Abstract

PETERSON, MATTHEW OWEN. Comprehension with Instructional Media for Middle School Science: Holistic Performative Design Strategy and Cognitive Load. (Under the direction of Meredith Davis.)

This study identifies three distinct levels of text-image integration in page design in a linear relationship of lesser to greater integration: prose primary, prose subsumed, and fully integrated strategies. Science textbook pages were redesigned according to these holistic design strategies for 158 7th-grade students. There were three separate treatment tests, as well as a pre-test and post-test, and pilot tests with both undergraduate students and the subjects themselves.

Subjects found the fully integrated strategy to produce the most visually interesting designs and the prose primary strategy to produce the least interesting, with prose subsumed definitively in between (according to 95% confidence intervals).

The strategy employed significantly altered interest in science subject matter in one of three treatments (ANOVA, \( P=0.0446 \)), where a Student’s \( t \)-test revealed that the prose subsumed strategy produced higher interest in subject matter than prose primary.

The strategy employed significantly altered comprehension of abstract relationships in one of three treatments (ANOVA, \( P=0.0202 \)), where a Student’s \( t \)-test revealed that the fully integrated strategy resulted in greater comprehension than prose primary. For the same treatment condition significant differences were found through ANOVA for factual-level knowledge (\( P=0.0289 \)) but not conceptual-level knowledge (\( P=0.0586 \)). For factual-level knowledge prose primary resulted in lesser comprehension than both prose subsumed and fully integrated. Comprehension is defined according to cognitive load theory.

No strategy impact on perception of task difficulty was found.

This study was approved by North Carolina State University’s Institutional Review Board and Wake County Public School System’s Research Review Committee.
Comprehension with Instructional Media for Middle School Science: Holistic Performative Design Strategy and Cognitive Load

by
Matthew Owen Peterson

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Design

Raleigh, North Carolina

2011

APPROVED BY:

__________________________  __________________________
Nilda Cosco               Meredith Davis
Chair of Advisory Committee

__________________________  __________________________
James Minogue              John Nietfeld
Biography

Matthew Peterson is Assistant Professor of Graphic Design at the University of Illinois. He holds Bachelor of Graphic Design and Master of Graphic Design degrees from North Carolina State University. Between his undergraduate and graduate degrees he worked at the Walker Art Center in Minneapolis as a design fellow and was principal of a design studio in Chicago called Field Study. He now serves with KT Meaney as principal of Terms & Conditions in a professional capacity. Before all of this Matthew was born in North Carolina.
Acknowledgments

This would not have been possible without the help of a number of mammals. Each knows his or her particular and likely sizable contribution.

- Meredith Davis, thank you
- Nilda Cosco, thank you
- James Minogue, thank you
- John Nietfeld, thank you
- Ada Lopez, thank you
- Vicky Staubach, thank you
- Elwood Peters, thank you
- Rachael Huston, thank you
- Amber Howard, thank you
- Britt Cobb, Stephanie Jansen, Kayce Lomas, Carrie Misenheimer, and David Mitchell, thank you all
- Bob, Lea, Bobby, Heather, and Colin Peterson, thank you all
- Tim and Hailey Baldwin, thank you both
- KT Meaney, thank you
- Mai Goodness, thank you
# Table of Contents

List of Tables ............................................................................................................. xi

List of Figures ............................................................................................................ xv

Section 1 — Conceptual Framework ...................................................................... 1

1.1 Problem and Significance ................................................................................. 1

1.2 Reader Experience ............................................................................................. 5

1.3 The Designer’s Role in Reader Experience ...................................................... 6

1.4 Designer Decision Making ............................................................................... 8

1.5 Reader Construction from Page Structure ....................................................... 9

1.6 Comprehension and Learning ......................................................................... 10

1.7 Argument in Summary .................................................................................... 10

1.8 Study Focus and Research Questions ............................................................. 11

Section 2 — Literature Review ............................................................................. 15

2.1 Constitutive Metaphor .................................................................................... 15

2.1.1 Beyond Figurative Metaphor ....................................................................... 15

2.1.2 Schema Theory .......................................................................................... 17

2.1.3 Image Schemas .......................................................................................... 19

2.1.4 Discussion .................................................................................................. 20

2.2 Creative Memory ............................................................................................ 20

2.2.1 Memory as Constructed ............................................................................. 20

2.2.2 Mnemonic Manipulation of Memory .......................................................... 21

2.2.3 Memory Modeled in the Book .................................................................. 22

2.2.4 Working Memory ...................................................................................... 24

2.2.5 Discussion .................................................................................................. 27

2.3 Cognitive Load Theory ................................................................................... 27

2.3.1 The Cognitive Load Construct .................................................................. 27

2.3.2 Learning and Understanding ...................................................................... 29

2.3.3 Expertise .................................................................................................... 31
2.7.2 A Design Performance Assessment Battery ........................................... 74
2.7.3 Discussion ................................................................................................. 77

Section 3 — Methodology .................................................................................. 78

3.1 Approval ......................................................................................................... 78
3.2 Research Design Logic ................................................................................... 79
    3.2.1 General Purpose ...................................................................................... 79
    3.2.2 Research Questions and Research Design Distinctions ......................... 79

3.3 Research Design ............................................................................................ 83
    3.3.1 Forms for Testing .................................................................................... 83
    3.3.2 Sequencing .............................................................................................. 83
    3.3.3 Sessions ................................................................................................. 84
    3.3.4 Dependent Variables .............................................................................. 85

3.4 Sampling ......................................................................................................... 86
    3.4.1 School Selection and Description ........................................................... 86
    3.4.2 Group Selection ...................................................................................... 87
    3.4.3 Group Assignment .................................................................................. 87

3.5 Form Development ......................................................................................... 88
    3.5.1 Equivalency of Forms ............................................................................ 88

3.6 Data Collection Stages .................................................................................. 91
    3.6.1 Plan of Sessions ...................................................................................... 91
    3.6.2 Instrumentation ...................................................................................... 92
    3.6.3 Pre-Existing Data ................................................................................... 92
    3.6.4 Pilot Tests with Undergraduates ............................................................ 93
    3.6.5 Pilot Test with the Subjects ................................................................... 93
    3.6.6 Pre-Test .................................................................................................. 94
    3.6.7 Treatment Tests ...................................................................................... 94
    3.6.8 Post-Test ................................................................................................ 95

3.7 Data Collection Components ....................................................................... 96
    3.7.1 Performance Item Development .............................................................. 96
4.3.4 Sequence Group Differences in Science Self-Efficacy ....................... 148

Section 5 — Results......................................................................................... 149

5.1 Reliability................................................................................................. 149

5.2 Inquiry Into the Secondary Variables ....................................................... 150

5.2.1 Secondary Performance Test Batteries .................................................. 150

5.2.2 Science Self-Efficacy Inventory ............................................................. 152

5.2.3 Session Duration ................................................................................. 154

5.3 Text–Image Integration Strategy’s Impact on Comprehension .................. 156

5.3.1 Reliability Estimates for Treatment Tests ............................................. 156

5.3.2 Review of Comprehension .................................................................... 162

5.3.3 Mean Differences in Comprehension ..................................................... 163

5.3.4 Hypothesis Tests for Comprehension .................................................... 164

5.3.5 Duration as a Factor in Comprehension ................................................ 165

5.3.6 Comprehension: Conclusion ................................................................. 165

5.4 Text–Image Integration Strategy’s Impact on Task Difficulty ................... 166

5.4.1 Reliability Estimates for Task Difficulty Scales .................................... 166

5.4.2 Review of Task Difficulty ...................................................................... 168

5.4.3 Mean Differences in Task Difficulty ...................................................... 168

5.4.4 Hypothesis Tests for Task Difficulty ..................................................... 169

5.4.5 Performance as a Factor in Task Difficulty ........................................... 169

5.4.6 Task Difficulty: Conclusion ................................................................. 170

5.5 Text–Image Integration Strategy’s Impact on Interest Level ..................... 170

5.5.1 Reliability Estimates for Interest Level Inventories ............................... 170

5.5.2 Review of Interest in Subject Matter .................................................... 172

5.5.3 Mean Differences in Indirect Interest Level ......................................... 172

5.5.4 Hypothesis Tests for Indirect Interest Level ......................................... 174

5.5.5 Indirect Interest Level: Conclusion ...................................................... 174

5.6 Interest in Text–Image Integration Strategy ............................................. 174

5.6.1 Reliability Estimate for Direct Interest Inventory ................................. 174
8.4.4 Treatment #3 Test Instrument ................................................................. 235
8.4.5 Post-Test Instrument .............................................................................. 237
8.5 Item Analysis Summary ........................................................................ 239
8.6 Item Reliability ......................................................................................... 240
  8.6.1 Correlation Matrices ........................................................................... 240
  8.6.2 Basic Statistics .................................................................................. 244
  8.6.1 Errata ................................................................................................ 247
  8.6.2 Reliability Summary ........................................................................... 248
8.7 Analysis of Variance (ANOVA) Tables .................................................. 250
  8.7.1 The Sex Variable (Where Significant) ............................................... 250
  8.7.2 The Class Variable (Where Significant) .............................................. 251
  8.7.3 Group Equivalence ........................................................................... 252
  8.7.1 Independent Variable Inquiry .......................................................... 253
8.8 Student’s T ............................................................................................... 257
  8.8.1 Group Equivalence ........................................................................... 257
  8.8.2 Independent Variable Inquiry .......................................................... 259
8.9 Linear Regression ...................................................................................... 260
  8.9.1 Assessing Task Difficulty According to Performance ...................... 260
8.10 Task Difficulty Explanations .................................................................. 262
List of Tables

Table 2.4.1: DCT codes and modes (copied from Sadoski & Paivio, 2006, p. 45) .................. 39
Table 2.5.1: Picture efficacy (Levin & Mayer, 1993, pp. 98–103) ................................. 54
Table 2.6.1: Bloom’s cognitive domain (Bloom et al., 1956) ........................................ 64
Table 2.6.2: Bloom’s Revised alterations to Bloom’s Taxonomy ..................................... 66
Table 2.6.3: Bloom’s Revised vs. The New Taxonomy ..................................................... 72
Table 3.4.1: Racial breakdown (WCPSS, 2010a & 2010b; USDOE, 2009) ......................... 87
Table 3.5.1: Word count by form ....................................................................................... 89
Table 3.9.1: Difficulty constraints on discrimination with 33% method ......................... 109
Table 3.9.2: Review test difficulty (p) and discrimination (D) ........................................ 113
Table 3.9.3: Review test item-total correlation ............................................................... 114
Table 3.9.4: Review test distracter analysis ..................................................................... 115
Table 3.9.5: Treatment #1 difficulty ................................................................................ 116
Table 3.9.6: Treatment #1 discrimination ...................................................................... 121
Table 3.9.7: Treatment #1 item-total correlation ............................................................. 122
Table 3.9.8: Factual #1.5 distracter analysis ................................................................... 123
Table 3.9.9: Conceptual #1.1 distracter analysis ............................................................. 123
Table 3.9.10: Conceptual #1.3 distracter analysis ........................................................... 123
Table 3.9.11: Treatment #2 difficulty .............................................................................. 124
Table 3.9.12: Treatment #2 discrimination .................................................................... 126
Table 3.9.13: Treatment #2 item-total correlation .......................................................... 127
Table 3.9.14: Conceptual #2.3 distracter analysis ............................................................ 128
Table 3.9.15: Treatment #3 difficulty .............................................................................. 129
Table 3.9.16: Treatment #3 discrimination .................................................................... 130
Table 3.9.17: Treatment #3 item-total correlation .......................................................... 132
Table 3.9.18: Conceptual #3.5 distracter analysis ........................................................... 133
Table 3.9.19: Difficulty indices compared ....................................................................... 135
Table 3.9.20: Item interrogation summary ...................................................................... 136
Table 8.5.1: Factual vs. conceptual set difficulty ................................................................. 239
Table 8.5.2: Item interrogation exclusions ............................................................................. 239
Table 8.5.3: Final item set counts ......................................................................................... 240
Table 8.6.1: Science self-efficacy correlations (all significant at \( P \leq 0.05 \)) .................. 240
Table 8.6.2: Review test correlations (bold is significant at \( P \leq 0.05 \)) ......................... 241
Table 8.6.3: Treatment #1 correlations \( (P \leq 0.05) \) ......................................................... 241
Table 8.6.4: Treatment #2 correlations \( (P \leq 0.05) \) ......................................................... 242
Table 8.6.5: Treatment #3 correlations \( (P \leq 0.05) \) ......................................................... 242
Table 8.6.6: Indirect interest correlations (all significant at \( P \leq 0.05 \)) ......................... 243
Table 8.6.7: Direct interest correlations (all significant at \( P \leq 0.05 \)) ......................... 243
Table 8.6.8: Pilot test correlations \( (P \leq 0.05) \) ................................................................. 244
Table 8.6.9: Science self-efficacy statistics .......................................................................... 244
Table 8.6.10: Indirect interest level statistics .................................................................... 245
Table 8.6.11: Direct interest level statistics ...................................................................... 246
Table 8.6.12: Difficulty assessment statistics ................................................................... 246
Table 8.6.13: Session duration statistics .......................................................................... 247
Table 8.6.14: Aggregate score correlations ..................................................................... 247
Table 8.6.15: Summary of reliability estimates ................................................................. 249
Table 8.7.1: ANOVA table, review test by sex ................................................................. 250
Table 8.7.2: ANOVA table, combined treatment tests by sex ........................................... 250
Table 8.7.3: ANOVA table, science self-efficacy by sex .................................................... 251
Table 8.7.4: ANOVA table, combined duration by sex ...................................................... 251
Table 8.7.5: ANOVA table, combined treatment tests by class ....................................... 251
Table 8.7.6: ANOVA table, combined duration by class .................................................. 252
Table 8.7.7: ANOVA table, review test by group .............................................................. 252
Table 8.7.8: ANOVA table, combined treatment tests by group ....................................... 252
Table 8.7.9: ANOVA table, science self-efficacy by group ................................................ 253
Table 8.7.10: ANOVA table, treatment #1 test by form .................................................... 253
Table 8.7.11: ANOVA table, treatment #2 test by form ..................................................... 253
Table 8.7.12: ANOVA table, factual #2 set by form ................................................................. 254
Table 8.7.13: ANOVA table, conceptual #2 set by form ......................................................... 254
Table 8.7.14: ANOVA table, treatment #3 test by form ......................................................... 254
Table 8.7.15: ANOVA table, difficulty #1 rating by form ....................................................... 255
Table 8.7.16: ANOVA table, difficulty #2 rating by form ....................................................... 255
Table 8.7.17: ANOVA table, difficulty #3 rating by form ....................................................... 255
Table 8.7.18: ANOVA table, indirect interest #1 rating by form ............................................. 256
Table 8.7.19: ANOVA table, indirect interest #2 rating by form ............................................. 256
Table 8.7.20: ANOVA table, indirect interest #3 rating by form ............................................. 256
Table 8.8.1: T-test, combined treatment tests by class ......................................................... 257
Table 8.8.2: T-test, session duration by class ........................................................................... 258
Table 8.8.3: T-test, treatment #2 test by form .......................................................................... 259
Table 8.8.4: T-test, factual #2 set by form .............................................................................. 259
Table 8.8.5: T-test, conceptual #2 set by form ........................................................................ 259
Table 8.8.6: T-test, indirect interest #3 by form ....................................................................... 260
Table 8.9.1: Difficulty performance-basis summary of fit ...................................................... 261
Table 8.9.2: Difficulty performance-basis ANOVA ................................................................. 261
Table 8.9.3: Difficulty performance-basis parameter estimate ............................................... 261
List of Figures

<table>
<thead>
<tr>
<th>Figure Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1.1</td>
<td>Textbook picture and organization (Bernstein et al., 1989, p. 7)</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2</td>
<td>Textbook complexity (Glencoe, 2005, pp. 174–177)</td>
<td>3</td>
</tr>
<tr>
<td>1.1.3</td>
<td>Illustration–caption system (Hill, 1971, p. 147)</td>
<td>4</td>
</tr>
<tr>
<td>2.4.1</td>
<td>DCT model (copied from Sasoski &amp; Paivio, 2006, p. 53)</td>
<td>42</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Visual–verbal integration (copied from Mayer et al., 1995, p. 32)</td>
<td>47</td>
</tr>
<tr>
<td>2.7.1</td>
<td>Design process (redrawn from Kimbell et al., 1991, p. 20)</td>
<td>75</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Three strategies (modified from McDougal Littell, 2005a, C18–C19)</td>
<td>82</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Order-based groups plan</td>
<td>84</td>
</tr>
<tr>
<td>3.6.1</td>
<td>Sessions plan</td>
<td>91</td>
</tr>
<tr>
<td>3.7.1</td>
<td>Difficulty assessment item</td>
<td>100</td>
</tr>
<tr>
<td>3.7.2</td>
<td>Difficulty explanation item</td>
<td>100</td>
</tr>
<tr>
<td>3.7.3</td>
<td>Indirect interest level item</td>
<td>102</td>
</tr>
<tr>
<td>3.7.4</td>
<td>Direct interest level item (randomized PP, FI, PS)</td>
<td>103</td>
</tr>
<tr>
<td>3.7.5</td>
<td>Review test item</td>
<td>104</td>
</tr>
<tr>
<td>3.7.6</td>
<td>Science self-efficacy item</td>
<td>105</td>
</tr>
<tr>
<td>3.9.1</td>
<td>Item analysis protocols</td>
<td>107</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Review test scores</td>
<td>151</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Science self-efficacy scores</td>
<td>153</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Session #1 duration record</td>
<td>155</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Session #2 duration record</td>
<td>155</td>
</tr>
<tr>
<td>5.2.5</td>
<td>Session #3 duration record</td>
<td>155</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Treatment #1 scores</td>
<td>158</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Treatment #2 scores</td>
<td>158</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Treatment #3 scores</td>
<td>159</td>
</tr>
<tr>
<td>5.3.4</td>
<td>Treatment test means by form visualized</td>
<td>164</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Difficulty scores</td>
<td>166</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Indirect interest scores</td>
<td>171</td>
</tr>
</tbody>
</table>
1.1 Problem and Significance

Textbooks are not merely ancillary materials in US education. Instruction often *relies upon, depends upon*, and *follows* textbooks (Woodward, 1993a, p. viii). Because of the textbook’s far reach and its role in molding so much instruction, its impact is considerable. Unfortunately, the production demands of textbook publishing can overpower instructional goals, too often rendering the product insufficient to the task of engaging young people and facilitating learning (Woodward, 1993b; Pettersson, 1998).

