-74

**Plate Tectonics Concept Test**

How many tectonic plates are present in this diagram?

- a. 2
- b. 3
- c. 4
- d. 4
- e. 5
- f. 6

---

Question I-75

**Plate Tectonics Concept Test**

Where is the shallowest seafloor?

- A.
- B.
- C.
- D.
- E.

---

Question I-76

**Plate Tectonics Concept Test**

Which pair of letters are getting closer together?

- a. A, C
- b. B, D
- c. C, E
- d. D, E
Question I-77

Plate Tectonics Concept Test

Which locations have deep earthquakes?

A. A, B  
B. A, C  
C. A, D  
D. A, E  
E. A, F  

I-78

More practice

How many plates are present?

A. 2  
B. 3  
C. 4  
D. 5  

I-79

More practice

Which location is most likely to experience deep earthquakes?

A.  
B.  
C.  
D.  
E.  
F.  

248
Question I-80

More practice

Which location is most likely to have active volcanoes?

A. B. C. D. E. F.

Question I-81

More practice

Which letters are moving toward one another?

A. A, C B. C, D C. D, E D. E, F

Question I-82

The schematic cross sections below illustrate four models of relative plate motions and plate boundary geometry. Which of the cross sections best represents the characteristics of the plate boundary the location shown on the map below?

[Map image]
Question I-83

The schematic cross sections below illustrate four models of relative plate motions and plate boundary geometry. Which of the cross sections best represents the characteristics of the plate boundary the location shown on the map below?

[Map image]

Question I-84

Which two locations are moving apart?

- a. Bombay and Sydney
- b. Hawaii and Tokyo
- c. New York and London
- d. Cape Town and Sydney

Question I-85

Magma Viscosity Concepttest

Which liquid has the higher viscosity?

A. Water
B. Shampoo
Question I-86

Magma Viscosity Concepttest

Which “eruption” would be the most explosive (as measured by the size of the gas bubbles)?

A. Water  
B. Shampoo

Question I-87

Volcanoes and Other Mountains Concepttest

How would the values h and d change if we used a larger block of the same wood?

A. h would increase, d would decrease  
B. h would decrease, d increase  
C. Both h and d would increase  
D. Both h and d would decrease

Question I-88

What happens if water is replaced with corn syrup (density ~ of 1.4 g/cm³)?

A. More wood above the equilibrium line  
B. Less wood above the equilibrium line  
C. Amount above the line stays the same
Question I-89

Crustal thickness map

Where are the highest mountains?

A. 
B. 
C. 
D.

Question I-90

What happens to mountain elevations when steep valleys erode in mountains?

A. Mountains rise
B. Mountains fall
C. Mountains stay same

Question I-91

Classification of Igneous rocks

Texture $\rightarrow$ size of crystals most important

Cooling rates
- Surface fast
- Small crystals
- Below surface slow
- Larger crystals

Which formed at surface?
Both are the same chemistry.

A. Rhyolite
B. Granite
Question I-92 (See I-30)

The lines shown on the idealized graph above depict time-temperature cooling histories for magma. What rock is best represented by cooling history A?

a. granite  
b. gabbro  
c. basalt

Question I-93

Classification of Igneous rocks

Color
Indicative of chemistry and temperatures of formation
Dark colors  
High temperatures (1000 to 1200 °C)
Low silica content
Light colors  
Lower temperatures  
High silica content

Which is the high silica rock?
Both formed below the surface

A: Gabbro
B: Granite

Question I-94

Identify the rock type associated with line 1 the diagram below.

a. andesite porphyry  
b. gabbro  
c. basalt  
d. obsidian
Question I-95

Identify the rock type associated with line 4 the diagram below.

- a. andesite porphory
- b. gabbro
- c. basalt
- d. obsidian

Question I-96

Concrete is formed by adding cement and water to a mixture of sand and gravel. How would you classify cement if it were a rock?

- a. clastic
- b. chemical
- c. biochemical

Question I-97

Select the correct letter

A sedimentary rock.

Detrital

Chemical

Biochemical

A: B: C: D: E: F: G:
Question I-98

Two of these rocks are igneous, 2 are metamorphic and one is sedimentary. Which rocks are most likely metamorphic?

- a. A & B
- b. B & C
- c. C & D
- d. D & E

Question I-99 (See I-48)

The diagram summarizes the rock cycle. Which process is step 5?

- a. Cementation and compaction ( lithification )
- b. Heat and Pressure
- c. Weathering, transportation, deposition
- d. Cooling and solidification
- e. Melting

Question I-100

Question I-101
Question I-102 (See I-8)

Review the map of the plates above and identify which pair of locations is moving closer together as a result of plate tectonics.

- Galapagos Islands and Hawaii
- Galapagos Islands and Rio de Janeiro
- New York and London
- New York and Mexico City

Question I-103

Plate Tectonics Concepttest

Examine the map below and answer the question that follows. Where is a mountain range similar to the Andes Mountains (South America) most likely to be present?

- continental crust
- oceanic crust
- oceanic ridge
- subduction zone

A. A
B. B
C. C
D. D
E. E

Question I-104

A and B are located on two plates separated by a transform boundary (see diagram below). What direction is plate B moving if plate A is moving northeast (NE)?

- a. northeast
- b. northwest
- c. southwest
- d. southeast
Question I-105

Which letter corresponds to the highest viscosity material?

Question I-106

Which letter corresponds to basalt?

Question I-107

Volcanoes and Other Mountains Concepttest

Examine the map below. Where are the highest mountains located?

A. 1, 5
B. 2, 7
C. 3, 8
D. 4, 6
E. 9, 10
Question I-108

**The Earthquake Machine**

**Match these statements**

1. Slip on a fault is represented by...
   - A. ... the stretching of the tubing
2. The steady build up of stress due to interactions of the plates of lithosphere is similar to...
   - B. ... the movement of the brick
3. The deformation of the rocks in the crust prior to the earthquake is similar to...
   - C. ... the turning of the handle

Question I-109

Which statement best summarizes the earthquake locations map you just analyzed?

A. All random
B. All in patterns
C. Some patterns, some random

Question I-110

**The Earthquake Machine**

What will happen if we place the brick on a surface covered in talcum powder?

The tubing will stretch and...

- A. the brick will move steadily forward
- B. the brick will initially not move, then jump forward
- C. the brick will move forward in a series of smaller jumps
Question I-111

Volcanoes and Other Mountains Concept Test

Place the following 4 materials — maple syrup, milk, peanut butter, frozen yogurt — in the correct positions (A, B, C, D) for their relative viscosity.

![Viscosity Increases](image)

Question I-112

Magma Sources and Magma Composition

Differentiate settings yield different magmas from different source rocks.

Which volcano is more likely to have an explosive eruption?
A. Mount St. Helens, WA
B. Kilauea, HI

Question I-113

Concept — Volcanoes

On the basis of what you know about volcanic eruptions, which city will face the greatest potential hazard if Mount Shasta erupts? Why?
A. Weed
B. Mount Shasta City
C. McCloud
D. Dunsmuir
APPENDIX E

BIVARIATE SCATTERPLOTS TO TEST MULTIVARIATE NORMALITY

<table>
<thead>
<tr>
<th>Instructor</th>
<th>Gender</th>
<th>ACT-Reading (Quartiles)</th>
<th>ACT-Math (Quartiles)</th>
<th>ACT-Science (Quartiles)</th>
<th>Text Questions - Percent Correct</th>
<th>Illustrated Questions - Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

260
APPENDIX F

Q-Q PLOTS TO TEST MULTIVARIATE NORMALITY

Normal Q-Q Plot of Text Questions - Percent Correct