Only certain factors are considered when textbooks are designed. Textbook publishers are meticulous with providing content to fit state standards (Glencoe, 2005, especially pp. NC8–NC14). But design decisions are generally applied after content is determined, with too little concern for the impact of visual structures of text and pictures. Textbooks are built from dominant, continuous prose, however heavily illustrated. A survey of textbooks found that most pictures were *content-supporting*, but almost none were *content-extending* (Woodward, 1993b), meaning that pictures were only relevant to subject
matter and didn’t contribute to the content in any meaningful way. This is perhaps a natural extension of a prose-dominant design strategy. The page from Bernstein et al. (1989, p. 7) reproduced in Figure 1.1.1 is from a section on units of measurement. The lone photographic illustration, besides reiterating an assertion from the text that there are different units of measurement (miles and kilometers), does nothing for the reader. It takes up space. Levin (1979) provides a typology of picture functions, identifying remunerative, decorative, and motivational functions as those that add nothing of value to the reader. The remunerative function stresses that pictures are often chosen simply because they might help sell books, which fails to translate into any cognitive gains (Pettersson, 1998, p. 7). Pictures don’t even appear to motivate readers inherently (Levin, 1979). Textbook designs focus on content inclusion (relevant but non-functional pictures, for instance), while ignoring cognitive outcomes.

Figure 1.1.1: Textbook picture and organization (Bernstein et al., 1989, p. 7)

A textbook’s reader must navigate the elements of its pages, which is no small feat given the chaos often present. Figure 1.1.2 illustrates the density of textbook layouts, in what
is not a particularly extreme example. Reader experience must be taken into account in order to assess a textbook’s value.

Figure 1.1.2: Textbook complexity (Glencoe, 2005, pp. 174–177)

Figure 1.1.1 provides a comical example of modeled reader experience (Bernstein et al., 1989, p. 7). This textbook religiously ends each passage (1–2 short paragraphs) with a review question that effectively asks the reader to repeat a phrase verbatim from the preceding text, sometimes in the sentence immediately preceding the question. The important sentence is then further repeated verbatim (its third appearance) under “What You Learned.” Any reasonable evaluation of this contrived system must be negative. At best the reader is taught to temporarily memorize a series of functionally meaningless phonemes; at worst the exercise is recognized by the reader and subsequently ignored. Reader experience must not only be modeled; it must produce desirable cognitive outcomes.

Figure 1.1.3 is an illustration–caption system from a mid-19th-century book on etiquette (Hill, 1971, p. 147). It provides a space in which a reader might inquire about bad manners at the dinner table. The main body of text is a numbered list that doesn’t promote a prose-style reading strategy. Instead it serves as a kind of database for the reader, who actively seeks examples of bad manners in the illustration and uses the numbered connections to check guesses or retrieve unknown information. “Bad Manners at the Table,”
as a text-and-picture integrated schematic, essentially models game play in its reader, a
cognitive outcome that should be desirable to those hoping to engage young people.

Figure 1.1.3: Illustration–caption system (Hill, 1971, p. 147)
Louise Rosenblatt (1978) provides a transactional theory of reading that begins to address the issue of reader experience. She assigns two core functions to the text, where the text is “a set or series of signs interpretable as linguistic symbols” (p. 12). First, the text activates “elements of the reader’s past experience,” functioning as a stimulus pattern (p. 11). Second, it “regulates what shall be in the forefront of the reader’s attention” (ibid.). Rosenblatt goes a step further in isolating the poem as distinctive from the text. The poem is any “literary work of art.” It is not a thing, but rather “an event in time” (p. 12). Without a reader, there is no poem. The relationship of text to poem, then, is that the poem is reader experience as guided by the text (ibid.).

Stanley Fish (1976), a literary critic, illustrates how a transactional model of reading changes the focus from information containment alone (as in the selection of content for textbooks) to reader experience. Fish reviews commentary on Milton’s Variorum sonnets by noting how his fellow critics seek final, correct answers to the meaning in the poems. The sonnets in question are typified by ambiguous meanings and references. For example, Milton’s twentieth sonnet concludes with this couplet:

He who of those delights can judge, and spare
To interpose them oft, is not unwise.

The sonnet is rumination on virtue and pleasure, and hinges on the meaning of “spare” in the penultimate line. It might mean either to “leave time for” or “refrain from” pleasures described earlier—with moral implications—both of which, dichotomous as the pair is, enjoyed healthy support from critics trying to “solve” the problem of inherent meaning (see Fish’s more extensive analysis, pp. 466–467). Fish points out that the conflicting solutions are equally valid, and turns his attention away from determining a winner in order to question the critical endeavor itself. Fish sees the interpretation of one of Milton’s sonnets as the most fruitful area of interest, and posits that Milton might have built
this ambiguity into the experience of reading the sonnet. Under a transmission model of communication (Emmert & Donaghy, 1981)—where information is unproblematically acquired from a source—it must be assumed that Milton provides an unambiguous message that is failing to reach all readers because of some kind of interference. But Fish removes the problem by redefining the exercise: “it is the structure of the reader’s experience rather than any structures available on the page that should be the object of description” (p. 468).

Fish changes the question from being what a poem means to how it goes about making meaning, or, as John Ciardi (1959) puts it: how a poem means. “What [a] poem is, is inseparable from its own performance of itself… Above all else, poetry is performance” (Ciardi, 1959, p. 668). J. L. Austin (1975), a linguist, coined the term performative to distinguish between language that merely describes reality (“constative utterances”) and language that is “aimed at getting something done” (Eagleton, 1983, p. 118). With this ring I thee wed is an example of a performative utterance because by virtue of being said, at least in a socially agreed-upon situation, something occurs. Ciardi and Fish view poems as performative. Educational media at the instructional level—be it print or screen-based—can also be understood in performative terms. To consider instructional print media as performative is to shine light on a reader’s cognitive processing through what Fish describes as the structure of reader experience.

1.3 The Designer’s Role in Reader Experience

Fish proposes that Milton was essentially creating an experience (as opposed to a poem-object). To do so, Milton must first have recognized poems as experiences. The worldview or conceptual system under which any person operates during an act of creation constrains what is done, in a multitude of ways. Mark Johnson (1987) contends that much (if not all) of cognition is structured metaphorically, and that the metaphor applied during a given task will constrain performance. (Constraint here should not be read with negative connotations. A metaphor can enhance performance, but it does so through guiding
constraints.) Johnson recounts a study (by Gentner & Gentner) that identified two metaphorical models people use to understand electricity: the fluid-flow model (electric current is like water flowing through pipes) and the moving-crowd model (electric current is like a crowd of people moving through passages and gates). Subjects in the study were taught and instructed to employ one of the two models, and the metaphor determined which kinds of problems they were able to solve and which ones led to errors (each model worked well for half of the problems but not the others, and neither model is objectively correct). Johnson concluded that the models people employ are not merely applications of knowledge, but are rather constitutive of thought, or are cognition itself (pp. 109–112).

A transmission model of communication (Emmert & Donaghy, 1981, p. 353) applied to a textbook, where information is passed unproblematically from source to receiver, would work with an underlying book-as-container metaphor: the book contains information that simply must be extracted at highest possible fidelity. This model underlies any assessment of textbooks that stops at cataloguing the requisite information. However, it fails to suggest direction for instructional objectives.

Under a transactional model, the same textbook could be considered in performative terms, or with a book-as-experience metaphor. This should have implications for creativity—specifically here the design of print media—as models can be constitutive: “giving formal, definite, or organized existence to something” (OED, 2009). A performative or experiential conception of instructional media derives its significance not only in a critical sense, then, but also in a generative sense—it helps us to understand media but also to create it. This offers some promise as a model for improving the instructional effectiveness of media.

In order to explore the experiential basis of instructional media, the time signature of the reader’s engagement must be considered. It is not sufficient to analyze page structure in its own terms (if such an exercise is ever meaningful). Cognitive load theory (CLT) addresses the unfolding of instruction through time. As such, it provides focus at the level at which performance operates.

Cognitive load theory (Schnotz & Kürschner, 2007; van Merriënboer & Sweller, 2005; Paas et al., 2003) explains how cognitive limitations determine instructional
effectiveness. When a learner is occupied with a task, all conscious processing occurs in a very limited working memory. Working memory can manipulate somewhere in the realm of only 4 to 7 elements of information (Cowan, 2000; Van Merriënboer & Sweller, 2005). The cognitive load of a task refers to its demands on working memory. Part of the cognitive load at any given time is due to intrinsic load, the inherent complexity of the task. Germane load is the portion of cognitive load that contributes to learning, which CLT defines as changes in long-term memory. The final portion of cognitive load is extraneous load. Extraneous load is the undesirable byproduct of instructional format (Schnotz & Kürschner, 2007). This is the area where design most explicitly operates, or at least can most gloriously fail. Poor design can render information less accessible, resulting in high extraneous load. The designer must be very concerned with extraneous load. But design is much more than a filter over pre-existing content.

1.4 Designer Decision Making

Dual-coding theory (Paivio, 1986; Sadoski & Paivio, 2001) is built on the same model of working memory (Baddeley, 1998) as cognitive load theory. It focuses on the representational formatting of memory and sensory stimuli (including the text and pictures of media design). The text and pictures on the page, and particularly their arrangement and integration, have direct consequences for cognition. However, many designers are unaware of specialized knowledge regarding interpretation and the integration of text and pictures. Design education promotes visceral decision-making that rarely goes beyond considerations such as the appropriateness of style and basic issues of legibility. The culture of design practice reinforces the shallow concerns of the typical design education by rewarding designers based on subjective visual interest and technical accomplishment. The arrangement and integration of text and pictures on the page has very real cognitive outcomes (meaning instructional implications for learning), and should not be left to intuition. At present, there is no adequate and coherent normative model for designing instructional media. This study
should contribute to such a model. Any discussion of page structure must always be aimed at reader experience. In Rosenblatt’s terms, the text must purposefully be made to function as the poem.

1.5 Reader Construction from Page Structure

This study focuses on the design of instructional print media for science. Learning about science cannot truly occur through memorization of the established facts of its disciplines. It involves recognition and application of underlying abstract principles. The student of science lessons must operate something like an actual scientist, by recognizing or constructing abstract principles from given evidence (induction) and using those principles to interpret and predict future evidence (deduction). Overarching abstract concepts cannot be summarized in a single picture or text passage for the reader to simply “receive.” The student, basically, must perform.

The trail sign in Figure 1.1.1 that displays distances in both miles and kilometers is meant to stand somehow for a concept underlying units of measurement in scientific practice. But this simple photograph cannot hope to do so. It does nothing to make concrete the difference between inconsistent “English” and decimal systems. To help a student reader actively construct an understanding of the implications of the metric system, the unpicturable, abstract concept that undergirds the harmony and utility of it must be supported by multiple elements (text, picture, and schematic) in meaningful relationships. This is why it is critical that any normative theory of instructional design deal with content relationships—the integration of elements on the page—rather than focusing too narrowly on the elements individually. Isolated elements will not prove sufficient to help readers construct abstract scientific principles related to such things as interdependency in ecological systems, the relationship between matter and energy, and economic trade-offs. Defining such things and rendering them evident are different tasks entirely.
1.6 Comprehension and Learning

Cognitive load theory is very specific in its definition of learning (Schnotz & Kürschner, 2007, pp. 477, 492–493). Learning is the product of germane cognitive load: the construction and automation of schemas—mental structures that organize facts into knowledge—in long-term memory. Comprehension, on the other hand, occurs entirely in working memory. Comprehension occurs when all task-relevant elements are successfully held and manipulated in working memory. Learning is the ultimate goal of instructional media. Page structure is connected to long-term memory structures (schemas) through comprehension processes in working memory. Comprehension, then, represents a more direct connection between page structure and cognition. A test for learning would occur after interaction with media; a test for comprehension should occur while the reader is involved in interpretation processes. Learning, of course, should be a result of comprehension, but the strength of memory traces is not a given, and it remains eminently useful to consider both comprehension and learning individually, and then to compare them.

1.7 Argument in Summary

I argue that a new framework for graphic design is necessary if textbooks and other instructional media are to address reader experience productively. After Austin (1975), Ciardi (1959) and Fish (1976), I will refer to this generative framework as performative design. Performative design is distinguished from more limited graphic design practice in its emphasis on the cognitive processes of the reader (where designers usually restrict their attention to style and legibility). The preceding logical argument can be summarized as follows:

1) Science textbooks—print media with broad influence—focus attention on the provision of relevant content to the exclusion of the relationships between text
and pictures (the representational manifestation of that content). (Pettersson, 1998; Woodward, 1993b.)

2) Transactional theories of reading, in contrast, stress that reading material cannot be understood by its physical attributes alone, or what information is “contained” therein, but is actually constituted of the reader’s experience.

3) The designer of instructional media has direct control over the extraneous cognitive load, and indirect control over the resultant germane cognitive load, that results from on-task reader–page interaction. Because overall cognitive load is limited, control of load has profound implications for comprehension and learning.

4) Designers need support in making decisions that will positively affect cognitive load.

5) Science lessons typically involve abstract principles and relationships, which cannot be represented adequately by individual components, be they textual or pictorial. Abstract principles and relationships must be constructed by the reader through engagement (or performance) with systematic structures of multiple components, or text–image relationships.

6) Comprehension, the successful manipulation of all task-requisite elements in working memory, and learning, construction of schemas or changes in long-term memory, are products of cognitive load.

7) Text–image relationships, as constructed by a designer, in part characterize the associated and resultant cognitive load, which impacts comprehension and learning.

1.8 Study Focus and Research Questions

This study focuses on science instruction because the subject is especially visual, requiring maps, diagrams, and causal representations to explain many of its concepts.
Scientific concepts often involve very abstract relationships among phenomena, difficult to “picture” in text. There are expected overlaps between good design methods for science and other subject areas, but still, each discipline of study is likely to present certain affinities with certain methods.

Cognitive load theory provides an appropriate framework for shaping instructional media design, however, differentiating types of cognitive load is problematic. Significant design alterations should manipulate extraneous load. Furthermore, they help determine the potential for germane load.

The primary dependent variable for this study is comprehension as defined by cognitive load theory. The independent variable is meaningful text–image relationships on the page, as created through a coherent design strategy.

Illustrated books, including textbooks, are usually driven by a continuous prose passage, from which images are referenced. This represents a coherent strategy that will be referred to as prose primary (strategies are visualized in Figure 8.2.1). Sometimes prose is subsumed under imagery, as is the case with captioned illustrations. A design strategy in which the main text is broken up and follows imagery, or is below it hierarchically, will be referred to as prose subsumed. Prose subsumed designs are prose-based, but the prose is not continuous; rather, it is structured and subdivided by images. One common example of prose subsumed design is a captioned illustration. Finally, text can be incorporated into image-defined spaces, as snippets of related textual information, rather than as a continuous structure of prose. A strategy that uses the entire space of the page as a diagram, rather than one that contains diagrams, will be referred to as fully integrated, in terms of text–image integration.

A prose primary design can include diagrams, which in turn include text (often as labels), but they are referenced sidebars to the meaning-driving continuous prose. A prose subsumed design can also include diagrams, and these diagrams function in part as structuring devices (like headers), but a strong prose is still available outside of their space. A fully integrated design is essentially a diagram, and though considerable text might be included, it is not accessed in a fully prescribed order—it is a truly parallel system. The
reader is more responsible for reading order. Note that these differentiations are beyond the level of simple layout, because a change in strategy involves more than moving elements around: the elements themselves must be altered to keep information, and so resultant constructed meaning, coherent.

These concerns lead to the primary research question:

- Research question #1: How does the strategy for text–image integration in instructional print media for middle school students impact their comprehension of abstract relationships in science material?

This study seeks authenticity by focusing not on an isolated design element but rather holistic strategies for producing complex print media. A strategy produces a sensible, consistent manipulation of variables. If a continuous prose is written before illustrations are considered, then a prose primary strategy is in effect. The employment of such a strategy will dictate or at least constrain countless design decisions.

Only print media (and no screen based media such as web sites or apps) will be addressed by this study.

All possible relationship strategies cannot be tested, so prose primary, prose subsumed, and fully integrated strategies will represent points on a spectrum of possibilities. Comprehension must be tested while students have access to instructional print media—while they are actively using it—while learning would have to be tested some time afterwards. Comprehension of textual and imaginal information will be measured by test performance with media.

A further measure, mental effort, could offer insight into the effect of media on comprehension (and by extension, learning) because it implicates difficulty. Task difficulty is the result of extraneous and intrinsic cognitive loads, and since cognitive load is additive and restricted, it limits germane load (and so learning). Subjects can estimate their mental effort as they perform print-based tasks. This leads to a further question regarding strategies for text–image integration:
• Research question #2: How does the strategy for text–image integration in instructional print media for middle school students impact their perception of task difficulty?

Perception of task difficulty should impact a student’s approach to instructional media. A related area of interest is interest level itself.

• Research question #3: How does the strategy for text–image integration in instructional print media for middle school students impact their interest level in subject matter?

It may be that a certain text–image integration strategy affects the degree to which a student is interested in the subject matter presented through instructional media. But it may also be the case that student interest is driven more directly by the visual qualities of the print media itself.

• Research question #4: How does the strategy for text–image integration in instructional print media for middle school students impact their interest level in instructional print media?

These research questions probe the instructional implications of text–image integration strategy.

The literature review elaborates upon the underlying issues of the conceptual framework. The subsequent section on methodology builds on the conclusion of this section and the definition of the research questions.
2.1 Constitutive Metaphor

2.1.1 Beyond Figurative Metaphor

Lakoff and Johnson (1980) identify metaphor as a constitutive aspect—or even as the primary structuring device—of our conceptual system. Metaphor is often considered a mere literary device. But Lakoff and Johnson offer overwhelming linguistic evidence that metaphor is not just a discursive technique and that it actually underlies our conceptual understanding. The linguistic evidence is in how we express different concepts, or more to the point, the consistent metaphors we use to express concepts. One example is the *theories (and arguments) are buildings* metaphor evident in the following language (Lakoff & Johnson, 1980, p. 46):

The theory needs more *support*. The argument is *shaky*. We need some more facts or the argument will *fall apart*. We need to *construct* a *strong*
argument for that. I haven’t figured out yet what the form of the argument will be. Here are some more facts to shore up the theory. We need to buttress the theory with solid arguments... The argument collapsed... We will show that theory to be without foundation. So far we have put together only the framework of the theory.

Any single example from this list would suggest nothing, but the consistency of the language is compelling.

The employ of metaphors such as theories are buildings is unconscious (p. 3). If we understand by way of metaphor, then a concept such as theory effectively is a metaphor. That is, it is metaphorically structured or constituted. By extension, potentially all thought is constituted of metaphorical association. We build new complex knowledge from existing knowledge. Metaphors are utilized in other metaphors. The conduit metaphor, for example, which describes “our language about language,” is constituted of: (a) ideas (or meanings) are objects, (b) linguistic expressions are containers, and (c) communication is sending (p. 10).

This naturally leads to speculation regarding root metaphors that are foundational to other metaphor use. Lakoff and Johnson state here that the experiential basis of metaphor is unknown, but Johnson (1987) would later provide this basis through image schemas.

The structure of individual metaphors is coherent by means of goodness-of-fit. A concept employed in metaphor is often multidimensional. For instance, in the argument is war metaphor, war is effectively mapped onto conversation to delineate argument. Select characteristics of war are co-opted based on goodness-of-fit. Not all aspects of war are co-opted. Lakoff and Johnson (1980) call the structure of counterpart an experiential gestalt (p. 81). There are structural characteristics of war, including participants (roles), parts (the two positions, planning strategy, attack, etc.), stages, linear sequence, causation, and purpose (victory) (pp. 80–81), that people involved in an argument must implicitly agree upon in order to even argue in the first place.

Coherence of metaphorical connections does not always involve consistency. In our conceptions of time, it is either a moving object coming at us or it is stationary and we are
moving towards it (p. 44). These two images are inconsistent, but nevertheless remain coherent from the point of view of the observer. There is still goodness-of-fit.

Most of the fundamental concepts identified by Lakoff and Johnson are organized into “spatialization” metaphors, such as good is up and bad is down (p. 17). These spatialization metaphors are systematic, such that “I’m feeling up” and “my spirits rose” both indicate happiness (the system is coherent through orientation). Furthermore, there is an “external systematicity” operating outside of individual spatialization metaphors. Up is considered positive (I can’t help but invoke the underlying metaphor in describing it) across many domains including happy, health, alive, and control (p. 18). Lakoff and Johnson suggest that the metaphors we use are based both physically and culturally. Johnson’s (1987) bodily schemas are suggested as the metaphors underlying other metaphors, with a physical basis for the spatialization of so many more complex metaphors.

2.1.2 Schema Theory

Spoehr and Lehmkuhle (1982) define schema in pictorial terms, as “a mental structuring of data that embodies necessary real-world constraints on the content and organization of [a] picture and thus contains expectations about what should appear in the picture” (p. 169). Jean Matter Mandler (1984) identifies some fundamental schema types—events, stories, scripts and scenes—which by the nature of their formatting extends the definition of schema beyond merely pictorial concerns. While scenes are indeed spatial and so pictorial by nature, events, stories and scripts are more temporal—or so they seem at first. Mandler acknowledges extra-schematic forms of knowledge in terms of organization, such as serial (representation of the alphabet), categorical (class-membership of animals) and perhaps matrix structures (pp. 3, 10–11). However, schemas are considered by some psychologists to be the “basic building block of cognition” (p. 2) and so may underlie these forms. Mandler offers as an example of serial structure, the child playing a piece on piano, who makes a mistake halfway through and can only continue by returning to the beginning (p. 12). Once the piece is more familiar to the child, a hierarchical structure might form so that it becomes
possible to re-enter the piece at some (still limited) number of points. This is chunking. The child can enter the serial structure at the beginning on any individual chunk. The resulting segmentation implies a temporal schema. While the serial structure of the piano piece is fully specified, a temporal schema (which could be generated from or for the same piano piece) is more a “class of events” (p. 12).

A schema provides a set of expectations to check phenomena against or to order phenomena. In general, schemas specify part–whole relations (p. 14). This means that there are connections between the items in any given structure. In event schemas these connections are serial (temporal), and they can be, in order from strongest to weakest, causal relations, temporal (invariant) relations, or arbitrary temporal connections (p. 14). In scene schemas, on the other hand, the connections are spatial or topological relations: up, down, next to, inside of, etc. (p. 15). Some relations are obligatory, such as a table resting on the floor, while others are unspecified or “optional” (p. 16). Schemas are more tightly interconnected than categorical organizations because connections are vertical as well as horizontal. In an event schema, horizontality refers to the serial connections and verticality refers to connections from part to whole. (In the piano example, the child’s expert ability to enter the piece at multiple points stems from part-to-whole connections.) A scene schema includes inventory information and spatial-relation (layout) information. There are more spatial connections in a scene schema than there are temporal connections in an event schema, leading Mandler to designate scene schemas as “heterarchies” with as many as three dimensions of connections (p. 16).

Ancient mnemonic techniques took advantage of the heterarchicality of scene schemas by encoding information (serial or no) into visual schematic arrangements (Carruthers, 2008). The method of loci involves use of a familiar space into which symbols are placed for the purpose of memorization and recall. With proper technique, any kind of information can be symbolized and so incorporated. Mnemonics was an integral part of argumentation: a speaker had to memorize speeches in chunks so that access wasn’t limited to a prescribed order, as in the case of the novice child pianist. The chunks of a speech encoded with the method of loci can be delivered in any order, no order more easily than
another, because the arrangement is spatial (pp. 89–98). (Mnemonics is covered further in the section on memory.) Though serial information can thus be spatialized, it’s important to reiterate that event schemas are themselves already schematic by definition and not simply serial. Mandler (1984) notes that temporal and spatial schemas are remarkably similar (p. 1). It may be appropriate, at least analogically, to consider all schemas to be spatial in nature.

Schemas are thus networks of information that vary somehow by type. Mark Johnson’s (1987) image schemas represent what is effectively a molecular schema theory as they are purported to be the basis of any understanding.

2.1.3 Image Schemas

Johnson (1987) offers a molecular theory tying together metaphor and schemas. Understanding must be built up from something, and from an earlier state. Johnson proposes that the foundation of knowledge structure is bodily experience. All that we know can be traced back to fundamental understandings of what it is to be embodied, in the world. These molar components are image schemas. Image schemas are preconceptual: they enable conceptualization. Image schemas are basic understandings that emerge out of bodily experience. Some image schemas are containment, compulsion, out, blockage, counterforce, path, cycle, scale, and center-periphery. The containment schema, for example, is the result of an intimate understanding of our bodies as containers into which we put things and out of which other things emerge. We experience liquid flowing in and out of external containers. We move in and out of rooms, vehicles and clothes (p. 21). The commonalities of these everpresent experiences serve as a way to metaphorically structure other experiences and to conceptualize such things as “being in love.” The containment schema is thus a product of being in the world that is then put to use. Image schemas have entailments, or consequences of their structural characteristics. For example, one entailment of the in–out orientation of the containment schema is that containment offers resistance to external forces (p. 22). The entailments of an image schema define it.
Image schemas are not mental pictures—they are more general and abstract than that (p. 24). Johnson describes them as *dynamic* in that they are “a primary means by which we *construct* and *constitute* order” while remaining contextually flexible (p. 30).

Image schemas are “highly *schematic* gestalts [thus the name] which capture the structural *contours* of sensory-motor experience, integrating information from multiple modalities” (Hampe, 2005, p. 1). Spatial and temporal orientation schemas are prevalent and underlie our use of language. A study of verb–particle constructions with *out* (600 cases) identified consistent patterns in use, based on only three different *out* schemas, implying that language (and so conceptualization) is constituted of image schemas (Johnson, 1987, p. 32).

### 2.1.4 Discussion

Metaphor involves projecting selected qualities of one structure onto another. For the use of metaphor to make any sense—or to be productive—the structures must be understandable in and of themselves. Schemas allow us to make sense of phenomena. Though image schemas—the molar basis of other schemas and ultimately all metaphor use—are not pictures, they are spatially (and temporally) defined—at least often if not always.

### 2.2 Creative Memory

#### 2.2.1 Memory as Constructed

Human memory is easily confused with computer memory, which results in misleading conceptions. A file saved on a computer, barring corruption or some other disastrous event, maintains its form perfectly and can be extracted (opened) as such. The research of Elizabeth Loftus on eyewitness testimony and “repressed” memories reveals human memory to be dissimilar to this carbon copying (Mook, 2004, pp. 199–202). Loftus shows that though memories point to past experience, they are constructed in the now, just as
they are “recalled.” She demonstrates that respondents report differentially on shared experience based on the questions asked them (she refers to these as leading questions). In one cited study, subjects who were shown a videotape of an auto collision varied their estimates of the speed of the cars predictably due to the terminology employed by the researchers to prompt an estimate, with “smashed into” (“how fast was the car traveling when it smashed into…”) resulting in higher speed predictions than “hit” (p. 200). Furthermore, leading questions about nonexistent events (the breaking of a headlight, for instance) resulted in respondents recalling them. Thus recall is seen as an act of reconstruction, rather than perfect-fidelity accessing.

Memory doesn’t suffer only from infidelity at the point of recall. Consistent with schema theory, we need to give information form as we encode it: “To organize is to memorize and to memorize is to organize” (Mandler, 1984, p. 8). Sensory information isn’t merely acquired as if it were poured into the head. Rather, it is either lost or it is converted into something usable or “memorable.” That which is committed to memory is given schematic structure. The classical and medieval practice of mnemonics was predicated on the purposeful manipulation of the schematic nature of memorizing or knowing. Carruthers (1998) characterizes this practice as “the construction of thinking,” equivalent to the contemporary definition of cognition (p. 2).

2.2.2 Mnemonic Manipulation of Memory

Mary Carruthers (1998 & 2008) studies medieval memory culture, or the mnemonic practice of medieval scholars. Medieval memory training was considered “an instrument of thought, employing particular devices for specific goals and uses” (Carruthers, 2008, p. x). While in contemporary culture mnemonics is relegated to parlor tricks, in medieval practice it was central and critical to intellectual practice. Carruthers (1998) is careful to ensure that her readers understand the medieval sense of memory (p. 4):
I must ask of my readers a considerable effort of imagination… to conceive of memory not only as ‘rote,’ the ability to reproduce something (whether a text, a formula, a list of items, an incident) but as the matrix of a reminiscing cogitation, shuffling and collating ‘things’ stored in a random-access memory scheme, or set of schemes—a memory architecture and a library built up during one’s lifetime with the express intention that it be used inventively… Meditation is a craft of thinking. People use it to make things, such as interpretations and ideas, as well as buildings and prayers.

This constructivist notion of memory as creative is entirely consistent with Loftus’ findings (Mook, 2004, pp. 199–202). The aforementioned method of loci (Carruthers, 2008, pp. 89–98), where information is encoded and mentally arranged in a complicated symbolic space, attests to the sense of memory as “architecture.” This allows even purportedly serial information to be extracted in any order as chunks, which is how such memory qualifies as random-access memory. It is the spatialization of information that made medieval memory practices possible.

Though there is any number of impressive examples of the power of mnemonics for the recall of information, the parading of memory tricks is beside the point, and perhaps even misleading. Many of the techniques employed are relatively cumbersome compared to simple rote memorization strategies. What is important is how mnemonic technique enables the use of memorized information. Information encoded into a random-access memory structure provides the mnemonist with a divisional system that not only provides free access in chunks but also imposes a kind of meaning onto it (pp. xii–xiii, 99–152). Mnemonic technique, then, is a purposeful means of utilizing schemas—schemas are usually unconscious in use.

2.2.3 Memory Modeled in the Book

The medieval manuscript, or any religious codex, is the object around which the most memory work was employed. Noakes (1988) describes reading in medieval times as a
particularly slow process. While modern books are read quickly (and written or designed \textit{to be read quickly}), the assumption held by a medieval scribe was that each page would be pored over exhaustively, read aloud in small groups, and discussed. The resultant page design is charged with thick meaning and operates with rhetorical devices immediately recognizable at the time by trained readers. The page of a medieval manuscript is best thought of as a complex space that offers much for its readers to do. It is actually misleading to think of the book as being \textit{read} in medieval times, since the contemporary conception of reading is both passive and comparatively straightforward; rather, the designed page instigated an ongoing involving act of labored interpretation.

The complexity of page design, with its illuminations, illustrations and marginalia, represented a strategy for learning. Control was exerted largely through the visual characteristics of the designed page with the understanding that a certain reading strategy would be employed. Lewis (1995) describes the strategy: “Pictures transform words into another realm of reality by moving the reader from the abstraction of written signs (words) to another level of comprehension where optical perception not only authenticates the experience but determines the direction of the reader’s understanding” (p. 10). Central to the practice of meditative reading was the use of imagery (Carruthers, 1998). The mnemonic practitioner was adept at converting information into mental pictures and organizing them schematically. The copious illustration of a medieval manuscript aided in this practice by providing material for memory work. It was even suggested that a scholar always use the same book (when alternates were available) for the study of a given subject, so that the visual characteristics of the one could better serve his or her purpose (note that books at this time were lettered and illustrated by hand, so no two copies would be the same).

There are many instances of linguistic imagery in medieval manuscripts. Oftentimes the border illustrations served as puns to involve the reader in a game of translation, where a utilized picture’s name punned on “border,” for example, and through further punning reinforced the gist of the accompanying passage (pp. 161–170). Dirk Bax (1978) provides convincing evidence that this practice of encoding Latin and vernacular language into pictures extends beyond books. The paintings of Hieronymus Bosch are constructed out of
puns, and so could be “read” by illiterate readers. In summary, the complementary use of language and image characterizes medieval graphic design, and served as a one-to-one analog of memory work, or cognition.

2.2.4 Working Memory

Anderson (1995; in Marzano, 2001, p.29) discredits the division of memory into separate types as short- and long-term memory. Instead, memory is of one type with different functions. Sensory memory temporarily contains relatively complete accounts of information that decay quickly unless encoded into long-term memory. The act of encoding, of course, determines the qualities of the resultant information in long-term memory. (Medieval mnemonists were especially concerned with this process.) Long-term memory is where information is effectively stored, in schematic and other structures. Working memory—the seat of consciousness—is where information is actually processed. Information is retained in working memory only through conscious focus, though its capacity is severely limited (pp. 29–31). If not rehearsed, information is lost in about twenty seconds (van Merriënboer & Sweller, 2005, p. 148). Loftus (Mook, 2004; Spoehr & Lehmkuhle, 1982) and Mandler (1984) note that information is manipulated (in working memory according to Anderson’s theory) when stored in permanent memory or “retrieved” from it.

There is copious evidence to suggest very real subdivisions of working memory. Baddeley (1998) presents a three-component model of working memory comprised of the visuospatial sketchpad, phonological loop, and central executive. Together these components act as a system of conscious awareness that “allows the various channels representing objects in the world to be bound together as a unitary experience, which in turn allows the constituents of an episode to be bound in [long-term] memory, facilitating subsequent recollection” (p. 168). Selective attention is a function of working memory that allows for discriminable amounts of information to be extracted from the cacophony of sensory experience. Working memory also supports a reflective capacity, so that material can be evaluated for efficacy and treated accordingly.
The central executive is required to coordinate multiple simultaneous tasks. The central executive is poorly understood and Baddeley refrains from postulating other functions, though he does suggest that it has sub-components. Van Merriënboer and Sweller (2005) identify schemas as playing the role of the central executive, as they organize the information that is to be processed in working memory (p. 149). Schemas effectively increase working memory capacity, as a schema, complex as it is containing multiple elements, can function as a single element. Through practice, schemas can become automated as the central executive (ibid.). The visuospatial sketchpad and phonological loop are more clearly defined than the central executive by empirical research. Their division has been established by studies showing little interference between linguistic and imaginal information (Baddeley, 1998, p. 169).

The visuospatial sketchpad is a complex of sub-systems that allows temporary maintenance of representations. The sub-systems are located in different regions of the brain, and maintain information such as shape and color and spatial arrangement.

The bulk of research on working memory has focused on the phonological loop, which deals with linguistic information. The phonological loop probably evolved for language acquisition. Adults who later in life experienced diminished cognitive capacities related to the phonological loop show only minor effects, while developing children with phonological loop deficiencies experience profound difficulty with language (p. 171). The phonological loop may have a role in comprehension, but it is not a critical one.

The phonological loop has two components: a temporary store and a rehearsal process. The phonological store has a somewhat auditory basis as suggested by the acoustic similarity effect: subjects who attempt to remember letter sequences will tend to make errors based on letter sounds, recalling $v$ as $b$, for instance (p. 170). Visual verbal information thus appears to be immediately encoded acoustically. Subjects attempting to remember words in working memory don’t tend to make errors of similarity of meaning, but they do when the exercise requires or allows use of long-term memory. This draws a stark contrast between working and long-term memory. The memory trace in the phonological store fades in just seconds, unless the rehearsal process is maintained. The rehearsal process is an active
repetition akin to sub-vocalization. While the temporary store fades quickly, the rehearsal process can last indefinitely with concentration. Subjects can recall about as much as they can say in two seconds’ time, suggesting a rough maximal length for the rehearsal process.

Working memory is defined by its very limited capacity. Its limitations are in fact integral to its functionality (van Merriënboer & Sweller, 2005, p. 155). Long-term memory must be stable and working memory must be dynamic. If working memory had greater capacity, it would destabilize long-term memory and result in an organism incapable of retaining beneficial adaptations. Three elements can be combined in six different ways, but ten elements have a staggering three million combinations, so even a small increase in working memory capacity can dramatically increase the rapidity of alterations to long-term memory and so destabilize it (ibid.).

Working memory then requires a low limit. The view of short-term memory limitation in the rehearsal process is well defined theoretically, and it’s not immediately clear how the famous “magical number seven plus or minus two” bits of information limit estimate resolves with it, if at all. Cowan (2000) calls this old estimate of short-term memory capacity (7±2) into question, suggesting that it was not even meant as a serious estimate in its original publication. Because of the associative nature of memory, the idea of a “bit” of information is problematic. What could be considered multiple bits are sometimes held together in chunks. Chunking extends capacity. A subject required to hold the letter sequence fbichsibmirs in short-term memory can extend his or her capability by recognizing familiar three-letter sequences embedded in the string: FBI, CBS, IBM and IRS (p. 90). Each three-letter sequence becomes one chunk in this example, as they are each already stored in long-term memory. When working memory utilizes information already stored in long-term memory, the characteristics of working memory are fundamentally altered (van Merriënboer & Sweller, 2005, p. 148). Cowan’s (2000) estimate for short-term capacity, supported by a body of research, is four plus or minus one items (or chunks). Van Merriënboer and Sweller (2005) estimate that working memory stores around seven “elements” but operates on just two to four (p. 148). Again, there is some level of dissonance between the specific
explanation of the phonological loop (whatever information fits within a two second cycle) and estimated counts of items or elements.

2.2.5 Discussion

Memory is far from passive or perfect. Memorization and retrieval both involve schematic manipulations of information. In medieval mnemonic practice, scholars recognized and made use of this fact by purposefully controlling their schematic organization of information. The resultant medieval plan of the book represents a remarkable complement of page design to cognition, where each one presupposes the other. Mnemonic book design promoted schema use to reduce the demands placed on working memory.

2.3 Cognitive Load Theory

2.3.1 The Cognitive Load Construct

Cognitive load theory (Schnotz & Kürschner, 2007; van Merriënboer & Sweller, 2005; Paas et al., 2003) accounts for working memory limitations—the theory is based on working memory assumptions detailed above—in order to improve educational instruction. According to cognitive load theory (CLT), “well-designed instruction should not only encourage schema construction but also schema automation for those aspects of a task that are consistent across problems” (van Merriënboer & Sweller, 2005, p. 149). Since novel information must be dealt with by a severely limited working memory, schema construction and automation is necessary to decrease the load (or demands on working memory) and enable expert behavior and more complex learning.

CLT focuses on a learner’s experience with a particular task. Cognitive task performance is defined as the transformation of a “mental representation from a less favorable state to a more favorable (goal) state by cognitive operations in working memory”
CLT seeks to assess the demands a learning task places on working memory. Task demands are determined both by the task itself and the individual learner. Some important task characteristics are “task format, task complexity, use of multimedia, time pressure, and pacing of instruction” (Paas et al., 2003, p. 64). Learner characteristics in part determine task difficulty, including expertise level, age, and spatial ability (ibid.). Interaction between task and learner produces a mental load, which represents the total demand on the learner’s working memory (ibid.). If mental load exceeds a learner’s total available cognitive load, then he cannot complete the task. Paas et al. (2003) identify the main hypothesis of CLT: “Differences in effectiveness between instructional formats are largely based on differences in memory load” (p. 69).

Available cognitive load can be decreased from its theoretical maximum by low mental effort. Mental effort is the amount of available cognitive load a learner commits to a task. It functions as a ceiling, an effective maximal cognitive load at any one time. Paas et al. (2005) suggest that CLT research has tended to ignore the impact of motivation on resultant mental effort because the laboratory (often the setting for CLT studies) enjoys a higher level of invested mental effort than does the classroom (pp. 26–27).

Whatever resources are available given the upward limit on available cognitive load (accounting for ceiling effects of mental effort) are divided amongst three distinct types of cognitive load: germane load, intrinsic load, and extraneous load.

Intrinsic cognitive load is the portion of cognitive load that results from the natural complexity of a learning task. Intrinsic load is a product of element-interactivity: “the number of cognitive elements that have to be held simultaneously in working memory” (Schnotz & Kürschner, 2007, p. 476). However, intrinsic load of a learning task is basically only intrinsic to, or fixed for, the learner’s level of expertise. Since cognitive schemas can function as individual elements, a more experienced learner—one with relevant schemas—might have to deal with functionally fewer interacting elements than a novice (ibid.). Intrinsic load can be
manipulated in the sense that the instructor can assign tasks based on learner expertise. However, *at a given level of learner expertise*, a task’s intrinsic load is fixed.

Germane cognitive load is related to schema construction and schema automation. Schema construction represents changes in long-term memory, which CLT defines as learning and gaining expertise (p. 477). Schema automation is the result of practice with constructed schemas, and represents an increase in expert performance at a task or class of tasks. Germane load is the result of “effortful learning,” differentiating the construct from unconscious learning (p. 476). Simple task performance is not germane load (p. 496). Germane cognitive load is constrained by intrinsic load because “it is not possible to reflect deeply (high germane load) about a very easy task (low intrinsic load)” (p. 497). A learner’s self-regulation delimits germane load (ibid.). The ultimate goal of CLT-inspired instruction is to maximize germane load, thereby maximizing learning.

Extraneous cognitive load is the portion of cognitive load generated exclusively by instructional format. Schnotz and Kürschner (2007) cite two working definitions of extraneous load by CLT researchers: (1) it is the result of “unnecessarily high element-interactivity in working memory” from instructional format; and (2) it is the result of irrelevant cognitive activities (p. 476). As it is unequivocally desirable to increase germane load, extraneous load should be reduced as much as possible.

### 2.3.2 Learning and Understanding

Cognitive load theory specifically differentiates learning and understanding. Learning is “an increase in expertise due to an alteration in long-term memory. If nothing has been altered in long-term memory, nothing has been learned” (Schnotz & Kürschner, 2007, p. 477). Learning then takes the form of schema construction and schema automation, which contribute to germane load. Schemas, once constructed, change what an individual “treats as an element” in working memory (ibid.). This effectively decreases the intrinsic load of a given learning task as a learner acquires relevant schemas—in other words, as the learner gains expertise. Schema automation is the product of practice with constructed schemas.
When a task becomes fully automated, it ceases to place demands on working memory (as working memory is conscious only), freeing its limited resources for other conscious activity (ibid.).

Understanding relates to element-interactivity. It is a state of the here-and-now in a learning task. Understanding “occurs when all relevant elements of information are processed simultaneously in working memory” (Schnotz & Kürschner, 2007, p. 477). Depending upon the difficulty of a given task, understanding may be contingent upon previous schema construction and automation, i.e., learning. However, understanding is differentiated from learning in the focus on working memory. The process of creating declarative knowledge structures in working memory is an achievement of understanding; “transforming” those structures into long-term memory schemas constitutes learning (p. 492). It should be theoretically possible for intrinsic and extraneous loads to render a student who achieves understanding during a task unable to learn from it. It is unclear exactly how much Schnotz and Kürschner differentiate between understanding and comprehension. Comprehension involves the construction of mental representations in working memory (ibid.). Comprehension will henceforth be used to stand for the CLT definition of understanding inclusive of the CLT definition of comprehension itself.

Comprehension is a factor of performance because it refers exclusively to manipulations in working memory (as opposed to long-term memory). Intrinsic and extraneous load are “performance-related concepts,” while germane load, again, is related to learning (p. 493). One important implication of this is that “making a task easier does not necessarily result in better learning” (ibid.).

CLT is only relevant for “biologically secondary knowledge,” that is, knowledge that is culturally mediated (p. 493). Biologically primary knowledge is acquired without conscious effort. For example, facial recognition is automatic and assumed separate from working memory (p. 474). Biologically secondary knowledge may also be acquired without conscious intent sometimes, meaning that not all biologically secondary knowledge can be explained by CLT. If a learner is unable to verbalize or unaware of what he has learned, then unconscious learning, or implicit learning, has occurred (p. 494).
2.3.3 Expertise

According to CLT, experts and novices differ in schemas. Experts have task-relevant and possibly automated schemas that can act as individual elements, thereby effectively reducing cognitive load. Novices lack task-relevant schemas, and can thus handle less element-interactivity. Schnotz and Kürschner (2007) use Vygotsky’s zone of proximal development to align expertise with task difficulty (p. 479). A task that is well matched for a learner’s level of expertise minimizes extraneous load, leaving load intrinsic to the task and, hopefully, resources for germane load. For a novice learner a simpler task is ideal, where an expert learner is best served by a more demanding task. Extraneous load increases not only when a task is too demanding for a novice, but also when that task is not challenging to an expert. Schnotz and Kürschner explain such misalignments through three demonstrable effects: split-attention, redundancy, and modality effects.

The *split-attention effect* occurs when multiple information sources that must ultimately be mentally integrated are presented in isolation of one another (p. 480). For instance, when a diagram and corresponding text are “unnecessarily” separated, extraneous load increases for novices, who must themselves cognitively integrate the sources (ibid.). The extraneous load is reduced if the diagram and text are physically close together. This is known as the *spatial contiguity principle*. However, for experts such physical integration on the page can actually produce extraneous load, because though the expert might not need the textual information in order to understand the diagram, he or she is likely to be distracted by it (p. 481).

The *redundancy effect* refers to the unneeded physical integration of information sources that are perfectly intelligible in isolation (ibid.). This effect is obviously derivative of the split-attention effect.

The *modality effect* occurs when pictures (visual modality) are supplemented with the spoken word (auditory modality) rather than text (visual modality) (ibid.). While picture and text split visual attention, picture and spoken word minimize split attention by presenting information in two separate modalities. However, as with the similar split-attention effect,
the benefits inherent in presentation of picture with spoken word disappear or even reverse with experts, who are basically forced to attend to unneeded verbal statements (ibid.).

Expertise and task difficulty misalignment, then, can generate extraneous load in numerous ways, including:

- Interactivity between relevant information: The learner “does not possess sufficient cognitive [schemas]” for a task that ultimately requires excessive element-interactivity. This is a case of task difficulty exceeding expertise (Schnotz & Kürschner, 2007, p. 481).
- Maintenance of relevant information: The learner overtaxes the “limited duration of information in working memory” because information that must be integrated is presented physically segregated. The extraneous load is a result of having to maintain elements in working memory (pp. 481–482).
- Interactivity between irrelevant information: Excess information is physically integrated when such integration is unnecessary, and so is the resulting high element-interactivity—this is the redundancy effect (p. 482).
- Waste of time and effort: The learner processes unneeded information that doesn’t result in high element-interactivity but simply distracts from other possible, more productive tasks (ibid.).

2.3.4 Instructional Implications

CLT is aimed at developing numerous instructional strategies of reasonable working memory demand (Paas & van Gog, 2006; van Merriënboer & Sluijsmans, 2008; Mayer & Moreno, 2003; Ayres, 2006). Learning occurs in environments, which, according to van Merriënboer’s four-component instructional design model (van Merriënboer & Sluijsmans, 2008), are constituted of:
• Learning tasks: Meaningful, “whole,” real-life tasks that “require the integrated use of knowledge, skills, and attitudes.”

• Supportive information: Instructional material that assists learners with task-relevant problem solving and reasoning.

• Procedural information: Provided information that “points out to learners how to perform the routine aspects” of the tasks.

• Part–task practice: Modeled routine practice that aids in automation. (p. 57)

Scaffolding reduces extraneous load by providing guidance as task difficulty increases, keeping learners “on-task” and away from irrelevant aspects (ibid.).

Van Merriënboer and Sweller (2005) summarize “major effects” that can increase germane load by decreasing extraneous load:

• Goal-free effect: “Focus learner’s attention on problem states and available operators” rather than on discrepancies between current problem states and distant goal states.

• Worked example effect: Instead of relying on the learner’s weak problem solving methods, draw attention to problem states and useful (provided) solution steps.

• Completion problem effect: Focus attention again on problem states and useful solution steps by providing a partial solution.

• Split attention effect: Reduce mental integration by physically integrating information sources on the page.

• Modality effect: Utilize both visual and auditory processors by accompanying visual non-verbal information with spoken information.

• Redundancy effect: “Replace multiple sources of information that are self-contained… with one source of information.” (p. 151)
Though intrinsic cognitive load is fixed for a task at a level of expertise, it can be manipulated through task modification. Intrinsic load can be decreased through two-stage instruction, one where only a small percentage of requisite information is covered, and a second where all information is covered holistically. Though the first stage is weak in authenticity, it enables the learner to develop schemas. These are then utilized in the next stage, where complex information is chunked and can then be brought to bear in a more authentic episode (van Merriënboer & Sweller, 2005, pp. 156–157). Gerjets et al. (2004) propose a similar strategy, where solution procedures are presented first from a molar view and subsequently from a modular view. This strategy is aimed at reducing intrinsic load in early instruction.

2.3.5 Measurement

Schnotz and Kürschner (2007) bemoan that the subdivisions of cognitive load cannot as yet be measured with reliability and validity (p. 500). Typically, attempts are kept to overall measures of cognitive load or mental effort. The “element-interactivity assumption” is that “cognitive load effects can only be demonstrated using material with high element interactivity” (p. 477). If intrinsic load is low, germane load may be available despite high extraneous load. Therefore, task demands must exceed overall mental load if ceiling effects are to manifest.

Cognitive load measures have varied in terms of objectivity and directness (Brünken et al., 2003). Indirect, subjective measures take the form of post-treatment Likert scales that ask subjects to assess their invested mental effort (p. 56). These ratings can be given in an authentic educational setting, which produces high ecological validity. Furthermore, results have proven reliable (Schnotz & Kürschner, 2007, p. 499). However, the relation between reliable subjective perception and actual cognitive load cannot be certain, especially because low mental effort might signify either low cognitive load or such high load that the subject disengages (Brünken et al., 2003, p. 56).
When subjects rate the difficulty of materials (as opposed to their own mental effort), the measure is subjective and directly related to “the cognitive load imposed” (p. 56). Task difficulty, learner expertise, and variation in “attentional processes” all pose a problem for such methods (ibid.).

Indirect but objective methods include performance outcome measures. Knowledge acquisition is measured post-test, for two or more conditions. The indirectness stems from dependent processes of information storage and retrieval (p. 56). Varying levels of learner expertise are a difficulty for performance measures. Physiological measures have been used as well. Galvanic skin response, pupillary dilation, and heart rate variability have all been indirectly tied to cognitive load, though not convincingly (Schnotz & Kürschner, 2007, pp. 499–500).

The dual-task paradigm provides a direct, objective measure of cognitive load (Brünken et al., 2003, p. 57). If a learner is required to perform two tasks simultaneously that utilize the same working memory resources (visual or verbal), and performance on the primary task is measured with and without secondary task interference, then secondary task load (the variable of interest) can be isolated (ibid.). This is an experimental method and is low on ecological validity. Furthermore, “one cannot exclude the possibility that the individual tries to keep [one] task performance constant at the expense of [the other]” (Schnotz & Kürschner, 2007, p. 500).

Brünken et al. (2003) suggest that neuroimaging techniques will offer a most direct, objective measure of cognitive load in the near future (pp. 56–57). But practical issues must be overcome, and ecological validity would be extremely low, considering regular and practical instructional techniques.

Paas and van Merriënboer (1993) combine mental effort and performance in a measure of efficiency. A learner is efficient if either (a) performance is higher than expected given mental effort, or (b) mental effort is lower than expected given performance (p. 738). Paas and van Merriënboer provide a statistical measure for calculating “relative condition efficiency” (pp. 739–742). Measurements finer than mental effort (that is, different kinds of load) remain a difficulty for CLT.
2.3.6 Discussion

Cognitive load theory remains in its formative stage (the first full book on the theory only emerged in 2010, after much of the preparation for field work for this dissertation was complete [Plass, Moreno & Brünken, 2010]). Methods of measurement are being refined and challenged. The inability to isolate one kind of load from others poses problems for interpretation. Inferences regarding extraneous load seem the most valid. Germane cognitive load was only added to the model in 1998 (Schnotz & Kürschner, 2007, p. 472; Sweller et al., 1998). Nevertheless, CLT is a remarkably coherent and elegant extension of working memory research to instructional design.

2.4 Dual-Coding Theory

2.4.1 Representational Coding

When information is obtained, its formatting must delimit its properties. Information is represented somehow in the brain. Representational format determines, in part, what information is explicit and accessible, and so it has implications for function (Kosslyn et al., 2006, p. 8). Mental representation is a matter of encoding. The question of representational format is not one of “superficial” input or output, but rather of truly fundamental distinctions.

Exclusively propositional (or verbal) representation is championed by some (notably Zenon Pylyshyn) as the representational format of information in the human brain. Propositions are something like verbal expressions, and so the propositional argument is a language-first argument. However, propositions are not equal to spoken language—they are more abstract (Spoehr & Lehmkuhle, 1982, p. 238). Propositional representations in a computer, for instance, use a notation similar to “On (Ball, Box),” as opposed to “the ball is on a box.” “On (Box)” is syntactically illogical because “on” defines a relation and so requires two objects (Kosslyn et al., 2006, p. 10). There are semantic limitations as well.
Propositions have truth-value: they are either true or false. Stored information usually takes the form of true propositions. Propositions operate according to set rules. They make assertions about entity relations, which are checked for truth-value using propositional rule structures (Spoehr & Lehmkuhle, 1982, p. 238). The requirement of truth-value for propositions themselves is a point of debate (Kosslyn et al., 2006, p. 11). Propositions can refer to classes as well as abstract entities, like sentimentality. Finally, they are not modality-specific, meaning that they can store types of information other than language, such as sight and touch (ibid.). The semantics of propositional encoding is arbitrary.

Proponents of sole propositional representation argue that the experience of imagery is in fact an illusion of a propositional network. Others propose an imaginal format, or mental imagery. Mental imagery is somewhat similar to pictures. It is not abstract but must be formed via resemblance in a particular modality (such as visual). Any “nonpicturable” concept can only be depicted indirectly, through association. There is spatial correspondence, such that represented distances must correspond to perceived distances of an object from a certain point of view (Kosslyn et al., 2006, pp. 11–14). An experiment by Tootell et al. (cited in Kosslyn et al., 2006) demonstrated that visual stimuli are represented topographically on the occipital cortex, strongly suggesting that, whatever the status of propositional representation, imaginal representation is very real (pp. 16–17).

2.4.2 Integration

Allan Paivio (1986) considers a proposition-only approach to be rationalist to a fault. His dual-coding theory (Paivio, 1986; Sadoski & Paivio, 2001) is based on the empirical evidence for propositional as well as imaginal formats, and suggests that the two operate together. If there are two representational systems at play, there must be an underlying “interlingua” for translating one to another (Spoehr & Lehmkuhle, 1982, pp. 239–240). The Stroop effect (pp. 242–245) attests to this. When Stroop asked subjects to identify which color words of conflicting color names were printed in (the word “red” printed in blue ink, for instance), a demonstrable interference occurred. Subjects couldn’t help but read the color
name, even when it hindered their naming of the ink color. If the visual system (represented by mental imagery) and the verbal system (represented propositionally) operated independently, then the modes should not interfere with one another. But they do.

Dual-coding theory emerged out of the study of mnemonic techniques that utilize imagery, such as the method of loci (covered in the section on memory) and the pegword method (Sadoski & Paivio, 2001, pp. 3–4). The pegword method is based on a standard rhymed verse that pairs numbers to nouns (beginning with “one is a bun, two is a shoe”). Once the rhyme is learned, its numbered lines serve as pegs onto which information (in the form of objects) is mentally “hung.” For a shopping list, hot dogs might be imaged in a bun, and a cat might be imaged eating out of a shoe (representing cat food) (Spoehr & Lehmkuhle, 1982, pp. 206–207). Though this essentially increases the amount of information considered, it works by forming referential connections, as well as taking advantage of already memorized information (the number rhyme). The correlation between increased information and greater memorability is surprising if one is working with a metaphor of mind-as-computer. But strength lies in interconnectivity.

Dual-coding theory suggests that mental representations retain holistic properties based in sensory modalities. Verbal modalities are derived from speech and writing, nonverbal modalities from nonlinguistic sight and other senses. While mental representations retain their holistic properties, they are also associated with other representations. These associations are critical to the theory, as is the recognition of the functional significance of the two representational formats—other associationist theories don’t stress format differences (Sadoski & Paivio, 2006, pp. 4–5).

### 2.4.3 Overview of Dual-Coding Theory

Sadoski and Paivio (2006) provide a brief overview of dual-coding theory (DCT) at times specific to reading and writing (pp. 42–66). Sensory information from the external world is captured as either verbal or nonverbal representations. This process is coding. Processing occurs when representations are activated within or between the dual systems.
The nonverbal coding system is referred to as the imagery system even though its representations are not always visually derived.

Within the two codes are multiple modalities. Each modality is based in sensory experience. This is a key proposition of dual-coding theory, that internal representations are not abstract but are rather corollaries of sensory data. Mental representations thus have a concrete quality and are considered continuous with perception and memory, or they are modality-specific. Verbal coding can take place in at least three sense modalities: the written word (visual), utterances (auditory), and Braille and handwriting (haptic). Nonverbal coding is more general to experience of the world and represents all senses (Table 2.4.1).

Table 2.4.1: DCT codes and modes (copied from Sadoski & Paivio, 2006, p. 45)

<table>
<thead>
<tr>
<th>Sense modality</th>
<th>Verbal code</th>
<th>Nonverbal code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual</td>
<td>Writing</td>
<td>Visual objects</td>
</tr>
<tr>
<td>Auditory</td>
<td>Speech</td>
<td>Environmental sounds</td>
</tr>
<tr>
<td>Haptic</td>
<td>Braille, handwriting</td>
<td>Feel of objects</td>
</tr>
<tr>
<td>Gustatory</td>
<td>none</td>
<td>Taste memories</td>
</tr>
<tr>
<td>Olfactory</td>
<td>none</td>
<td>Smell memories</td>
</tr>
</tbody>
</table>

Mental representations are made up of nonarbitrary, concrete, modality-specific building blocks. The basic unit of verbal representation is the logogen (meaning language generator). Logogens are units of various sizes. A letter, a syllable, a word, and a familiar expression are all logogens. The basic unit of nonverbal representation is the imagen (meaning image generator). Imagens are nested, such as with a face containing eyes and a body containing the face. There are scale limitations for logogens and imagens, reflecting their concrete nature. A long word cannot be “seen” all at once mentally; it must be scanned to be “read.” A mental image has limits as to what or how much can be salient at any one time, with similar scanning requirements for “viewing.” Medieval mnemonic practitioners
accounted for the limited visual field apparent in the manipulation of mental imagery (Carruthers, 2008). The relative depicted size of memory objects was known to affect memory trace.

Logogens are arranged serially, as language is temporally situated and so serial itself. It is more difficult to spell a long word backwards than forwards because of the directional constraints of language use. In order to access a nested logogen it is often necessary to run a nesting sequence, as when the child must return to the beginning of an aborted piano piece rather than continue where he left off. Imagens are synchronous, such that all information is available at any time. An imaged room or a visualized process can be traversed mentally through any path without the constraints on a familiar order in the experience of logogens. The “visual field” puts limitations on access to imagens, meaning that although any information is eminently available, all information cannot be processed at once. The spaces involved must be traversed. Imagens are nested in the most conventional sense, where logogens are more embedded in sequence.

Figure 2.4.1 summarizes the connectivity assumed in dual-coding theory. All meaning resides in the various connections proposed by the theory. The verbal and nonverbal systems can operate independently (one at a time), in parallel (independent of one another), or in concert. Both logogens and imagens exist within hierarchical arrangements. But, again, they differ in arrangement constraints, with logogens arranged sequentially and imagens arranged through overlapping and true nesting. Some (but not all) logogens are connected referentially to imagens, and necessarily vice versa. (Highly abstract “function” words such as the and of have no referential connections as they are pure language units; likewise, some imagens are “nameless.”) Logogen associations are created when different logogens are commonly used together, or simply by a detected similarity (p. 55). Imagens usually form in context as part of a hierarchical environment (p. 56). Connections can be cross-modality as well as cross-code.

The activation of mental representations and the connections between them is processing. There are three basic levels of processing: representational processing, referential
processing, and associative processing (p. 58). Higher order processing is organizational or transformational.

Representational processing involves the activation of mental representations by sensory information. The activation of any given representation is probabilistic, based on both the stimulus situation, which includes stimuli characteristics and context, and individual differences, which includes personal experience in interaction with cognitive styles and limitations (p. 58).

Referential processing involves the activation of representations in one system by previously activated representations in the other (p. 59). It is thus indirect—representational processing of the picture of a cup will activate related imagens, which in turn may referentially activate the logogen of the name “cup.” Referential connections are one-to-many in both directions, meaning that one imagen is often connected with multiple logogens, and vice versa (ibid.). In reading, mental imagery “is more likely to be evoked by concrete language than abstract language” in text, and all the more likely to be evoked by pictures (ibid.).

Associative processing occurs when one representation activates others in the same system. Nonverbal associations are part of “mental sets.” These are spatial and/or temporal.

Concepts, episodes and domains are higher order mental structures formed by organizational and transformational processing. The sequential constraints in the verbal system continue to operate at higher levels. Story grammars and textual cohesion are higher order verbal organizations. In the nonverbal system events can be organized into compound images that act as holistic representations. Imagery transformation involves mental shifting of perspective or mental rotation of imagens. Mental imagery serves like a functional analog of sensory experience, and many of the constraints and affordances of observing the world are reflected in the experience of mental imagery.

The structure of the dual systems and the mental processes involved “serve important mediating roles in the performance of literary tasks and the guidance of literacy behaviors” (p. 63). There is a mnemonic function implied by dual-coding theory because the two codes are thought of as independent and additive. Encoding information both verbally and
nonverbally results in better memory. Concrete language is more memorable than abstract language. The conceptual peg hypothesis conceptualizes imagery as a peg onto which related information can be hung (ibid.). Evaluative and monitoring functions are at play whenever a reader questions his comprehension through internal dialogue. Spatial reasoning is often employed to solve problems. Imagery is often used to track the events of a story. The experience of imagery while reading can serve emotional and motivational functions. Concrete texts, again, promote imagery. Abstract terms can be affective though, as in the case of beauty or honesty (p. 64). Verbal and nonverbal processes also have a creative function. Verbal creativity involves novel associations within an accepted framework. Imagery is used to simulate manipulations of the world and to create new worlds.

Figure 2.4.1: DCT model (copied from Sasoski & Paivio, 2006, p. 53)
2.4.4 Discussion

The designed page often includes text and image in some kind of arrangement. Even in cases where no images are provided, concrete text promotes mental imagery. Dual-coding theory is first and foremost about connections between and within codes and modalities. An important claim is that verbal and nonverbal codes are additive, resulting in a stronger memory trace when operating in concert. Furthermore, mental processing (beyond and including remembering) is positively affected by interconnectivity. A body of literature on textbook illustration has emerged in part from these central dual-coding theory claims.

2.5 Textbook Illustration

2.5.1 Classifying Illustration

Many forms of ill-defined text—picture structures are referred to in the textbook illustration literature with inconsistent terminology. Newton (1993) provides a formal classification system (citing a dissertation by P. J. Roe) that includes: *representational drawings, photographs, exploded diagrams, symbolic diagrams, tables and charts, and graphs*. However, it is not finely detailed.

Vekiri (2002) provides a taxonomy of graphics (p. 265) that is brief but detailed. Vekiri identifies two graphic types with non-arbitrary symbol systems—diagrams and maps—and two with arbitrary symbol systems—graphs and charts. Diagrams show “parts, structure and operation of real objects or abstract entities; process”—they are schematic and often utilize arrows and lines (ibid.). Maps are more direct representations of a “real territory,” with relative isomorphism. Graphs (arbitrary symbol systems) present quantitative data in such a way that readers can actively compare “relations among variables.” Charts, including tree diagrams, concept maps and matrices, show relations among concepts or a
sequence of events. Hybrid forms, such as charts with non-arbitrary symbols, are not discussed.

Other authors in the literature do not necessarily subscribe to Vekiri’s definitions. In an attempt to defer to the researchers’ determinations, varying terms will be employed throughout the follow subsections as they “travel” with the citations. It is hoped that terms such as map, though not explicitly defined in many cases, are sufficient nonetheless.

### 2.5.2 Reader Characteristics and Efficacy

Learning from illustrations is contingent upon characteristics of learners (such as visuospatial ability, reading ability, reasoning ability, and content knowledge), characteristics of learning materials (such as text difficulty and illustration qualities), and the fit to the targeted learning activities themselves (Hannus & Hyönä, 1999, p. 97; Vekiri, 2002, p. 275). The study of high-ability readers implies some effective reading strategies. It is good to: reread text; study pertinent rather than less pertinent details (which means critical differentiation is at play); and move back and forth between text and relevant illustration with higher frequency. In other words, high-ability reading entails “concentration on pertinent information and on integrative processing” (Hannus & Hyönä, 1999, p. 118). The demands heavily illustrated text puts on a reader are considerable. Study time equates with success.

One eye movement study found that readers make local diagram inspections while reading (shifting from text to accompanying displays) and then make global inspections afterwards, aimed at combining the components into a whole (Vekiri, 2002, p. 277). Low-knowledge readers have difficulty determining relevant information; they need prompts (p. 278). Diagrams are difficult (and less effective) for students with low visuospatial ability (p. 279). Good reading strategy can be facilitated by “breaking down the information in multiple displays and by using cues (such as arrows or descriptors embedded in the display) and labels that direct readers to the parts of the display that are important” (p. 280). Text illustrations must be related to the text and must seem to be related to the text in order to have an effect.
(Levin & Mayer, 1993, pp. 105–106). There must be cues to guide the reader from text to illustration. In order to keep intent and perception in sync, captions should support intended interpretations (Pettersson, 1998, p. 6).

Though pictures facilitate learning in many ways (including motivation), readers might view them as “easy” material and attend only superficially (Carney & Levin, 2002, p. 10). Since so many pictures in textbooks are decorative, readers experience “image overload.” “Regardless of the intended functions pictures are not always used in an active way at school. On the contrary pictures in textbooks are often ignored and ‘skipped.’ Most students do not attend to the visuals unless they are instructed to do so” (Pettersson, 1998, p. 6)—thus the importance of providing cues.

Pettersson (1998) speculates that interest—clearly not a given—might increase with a mixture of different types of visuals (p. 7).

Teachers should be instructed through a guide in the use or purpose of provided visuals so that they can facilitate high-ability reading strategies and model a productive experience. Pettersson provides a list of the experiential requirements of readers for illustration efficacy. Readers must:

- “See or rather ‘discover’ the picture.
- Pay attention to the picture.
- Read the picture in an active and selective way.
- Process the information mentally.” (p. 7)

Hannus and Hyönä (1999) provide an explanation as to what reading “in an active and selective way” might entail. Successful comprehension requires that the reader:

- Comprehend abstract concepts in text and illustration.
- Decide upon a reading sequence.
- Distinguish between pertinent and less pertinent information within text and illustrations.
• Determine the relationships between different pieces of information and how they “converge” on a central concept or principle.
• Integrate related pieces of information into a coherent internal representation.

(p. 98)

2.5.3  Text–Image Integration

“In a generative theory of textbook design, learning is viewed as a constructive process in which learners select and build cognitive connections among pieces of knowledge” (Mayer et al., 1995, p. 32). Integrating happens between verbally- and visually-based models—Figure 2.5.1, based on Paivio’s dual-coding theory—and forms referential connections. Critical to the central idea is that this integration must happen in working memory, with its notable limitations on load. Mayer et al.’s (1995) major proposition is that when illustration and text are more integrated on the page, it is easier for the reader to integrate them mentally (p. 33)—this is the contiguity principle (Vekiri, 2002, p. 275). The contiguity principle states: “in order to minimize the cognitive load associated with mental integration of information, new material should be provided in different modalities and coordinated in space and time” (ibid.). Too much separation of illustration and text requires the reader to hold one component in working memory while attending to the other—and it is more difficult to hold textual information in working memory (pp. 276, 295). Learning is better with auditory verbal information than text, because the latter requires the visual system in processing, and this mixes modalities. (Visual and auditory information in concert doesn’t overload working memory, while it provides more connections.) Integration reduces the visual demands of text by limiting the need for short-term retention. “Short explanatory texts… are more effective for the retention and transfer of explanatory information than are longer texts but only when coordinated with visual information” (p. 272). Processing demands are decreased when different kinds of representation are integrated into a single representation, as in text embedded in graphical displays (p. 303). Coding simultaneously in both representational formats (propositional and pictorial) provides the reader with richer
detail (Hannus & Hyönä, 1999, p. 96). There are indications that the one representational
code can be co-opted for the other’s use (Vekiri, 2002, p. 269). Integration of text and image
is effective, but needlessly redundant information is overly taxing, and “segments that can be
understood in isolation” should be presented as such (p. 277).

Figure 2.5.1: Visual–verbal integration (copied from Mayer et al., 1995, p. 32)

Utilizing two codes in instructional material increases retention because visuals
increase concreteness and they lead to better generation of mental imagery (Vekiri, 2002, p.
267). Mental imagery manipulation (rotating, zooming, etc.) has been found equivalent to
external image manipulation (p. 268). The brain appears to manipulate external and mental
images with the same hardware—brain-damaged patients with deficiencies in the left visual
field reported corresponding difficulty with mental imagery (p. 269).

Hannus and Hyönä (1999) criticize much of the research literature on textbook
illustration as inauthentic, because experiments often present a text passage with a single
illustration, where authentic textbook materials utilize multiple illustrations (they use
authentic materials in their studies) (p. 97). Authentic materials require the reader to make
constant decisions regarding engagement. “Illustrated textbook materials can be regarded as
highly complex stimulus environments for several reasons,” including integration and
synthesis activities, determination of sequence (reading strategy), and the difficulty inherent
in interpreting visual material (such as diagrams) (p. 98). The reader must attend to the
relevant components of a picture and cross-reference them with separated textual content. Hannus and Hyönä report a remarkably confused body of literature that makes conflicting claims as to whether frequent or infrequent shifting from text to image is more successful. Their “authentic” tests indicate that illustrations improve subjects’ answers to detailed questions, but only high-ability subjects’ answers to comprehension questions (p. 107). They speculate that comprehension questions are more demanding of integration, and only high-ability students are capable of such integration. The researchers found no evidence that illustrations distract or otherwise stymie low-ability students. However, the same study found a “discouragingly” small effect size for illustrations in general.

A follow-up eye movement study revealed that high-ability students spend slightly longer with text, but read faster—meaning they devote more time to rereading (p. 115). (There were no significant differences between groups regarding attention to illustrations or captions in time spent or fixation frequency.) Low-ability students spend more time gazing at irrelevant or blank spaces, which is possibly an indication of distraction. Only 6% of study time was devoted to illustrations, despite the heavy use of illustrations in the materials—twice as much time was spent on the captions (the importance of cues will be addressed shortly). While this may mean students don’t attend to illustrations, it could also mean that use of imagery is highly efficient, and less demanding of time. When the analysis differentiated between pertinent and less pertinent segments of illustrations or text passages, high-ability students showed a significantly higher level of discrimination. Surprisingly, memory for illustrations correlated more highly with time spent on text than time spent on illustration, leading the researchers to conclude that “learning [is] heavily text-based” (pp. 116–118).

Mayer et al. (1995) reported a greater effectiveness of integrated designs for low-experience learners only. No difference was found for high-experience learners.

Illustrations are more effective when explicit instructions (cues) for engaging are given, since it appears that text drives reading strategy (Duffy, 1992; Hannus & Hyönä, 1999; Carney & Levin, 2002). Directives for reading images have variable results. In order to improve learning, “illustrations need to be directly relevant to some aspect of the text
content” rather than arbitrary or isolated (Hannus & Hyönlä, 1999, p. 97)—that is, the relationship between text and picture should be meaningful.

Mayer et al. (1995) attempt to parse out caption effects from annotation presence, to distance from annotation, to illustration. Annotation proved more effective than proximity, with the strongest result from combination. Proximity without annotation was not different from non-annotated–distant. However, the researchers stress that it is inappropriate to conclude that distance is not important (p. 40). Mayer et al. conclude that understanding of scientific text and problem-solving transfer improves with “multi-frame illustrations” for cause–effect systems. Though cause–effect systems are predominant in science, such illustrations are uncommon in textbooks (though the study was conducted in 1995). Even “modest” annotative adjustments to current textbook illustrations could improve comprehension (p. 39). In short, the dual-coding model (Figure 2.4.1) should be utilized.

2.5.4 Text–Image Integration in Action

Carroll et al. (1992) tracked eye movement through single-frame cartoons (The Far Side), which are predicated on incongruity between caption and image and a resulting “correct” interpretation via conceptual leap (pp. 446–447). An initial anecdotal study found variation in the order in which readers consult the modal components (picture and caption). However, reading in the two modes proved to be relatively isolated events: readers did not repeatedly shift back-and-forth from picture to caption. Partial inspection of the picture often precedes caption reading. The caption is pored over and engagement includes regressions—returning to a point already scanned—that are separate from picture viewing. Carroll et al. created experimental conditions where subjects were forced to view captions and pictures in isolation, in both possible orders.

Readers very quickly assess the gist of a scene based on its spatial characteristics, presumably limited by their visual processing capabilities. In the early moments of engagement with a picture, readers seek out a schema that is consistent with the picture (p. 453). Even when readers first read a caption, this schema-seeking stage persists—seemingly
isolated from whatever the caption may suggest. As this initial stage ends (at around 6–7 fixations in the cartoon study), a second stage begins where the reader attempts to resolve any incongruities between picture and caption. In terms of individual modal performance, forming a memory representation of a picture is quicker and probably easier than of text (p. 455).

Carroll et al. equate “getting the joke” of a cartoon with solving a problem or puzzle. The relative simplicity of the picture and its caption require the reader to mentally complete the scenario—to employ imagination.

2.5.5 Picture Functions

Numerous picture function typologies exist, each reflective of a unique purpose. Levin and Mayer (1993) state that illustrations can have cognitive, aesthetic, attitudinal and social outcomes. Hannus and Hyönen (1999) provide a similar list of general functions of illustrations: attention guiding, affective, cognitive, and compensatory functions. Motivation falls under the affective function. Compensatory function refers to the assumption (unresolved in the literature) that low-ability students benefit more from illustrated text (as scaffolding) (p. 95). The compensatory function appears to stress the relationship of text to illustration. Regarding that relationship, pictures can make text more: concentrated, compact/concise, concrete, coherent, comprehensible, correspondent, and codable (Carney & Levin, 2002, p. 9; Levin & Mayer, 1993). These “seven C’s of picture facilitation” are aligned with a more detailed and complete list of picture functions detailed by Levin and collaborators (Levin, 1979; Levin & Mayer, 1993; Carney & Levin, 2002). The “seven C’s” will be discussed along with the “prose pictures” function types, which are, in order of increasing recall benefits (Levin, 1979) as well as increasing text-learning effect (Carney & Levin, 2002; who excluded the remunerative, motivational and reiterative functions and found decorative to have no benefit):
1) Decorative: “Pictures increase a text’s attractiveness [only].”
2) Remunerative: “Pictures increase publisher’s sales.”
3) Motivational: “Pictures increase children’s interest in the text.”
4) Reiterative: “Pictures provide additional exposures of the text.”
5) Representational: “Pictures make the text information more concrete.”
6) Organizational: “Pictures make the text information more integrated.”
7) Interpretational: “Pictures make the text information more comprehensible.”
8) Transformational: “Pictures make the text information more memorable.”

(Levin, 1979, p. 14)

Additional selection and integration functions are identified and an efficiency function implied as well (Levin & Mayer, 1993) and will now be discussed along with the illustration efficacy (seven C’s) and picture function types.

Decorative, remunerative and motivational functions are treated as chimeras—they don’t function, per se. Pettersson (1998) is clearly no fan of decorative pictures: “Cognitive and decorative functions should never be confused or mixed” (p. 7). Remunerative is clearly confounded with decorative, save that it specifically represents the publisher’s agenda. Levin (1979) states that the literature does not well support a motivation function to pictures—in absence of content-appropriate characteristics—that results in better recall (p. 16). The qualification here is that there is not a lone motivation-to-recall effect supported by the literature, but this doesn’t mean that motivation isn’t a contributing factor in many possible cases. (Note that recall is often, but not always, the object of interest in Levin et al.’s work—which belies a narrow consideration of function at times.)

Regarding reiteration, Levin (1979) concludes that repetition of text does improve recall, but only when recall is cued by verbatim (to the text source) questions. Pictures facilitated paraphrased questions as well as verbatim questions, and so the benefits of reiterative pictures stem not merely from a repetition function but include other effects (p. 17). Woodward (1993b) defines textbook illustration functions by illustration content relevance, which ties into reiteration as well as others of Levin et al.’s functions. Woodward
identifies non-content-related visuals (irrelevant, or decorative); tangentially content-related visuals (connected to broad subject matter only); content-supporting visuals (repetitive, or reiterative); and content-extending visual (additional information that enhances and extends, which might align with some of Levin et al.’s types) (p. 128). Woodward surveyed textbooks and found most visuals to be content supporting, with almost no examples of content extending. In order to support or extend content, Woodward concludes that captions must be studied along with illustrations. He considers five possibilities for picture captions: no caption whatsoever; title and name only; repetition of main text; extension of main text (providing new information); and question to the reader (pp. 125–126). In some cases title and name only can be very effective in drawing connections to the text, as in when an unfamiliar object is pictured, though titles are often superfluous. Paraphrased or repeated text in captions draws the strongest connections. Woodward’s survey of two textbook series found titles (29% and 81%) and repetition (39% and 19%) to be the primary caption method, accounting for 68% and 100% of instances. Only 19% and 13% of captions in one series asked questions of the reader or provided further information, respectively. In the other series no captions did either (pp. 126–127). (As with other results, these figures should be tempered by the knowledge that the study was published in 1993 and the textbooks, all the more dated, were themselves published in 1985.)

None of the functions described up to this point are aligned with the illustration efficacy types. They will thus be considered low functioning types. The remainder of the functions will be subsumed within the efficacy types (picture efficacy is reviewed in Table 2.5.1).

Pictures can make text more concentrated. This means of facilitation is aligned with a selection function of illustrations (Levin & Mayer, 1993; Mayer, 1989). A selective reader is able to discriminate between critical and non-critical textual information. Illustrations can facilitate this proficiency when somehow foregrounding the “essence” of a passage for purposes of recall and comprehension (Levin & Mayer, 1993, p. 98).

Pictures can make text more compact/concise. In many cases, information is inherently spatial and a text-based explanation is necessarily ponderous. In such cases,
illustration performs an *efficiency* function for the reader. Levin and Mayer (1993) use an illustration of a three-pulley, two-weight system as an example, and as a counter-example provide a textual description of the system, which though *complete*, is far from compact or concise (pp. 98–99).

*Pictures can make text more concrete.* The *representation* function refers to a picture’s ability to render textual information more *specific* as well as its provision of “a second modality through which the text can be directly represented in the brain,” being dual-coded (Levin, 1979, p. 18). Levin characterizes pictures in this light as leaving a “memory trace” superior to verbal representation alone. Carney and Levin (2002) found that narratives accompanied by representational pictures promote detail recall better than those without (p. 11).

*Pictures can make text more coherent.* Coherence is similar to *compact/conciseness*. Illustrative organization potentially instills a text with “internal connections.” Thus coherence efficacy is aligned with *organizational* imagery (Levin & Mayer, 1993, p. 99). The organization function is most appropriate when the accompanying prose passage is not itself well structured (Levin, 1979). In such a case the associated pictures serve as a surrogate structure for the text itself. Low-ability students are particularly aided by organizational imagery because otherwise they are likely to encode information in “fragmentary bits and pieces” (p. 19).

*Pictures can make text more comprehensible.* Some texts, particularly science texts, can be unintelligible without a great deal of expertise in a particular area. When a reader does not possess requisite knowledge for productive interpretation, complexity can be managed and clarity enhanced by accompanying illustrations. The *interpretation* function makes text more comprehensible in such situations (Levin & Mayer, 1993, pp. 99–100). Without a difficult text, the interpretation function is likely minimally effective (Levin, 1979, pp. 20–21).

*Pictures can make text more correspondent.* Unfamiliar concepts can sometimes be understood through analogous familiar concepts. Correspondence is effectively metaphor application: the structure of the familiar (the eye, for example) is mapped onto the unfamiliar
(the aperture of a camera). This involves building direct connections, and as such is considered to have an integration function (Levin & Mayer, 1993, pp. 100–101).

Pictures can make text more codable. Codability here refers to representational (propositional and depictive) codes and the transformation of one into the other. This illustration efficacy then is aligned with the transformation function as well as mnemonics. Mnemonic illustrations can improve not only memory for facts, but also reasoning ability and application (Levin & Mayer, 1993, pp. 101–103). The transformation function is not necessarily associated with difficult-to-comprehend material, but rather difficult-to-memorize material, such as historical passages (which often come down to a parade of names, dates, etc.) and medical texts (replete with specialized terminology) (Levin, 1979, pp. 21–22).

Table 2.5.1: Picture efficacy (Levin & Mayer, 1993, pp. 98–103)

<table>
<thead>
<tr>
<th>Pictures make text more</th>
<th>By means of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concentrated...</td>
<td>Selection</td>
</tr>
<tr>
<td>Compact/concise...</td>
<td>Efficiency</td>
</tr>
<tr>
<td>Concrete...</td>
<td>Representation</td>
</tr>
<tr>
<td>Coherent...</td>
<td>Organization</td>
</tr>
<tr>
<td>Comprehensible...</td>
<td>Interpretation</td>
</tr>
<tr>
<td>Correspondent...</td>
<td>Integration</td>
</tr>
<tr>
<td>CODable...</td>
<td>Transformation</td>
</tr>
</tbody>
</table>

Levin and Mayer (1993) note an eighth possible illustration efficacy type: collective recollection. Collective recollection refers to culturally pervasive symbols. The benefit of this “legacy” is not explained (p. 103). Arguably, collective recollection is what makes visual communication possible without text, and so this efficacy type is broad and diffuse.
2.5.6 External Cognition

Graphical representations are effective because they are visuospatial and require “fewer cognitive transformations” than reading text. The co-presentation of elements and their relations in graphical displays (perceptual enhancement) makes it easier for readers to “perceive or draw inferences about these relations” than text (Vekiri, 2002, p. 281). Norman (1993) refers to such graphical displays as “cognitive artifacts”. Searching through text is demanding of working memory because the arrangement (sequence) is not conducive to searching, while information arranged spatially acts as a kind of extended mental workspace relieving working memory. This amounts to a “computational advantage” for graphical displays—readers are able to utilize perceptual mechanisms rather than relying on “interpretation processes” (Vekiri, 2002, pp. 281–282). It is important that information be presented in “visual chunks” so pattern perception processes are used rather than “complex data transformations” (p. 289). Visual displays function as “external cognition” by guiding cognitive behavior. There is directness in being able to think through the images on a page rather than having to maintain representations in working memory (p. 282).

As always, the affordances of a graphic display must align with instructional goals. No design strategy is good for all applications, so it’s important to note the strengths of different display types. “Parts-and-steps” pictures, for instance, improve both recall and problem solving by providing “runnable mental models” through which students might reason (Carney & Levin, 2002, pp. 13–14). “Explanative” pictures can best function as mental models when cause-and-effect systems or complex processes are the objects of instruction (p. 21).

When maps are encoded as intact, holistic units they preserve their visuospatial properties: they include information about individual features as well as structural information about relationships (Vekiri, 2002, p. 292). There is thus a kind of isomorphism (a second-order isomorphism) between structural depictions on the page and mental representations, which seems to carry implications for efficiency. However, if the structural information in maps is insufficient, advantages disappear as working memory is overtaxed.
Furthermore, when using a map in relation with a text, reading order holds influence. When a map is studied \textit{before} accompanying text, readers have higher recall, make more inferences, and are able to reconstruct the map more accurately than if the order is reversed (Vekiri, 2002, p. 295). The maps-first results come about because the mental imagery of the map is activated in working memory \textit{during} reading of the text. The opposite does not hold true: holding the text in working memory is much more demanding because it must be processed serially. Maps are most effective when they minimize the processing necessary to read them (pp. 298–299). (This “thinking through imagery” is a principle behind medieval mnemonic practice.) Maps are more beneficial when students are familiar with them (p. 296). (This too is reminiscent of mnemonics, particularly techniques such as the method of loci and the pegword method, which utilize familiar spaces or other familiar constructs to facilitate working through novel information.) Student recall of information is contingent upon how the structure of study-aiding maps is encoded. Nonintegrated maps—where related text is not explicitly connected to visuals—are ineffective (p. 293).

\textbf{2.5.7 \textit{Higher-Order Objectives}}

Much of the illustration literature appears to focus on recall, which is considered to be “lower-order thinking.” There are, however, many cases where higher-order objectives are addressed.

The concept of “higher-order thinking” is itself problematic. Anderson (1993) makes an analogous but better-defined division of contextually rich and contextually limited knowledge. Contextually rich knowledge occurs when a student understands how, when and why to use knowledge. Contextually limited knowledge occurs when a student is taught some task in isolation and then might only apply that knowledge in a starkly similar situation. (There isn’t transfer.) Contextually rich knowledge is encoded through multiple conceptual relationships, meaning that there are more “paths along which reasoning may travel from idea to idea” (pp. 138–139). Otherwise there are fewer routes and it becomes less likely that a student will utilize the knowledge in a novel situation.
Auxiliary materials—end-of-chapter activities and other supporting materials that accompany textbooks—too often promote contextually limited knowledge by virtue of conceptual incoherence. Auxiliary textbook materials tend to focus on controlled release of content, rather than “the ways that learners are making sense of the content they are encountering” (p. 140). A task that requires students to select a correct option from a list builds only contextually limited knowledge because: it reinforces that the activity is tied to instructional materials; there is one correct answer; and the correct answer is given rather than constructed by the student (p. 141). Such tasks fail to promote true comprehension or synthesis—use of the details to create new knowledge structures. Worse yet, they train students in bad practices: “far too many students learn from science textbooks that answers to science questions can be found by attending to the words in boldface type or by memorizing definitions” (p. 145). Rich contextualization occurs only when ideas are encountered in different situations and with varying application. Multiple representations should be used. Tasks should be situated within a social setting that facilitates dialogue—that is, the classroom.

In order for students to successfully build contextually rich knowledge, teachers or learning materials should provide scaffolding. Scaffolding is anything external that purposefully assists a student in moving from a beginning state to higher understanding. To accomplish this, the activity in question must operate in the “zone of proximal development”, where tasks are challenging but not prohibitive. If a task is too difficult, a student will likely fail to link new information to prior knowledge, despite teacher assistance (p. 151). To make scaffolded auxiliary materials the required conceptual support must be accurately identified—that is, productive behavior must be modeled. Some specific strategies (or design forms) for achieving contextually rich or higher-order thinking have been discussed in the textbook illustration literature.

Since mnemonics often involves schema manipulation, its practice extends beyond mere rote memorization. Use of a mnemonic taxonomy (as content organization) for a botany lesson improved both information reconstruction and application performance (Carney & Levin, 2002, pp. 17–18). Mnemonic strategies make information more accessible, aiding in
“acquisition of high order concepts and skills” (p. 20). College students using mnemonic study guides (versus students in “free study”) exhibited better performance not only with “factual memory” but also the higher-level task of essay writing (ibid.).

Pictorial models and analogies improve conceptual understanding and problem solving but not verbatim memory (Levin & Mayer, 1993, p. 105). In this sense purposeful illustrations provide a way to ensure that rote recall doesn’t pay off as a strategy.

Vekiri (2002) evaluates and provides guidelines for various graphic formats. Graphic organizers help students make complex inferences, integrate provided information, and identify patterns among concepts. They also encourage non-memorization strategies when used for review. Important relations in graphic organizers must be salient (p. 287). Salience is critical as graphics help with complex interpretation tasks because they allow the reader to avoid complex cognitive tasks—not the case if the graphics demand too much difficult processing (p. 289). Thus, use of gestalt principles produces more effective graphic organizers. Too much complexity in a knowledge map hinders learning (p. 288). (Though hierarchy manipulation can help to manage complexity.) Ultimately, graphical displays “are effective when they address the limitations of working memory” (pp. 302–303).

2.5.8 When to Use Illustrations

Illustrations should not be used in every situation. With learning to read, illustrations are obstructive. With reading to learn, illustrations are beneficial (Levin & Mayer, 1993, pp. 96–97). A study by Newton (1983) found that readability is higher for text accompanied by illustration than without. Factors that may determine illustration’s effect on readability are: “the degree of integration of illustrations with text” (effectively confirmed in other studies and mentioned above); “the style of illustrations”; and “the density of illustrations” (though density is not clearly defined) (p. 50).

Desired performance outcomes, nature of illustrations, nature of text, and learner characteristics determine picture facilitation (Carney & Levin, 2002, pp. 9–10). The former consideration is critical—cognitive outcomes should drive illustration selection and align
with intended function. Levin and Mayer (1993) list potential outcomes as: understanding or comprehension (especially when the text is difficult); remembering; and applying (p. 104). For *understanding*, concise representations or comprehensible pictorial models and analogies are most effective. For *remembering*, use connected “pictorial organizations” and recoded mnemonic illustrations. For *applying*, use “illustrations that foster factual and conceptual transfer,” which has worked with model, analogical and mnemonic illustrations. Illustrations should function as desired rather than merely sell textbooks. Illustration needs should take into account text difficulty and other qualities. Positive effects of illustrations are observed if the text isn’t so easy as to render facilitation unnecessary, and the text isn’t so difficult as to render the endeavor too daunting overall (Newton, 1983, pp. 50–51). “Concrete and engaging” passages will not benefit from accompanying pictures, as such text already promotes image use in the reader (Carney & Levin, 2002, p. 20). Illustrations are beneficial for complex, unfamiliar and explanatory texts. Illustration benefits are greatest for content-underexposed and less proficient students. However, students must possess adequate interpretative skills to reap any benefits (Levin & Mayer, 1993, pp. 109–110).

Levin and Mayer (1993) offer guidelines for some specific illustration types. Pictorial maps improve recall of picture-relevant information, but not picture-irrelevant information. Pictorial models are more beneficial for text information that can be directly pictured. Pictorial analogies are more beneficial for text information that cannot be directly pictured. Mnemonic illustrations improve associative but not non-associative memory (p. 105).

Pettersson (1998), citing Marsh (1983), offers eight general guidelines for “selecting a visual channel for a message”:

- When messages are complex.
- When referability is important.
- When messages are long.
- When environment is noisy.
- When arrangement is complicated.
• When precise spatial discrimination is important.
• When simultaneous presentation is desired.
• When more dimensions are required.” (p. 8)

Ultimately, careful consideration is required when determining illustration use. There are indications that picture selection in the textbook industry is largely a result of production limitations, and not theoretically grounded image function considerations (Woodward, 1993b; Pettersson, 1998).

2.5.9 Textbook Publishing Industry Considerations

Textbooks “attempt to distill and interpret the knowledge and scholarship of a discipline” (Woodward, 1993b, p. 115). This is a tall order. According to Woodward, market pressures and production issues result in textbooks that tend: to be overly simplistic; to be choppy in prose; to be guided by marketing rather than instructional purposes; and to cover content superficially. Practice isn’t guided by research (p. 116).

Interviews with those in the Swedish textbook publishing industry revealed intent regarding the use of pictures: pictures are used “only” to stimulate the reader; they have a “life of their own”; they provide “breathing space” (Pettersson, 1998, p. 6). Pettersson lists the questions that textbook publishers (editors, art directors and designers) ask themselves when selecting images as:

• “Does the picture depict the right thing?
• Is the presentation of the subject satisfactory?
• Is the picture technically acceptable?
• Is the picture aesthetically satisfactory?
• Is the picture ‘flexible,’ i.e., will it work with different formats?
• Will the picture fit into a given area?
• Will the picture fit in with the other pictures on the same page?” (p. 7)
None of these questions addresses picture function or instructional objectives. They are solely guided by the considerable production limitations of textbook publishing.

2.5.10 Discussion

Vekiri (2002) criticizes much of the literature on textbook illustration as overly simplistic, and doubts that many of the results can be generalized to authentic materials and situations (p. 270). (This literature review has attempted to foreground more authentic studies, but some cited here may be suspect.) Because informational design products are so complex and holistically constituted, generalization is remarkably problematic. Precision must be softly sacrificed for authenticity.

The textbook illustration literature hinges on the relationship of illustration to a primary text. Dual-coding theory most often serves as the theoretical basis for inquiry into this relationship. Pictures, in short, have certain affordances based in their spatial nature that enable them to function in ways that text alone cannot. The textbook illustration literature outlines these affordances and they are summarized in the efficacy and function typologies. It should be noted that the literature affords text primary status, and never purposefully considers text as a possible adjunct to imagery. However, image-caption structure, considered as a complex separate from a containing layout, does presume primacy for illustration, and is addressed in the literature.

2.5.11 Truth Claims from the Literature

Some claims from the textbook illustration literature are articulated here in isolation for the purpose of focusing the conclusions for this particular study. This list is by no means exhaustive.

1) Reader visuospatial ability is a factor in illustration function. (Mayer et al., 1995)
2) Well-crafted and easy text enjoys little benefit from illustration. (Levin, 1979)
3) Difficult text requires mnemonic or other intervention. (Levin, 1979)
4) Students have been trained to bypass textbook illustrations. (Pettersson, 1998; Carney & Levin, 2002)
5) Visual displays render abstract concepts more concrete. (Vekiri, 2002)
6) Displays must be aligned with goals to be effective. (Vekiri, 2002)
7) Time spent on relevant info (selective attention) equates with strength of memory representation. (Reynolds, 1992)
8) Illustrations can provide runnable mental models. (Carney & Levin, 2002)
9) Gestalt principles minimize cognitive processing demands when used to construct visuals. (Vekiri, 2002)
10) Illustrations are more effective with explicit instructions for interpretation. (Pettersson, 1998; Carney & Levin, 2002)
11) Text–picture integrated designs are more effective. (Mayer et al., 1995; Vekiri, 2002)
12) Mental integration of language and image is facilitated by design integration. (Carney & Levin, 2002; Vekiri, 2002)
13) Illustrations improve learning of explanatory information and problem solving if labeled. (Vekiri, 2002)
14) Short explanatory texts are more effective for retention and transfer than longer texts if accompanied by illustrations. (Vekiri, 2002)
15) Explanatory pictures function best with cause–effect systems or complex processes. (Carney & Levin, 2002)
16) Picture-based reiteration functions as well as text reiteration, without being verbatim. (Levin, 1979)
17) Annotated multi-frame illustrations for cause–effect systems improve problem solving. (Mayer et al., 1995)
18) Mnemonic strategies aid in acquisition of higher-order concepts and skills. (Carney & Levin, 2002)
2.6 Learning Typologies

2.6.1 Bloom’s Taxonomy

Bloom’s Taxonomy (Bloom et al., 1956) provides a means to assess learning by giving coherent structure to educational objectives. It has been widely cited in the half century since its publication. Bloom et al. produced taxonomies for both the affective and cognitive domains for educational objectives. The cognitive domain is what is most commonly referred to as Bloom’s Taxonomy. The taxonomy is divided into six classes of objectives, five of which are subdivided (Table 2.6.1).

In some cases there are further subdivisions. For instance, knowledge of specifics (1.10) is subdivided into knowledge of terminology (1.11) and knowledge of specific facts (1.12). Knowledge is lowest in the hierarchy of Bloom’s Taxonomy. The upper five classes are sometimes referred to as the “higher order” thinking skills (while in other cases that identification is reserved for the upper three). Knowledge relates to recall from a static store of information, while the five classes above it represent “the intellectual abilities and skills” or “organized modes of operation and generalized techniques for dealing with materials and problems” (p. 189). Comprehension involves understanding of communication as well as use of communication, but stops short of “relating it to other material or seeing its fullest implications”—those abilities are higher yet (p. 190). Application refers to the ability to utilize knowledge in real situations. Analysis involves breaking information down and organizing it for the sake of clarity and conveyance. Synthesis involves creating new patterns from previously disorganized parts. (Synthesis is clearly creative in nature.) Evaluation involves judgments and is reminiscent of critical thinking skills. The hierarchical arrangement, with knowledge lowest and evaluation highest, is meant to signify behavioral complexity, where knowledge is the simplest (p. 15).
Table 2.6.1: Bloom’s cognitive domain (Bloom et al., 1956)

<table>
<thead>
<tr>
<th>Level</th>
<th>Domain</th>
<th>Sublevel 1</th>
<th>Sublevel 2</th>
<th>Sublevel 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>Knowledge</td>
<td>1.10 Knowledge of specifics</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.20 Knowledge of ways and means of dealing with specifics</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.30 Knowledge of the universals and abstractions in a field</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.00</td>
<td>Comprehension</td>
<td>2.10 Translation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.20 Interpretation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.30 Extrapolation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.00</td>
<td>Application</td>
<td>3.— No subdivisions specified.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.00</td>
<td>Analysis</td>
<td>4.10 Analysis of elements</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.20 Analysis of relationships</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.30 Analysis or organizational principles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.00</td>
<td>Synthesis</td>
<td>5.10 Production of a unique communication</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.20 Production of a plan, or proposed set of operations</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.30 Derivation of a set of abstract relations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.00</td>
<td>Evaluation</td>
<td>6.10 Judgments in terms of internal evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.20 Judgments in terms of external criteria</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Marzano (2001) proposes a significant reworking (or replacement) of Bloom’s Taxonomy. In so doing he interprets and critiques the taxonomy. Marzano downplays the effect of Bloom’s Taxonomy on curriculum planning, offering instead Gagne’s framework as detailed in *The Conditions of Learning* (Gagne, 1977). Bloom’s Taxonomy is most influential, instead, with evaluation, where it provides “detail in stating goals” (Marzano, 2001, p. 2).

The hierarchical basis of Bloom’s Taxonomy is the major target of Marzano’s critique. As evidenced in research, Bloom’s domains are not in fact hierarchical (pp. 8, 10). Marzano states that a taxonomy cannot be based on “difficulty of mental processes” because difficulty is relative to familiarity (p. 10). (Marzano is not entirely convincing in this because
he does acknowledge, alongside the familiarity factor, that complexity of processes is fixed.)

Marzano characterizes Bloom’s Taxonomy as a framework and not a model or theory, indicating that it is unable to predict behavior (p. 12).

2.6.2 Bloom’s Revised

Anderson and Krathwohl (2001) updated Bloom’s Taxonomy—Krathwohl was co-author of the original Bloom’s publication. Bloom’s Taxonomy Revised is considerably more tightly structured and expansive than the original taxonomy, and notably separates knowledge from the objectives classes and expands it. (A schematic of the taxonomy is provided as Figure 8.1.1 in the appendix.)

As in the original taxonomy, Bloom’s Revised gives structure to educational objectives. Educational objectives are situated between the superior category of global objectives and the inferior instructional objectives. Global objectives are broad and cover years of curricular planning. Educational objectives plan units of instruction—this is the level at which Bloom’s Revised is meant to operate. Instructional objectives involve day-to-day activities. Instructional objectives should serve educational objectives, which in turn contribute to global objectives.

Bloom’s Revised has two dimensions: cognitive processes and knowledge domains. Educational objectives are articulated as: “The student is able to [cognitive process] (a verb) [specific knowledge] (a noun),” within a subject area lesson. For instance: “The student is able to categorize knowledge of sustainability principles.” In this example, the student must be aware of “sustainability principles” and demonstrate an ability to categorize them.

The most striking difference between Bloom’s Taxonomy and Bloom’s Revised is the extraction of knowledge from the cognitive process dimension into its own second dimension. Four classes of knowledge are identified, each of them subdivided into types. Factual knowledge includes knowledge of terminology, specific details and elements. Factual knowledge is at a low level of abstraction, representing straightforward information. Conceptual knowledge is more complex and organized, and includes knowledge of
classifications, categories, principles, generalizations, theories, models and structures. Each type has its own characteristics, which are detailed in the taxonomy. Procedural knowledge refers to conditionalized performance and metacognitive knowledge is knowledge of one’s own cognitive strategies. Theoretically, any class of knowledge can be manipulated in any cognitive process.

The cognitive processes in Bloom’s Revised are, after the original taxonomy, arranged in order of increasing cognitive complexity. Where knowledge formed the lowest rung of Bloom’s Taxonomy, remember—distinctly process-based where knowledge was atypical—is the base of Bloom’s Revised. Remember includes recognizing and recalling. The higher-order classes have one-to-one equivalence to the original higher-order classes, with minor changes in terminology (including a consistent shift to action verbs) and one case of reordering. These changes are summarized in Table 2.6.2.

<table>
<thead>
<tr>
<th>Original</th>
<th>Revised</th>
<th>Significant alteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.00</td>
<td>Evaluation</td>
<td>Create</td>
</tr>
<tr>
<td>5.00</td>
<td>Synthesis</td>
<td>Evaluate</td>
</tr>
<tr>
<td>4.00</td>
<td>Analysis</td>
<td>Analyze</td>
</tr>
<tr>
<td>3.00</td>
<td>Application</td>
<td>Apply</td>
</tr>
<tr>
<td>2.00</td>
<td>Comprehension</td>
<td>Understand</td>
</tr>
<tr>
<td>1.00</td>
<td>Knowledge</td>
<td>Remember</td>
</tr>
</tbody>
</table>

At the level of cognitive process type Bloom’s Revised expands somewhat upon the original taxonomy. Where application included no types, apply has two (executing and implementing); where comprehension had only three types, understand combines two into one and adds five new types. All revised types, even when equivalent to original types, are more elegantly named and better described. One or more action verbs are associated with each type for clarification. For instance, inferring can entail concluding, extrapolating,
interpolating, and predicting. These alterations render Bloom’s Revised considerably more comprehensible and operational than its predecessor.

As of 2008, Bloom’s Revised was used within the North Carolina public school system.

2.6.3 Enduring Understanding

Wiggins and McTighe (1998) propose a multifaceted view of enduring understanding. Deep understanding can be exhibited by:

1) Explaining meaningful relationships,
2) Interpreting ambiguous material,
3) Applying knowledge to new situations,
4) Holding a perspective and acknowledging point-of-view,
5) Empathizing with others’ worldviews, and
6) Knowing and managing one’s own thinking patterns. (pp. 45–60)

Guiding verbs that illustrate what students should exhibit are associated with each type. Explanation, for instance, includes “explain, justify, generalize, predict support, verify, prove, and substantiate” (p. 47). Though the latter facets of enduring understanding may be more difficult to achieve (at least in some situations), the list is neither hierarchical nor developmental. These are facets, not hard-and-fast categories, and not stages. As worded, the facets of enduring understanding are observable. They are also all considered to be “higher” than fact acquisition or rote memorization. That is, the qualifier enduring means that this is not a full typology of learning, but rather desirable types of learning or evidence of real understanding.

Wiggins and McTighe are explicit that classroom activity must be aimed at an outcome. The materials utilized in class must be selected in order to achieve a standard—this is referred to as backward design (p. 8). Classroom activities must point to an ultimate
assessment, rather than assessment being tacked on at the end. A curriculum planner must know what evidence of student learning he or she wants, and the six facets of enduring understanding are meant to serve in that capacity. *Enduring understanding* is like Bloom’s Taxonomy in general use, but the strong emphasis on backward design is Wiggins and McTighe’s special contribution.

2.6.4 The New Taxonomy

Marzano’s (2001; Marzano & Kendall, 2007) New Taxonomy stems from a theoretical model of *control* in mental processes. (A schematic of the New Taxonomy is provided as Figure 8.1.2 in the appendix.) Three mental systems are nested or connected through a learning sequence. The *self-system* is the first in the sequence and determines if the learner will become engaged in a given activity. Motivation resides in the self-system. The self-system works by assessing the benefits and costs associated with a task, making probabilistic determinations and responding accordingly. A response of “yes” (to engagement) feeds into the next system, the *metacognitive system*. The metacognitive system is involved in determined goals and strategies for a task. It continues to interact with the next system, the *cognitive system*, throughout the task. The cognitive system handles the processing that was essentially ordered by the metacognitive system. The metacognitive system is analytical and involves “making inference, comparing, classifying, and the like” (Marzano, 2001, p. 12). *Knowledge* is a final component of Marzano’s underlying theoretical model, and though his schematic fails to indicate this, it is used by all three systems (the schematic depicts knowledge affecting only the cognitive system and as an outside influence).

Knowledge is composed of three domains: *information, mental procedures*, and *psychomotor procedures*. Different subject areas comprise different balances of knowledge domains. The information domain—or *declarative information*—is predicated on accuracy. Marzano lists two information categories: *details* and *organizing ideas*. Details include *episodes, cause–effect sequences, time sequences, facts*, and *vocabulary terms*. Organizing
ideas are either *principles* or *generalizations* (which is considered equivalent to *concepts*) (pp. 17–18).

Procedural knowledge is the “how-to” to the “what” of declarative knowledge (p. 23). Mental procedures are described as propositional, made up of *productions* (if–then “structures”). Procedures become internalized through a three-stage process. The first stage is cognitive, where the activity must be consciously attended to; the second stage is *associative*, with errors managed; and the final stage is *autonomous*, where internalization is completed. *Procedural knowledge in the first stage only is equivalent to information knowledge*. There is a hierarchy to procedural knowledge: *processes* and *skills* are superordinate types. Processes contain only (and so *are*) *macroprocedures*, which are complex, open-ended, and must be controlled. Skills include *tactics*, *algorithms*, and *single rules*, which are differentiated by the complexity of the steps (single rules being single steps).

Psychomotor procedures, not particularly relevant to this survey, involve *processes* (or *complex combinational procedures*) and *skills* (*simple combination procedures* and *foundational procedures*).

Citing a recent dissolution of the conception of a short- and long-term memory split, the New Taxonomy is based on three *functions* of memory. *Sensory memory* stores relatively complete records for a short amount of time. If attended to, the records of sensory memory can be partially encoded into permanent memory. Otherwise, what is represented in sensory memory is lost. *Permanent memory* is the knowledge store. *Working memory* utilizes both sensory and permanent memory in processing (pp. 29–30). Memory is a core element throughout the levels of the New Taxonomy.

There are six levels to the New Taxonomy, as in Bloom’s Taxonomy, Bloom’s Revised and Enduring Understanding. The three mental systems are distributed among the six levels, with the cognitive system subdivided into four levels. Marzano arranges the systems in reverse order in his schematic, so that level 6 is highest. The six levels of the New Taxonomy are: (6) *self-system*, (5) *metacognitive system*, (4) *knowledge utilization* (the first of four levels of the cognitive system), (3) *analysis*, (2) *comprehension*, and (1) *retrieval*. 

69
Self-system thinking (level 6) includes *examining importance, examining efficacy, examining emotional response, and examining overall motivation*. The first three types of self-system thinking determine the latter: motivation (pp. 50–54). The metacognitive system (level 5) is responsible for executive control. It is comprised of: *goal specification, process monitoring, monitoring clarity, and monitoring accuracy*. Neither self-system thinking nor metacognition have correlates in Bloom’s Taxonomy (pp. 48–50). The remaining four levels of the New Taxonomy are considered to be the cognitive system.

“Knowledge utilization processes [level 4] are those that individuals employ when they wish to accomplish a specific task” (p. 45). Knowledge utilization generates new “products.” It includes: *decision making, problem solving, experimental inquiry, and investigation*. Analysis (level 3) differs from localized (on-the-spot) inferences in that it is a reasoned *extension* of knowledge previously encoded. Analysis includes: *matching* (“identification of similarities and differences between knowledge components”), *classification, error analysis* (“reasonableness” of knowledge), *generalization* (more inductive than deductive), and *specification* (more deductive than inductive). As it is impossible (and probably undesirable) to store all incoming information faithfully in permanent memory, we must translate it so as to manage the information (retaining some and discarding the rest). Comprehension (level 2) refers to the *parsimonious* storage of information in permanent memory. Comprehension is either a process of *synthesis* (distilling knowledge) or *representation* (the creation of knowledge analogs with mental imagery). Retrieval (level 1) has the lowest level of conscious awareness. In fact, Marzano states that his hierarchy is organized in terms of consciousness, with level 6 the most conscious and level 1 (retrieval) the most automatic (p. 55). Retrieval involves transfer of information from permanent memory to working memory. It is not direct (that is, not consistent with a conception of instruction as transmission), as inferred information will be encoded along with provided information. “Stored” information comes in the form of *microstructures*, which can include both default and reasoned inferences. The former utilizes general knowledge while the latter utilizes more specific knowledge and involves induction.
In addition to the New Taxonomy’s consciousness hierarchy, it has a flow-of-processing hierarchy, with self-system thinking (level 6) as the initial stage (pp. 54–55). The New Taxonomy, as Marzano stresses, is not hierarchical in terms of complexity. There is no assumption that complexity varies among the types (p. 55).

2.6.5 The New Taxonomy vs. Bloom’s Revised

The New Taxonomy is in many ways equivalent to Bloom’s Revised. The differences are largely structural, which is significant because the structure of these taxonomies defines a model of thinking and learning. Both include metacognition, but while Bloom’s Revised conceptualizes it as a type of knowledge, the New Taxonomy breaks it down into processes, all of which can relate to any kind of knowledge. Self-system thinking in the New Taxonomy is a model for motivation, integrated into the flow of the other systems of thought. Motivation is not incorporated into Bloom’s Revised. The remaining four subsystems in the New Taxonomy’s cognitive system are compared with the six hierarchical levels in Bloom’s Revised in Table 2.6.3.

A plus-minus symbol (±) with italics denotes similarity. Many of the equivalent types exist at very different levels of the two taxonomies, as can be seen by comparing level numbers (the table is fixed to the levels of the New Taxonomy). A question mark and italics denotes types in Bloom’s Revised that are not necessarily equivalent to types in the New Taxonomy. However, many of these types are apparently confounded with others, indicating that they may not be missing, per se, from the New Taxonomy.
Table 2.6.3: Bloom’s Revised vs. The New Taxonomy

<table>
<thead>
<tr>
<th>Level</th>
<th>Bloom’s Revised</th>
<th>New Taxonomy</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Producing</td>
<td>?</td>
<td>None (ill-defined)</td>
</tr>
<tr>
<td>6</td>
<td>Generating</td>
<td>= Decision Making</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Planning</td>
<td>= Problem Solving</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>Checking</td>
<td>= Experimenting</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Inferring</td>
<td>?</td>
<td>None (confounded with Experimenting?)</td>
</tr>
<tr>
<td>5</td>
<td>Critiquing</td>
<td>= Investigating</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Explaining</td>
<td>?</td>
<td>None (confounded with Investigating?)</td>
</tr>
<tr>
<td>2</td>
<td>Comparing</td>
<td>= Matching</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Classifying</td>
<td>= Classifying</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Exemplifying</td>
<td>?</td>
<td>None (confounded with Classifying?)</td>
</tr>
<tr>
<td>4</td>
<td>Attributing</td>
<td>± Analyzing Errors</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Summarizing</td>
<td>= Generalizing</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Organizing</td>
<td>± Specifying</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Differentiating</td>
<td>= Integrating</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>Interpreting</td>
<td>= Symbolizing</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>Recalling</td>
<td>= Recall</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>Recognizing</td>
<td>= Recognition</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Executing</td>
<td>= Execution</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Implementing</td>
<td>?</td>
<td>None (specific to mental procedures)</td>
</tr>
</tbody>
</table>

2.6.6 Discussion

Bloom’s Taxonomy is ultimately ambiguous. The major terms (analysis, synthesis, etc.) are familiar but not technically defined. This leaves it open to interpretation to a degree that weakens it as a device for standardization and precision. Wiggins and McTighe’s Enduring Understanding is perhaps better defined, but is also too simple (six types with no subdivisions).

Bloom’s Revised and The New Taxonomy are comparably complex. Marzano’s New Taxonomy serves as a model for processing and accounts for motivation. Bloom’s Revised
does neither of these things, but is more straightforward and comprehensible. Bloom’s Revised also specifies more cognitive types, but some may be redundant. It also includes two problematic types: producing and implementing. Producing is “inventing a product,” which doesn’t seem to warrant a separate type but rather focuses on a performance that necessarily incorporates, and is thus defined by, other objectives. Implementing is specific to procedural knowledge, but is the only type so constrained, which represents an inconsistency in the model. Ultimately, the greater number of cognitive types in Bloom’s Revised doesn’t appear to equate with any greater breadth.

While the New Taxonomy is strong for providing a flow of information hierarchy, its second hierarchy of level of consciousness is dubious. Self-system thinking should be the most conscious, followed by metacognition. At some times this may be true, but certainly efficacy judgments, for example, are most often done unconsciously. Bloom’s Revised uses a cognitive complexity hierarchy, which is theoretically problematic. Especially suspect is executing, which is positioned relatively high at the third level yet involves the automatic use of skills and algorithms to provide a predetermined answer or perform a familiar task. It is hard to justify this as a more demanding type than inferring or interpreting, yet those two are placed at the second level.

Finally, it should be reiterated that Bloom’s Revised is used to describe the standards in North Carolina public schools.

2.7 Performance Evaluation

2.7.1 Kinds of Assessment

Bloom’s Revised and The New Taxonomy are both designed to help teachers and administrators tie instruction to desired outcomes. Instructional changes should be evaluated, as should student performance. Baron (1987) provides four dimensions relevant to instructional evaluation:
1) Formative–summative. Formative evaluation occurs during an instructional program and enables the teacher to make adjustments mid-stream. It is used to improve the instructional program. Summative evaluation occurs once the program is complete, and is aimed at “net effectiveness.”

2) Product–process. Product evaluation focuses on the work outcomes of students, while process evaluation focuses on “the internal workings of the situation” at hand.

3) Qualitative–quantitative. Qualitative evaluation is concerned with description and depth, while quantitative evaluation is standardized and makes it easy to compare students (though without the qualitative depth).

4) Experimental–quasi-experimental designs. Like the qualitative–quantitative dimension, this dimension is methodological. True experiments must utilize random sampling and control groups. In education this is rarely possible, especially when authenticity or ecological validity is a concern. Quasi-experiments attempt to make the best of a real situation through various methods. (pp. 223–225)

These dimensions are largely just as relevant for student performance evaluations. If a teacher wishes to determine how a student arrives at his or her products, then there must be a way to capture something of process (in the product–process dimension).

2.7.2 A Design Performance Assessment Battery

Kimbell et al. (1991) were charged with assessing the British Design and Technology program, determining to focus on the 14–16 age range. Their study was meant to “provide support for curriculum development” in an interventionist capacity: they wanted to understand why students performed as they did at novel design tasks (p. 11). This meant that they had to capture process in some manner.
Kimbell et al. developed a model of design process (Figure 2.7.1) that accounts for both mental imaging and physical modeling.

**Figure 2.7.1: Design process (redrawn from Kimbell et al., 1991, p. 20)**

Modeling is important in design because of the limitations of mental imagery, both in terms of processing demands and the natural “haziness” of mental imagery. The “hazy impressions” in an early stage of a design process are brought out into the open through sketches and discussion, which provide clarification and a means for eliciting input from others. While it is impossible to see what is inside a student’s head during conceptualization, sketching and discussion is its direct product. Formative products, if properly modeled, can be used to infer process. Indeed, the manipulation of mental imagery in a design process is considered to be “cognitive modeling,” paired with modeling “in reality” (by hand) (p. 21). In the context of a design project, then, the two types of modeling are unified and represent
“thought in action,” rendering design process theoretically observable (though in a limited capacity to be sure).

Design projects are too involved and extended to capture in a test, so Kimbell et al. conceptualized their tests as opening a window into a manageable part of the process. They broke a general process into four stages—identifying and clarifying tasks, investigating, generating and developing solutions, and appraising/evaluating—and tested on each stage individually (p. 29). A test on the second stage, for instance, provides the students with a completed first stage, from which they base their work. The test itself took the form of a folded worksheet that presented steps for the students to work through. These involved writing and modeling tasks that lead towards a solution proposal, along with reflective self-evaluation.

The marks account for both sides of Baron’s qualitative–quantitative dimension. Raters made holistic judgments on a scale from 0 to 5. They also judged against specific descriptors in 13 categories, which by virtue of such systematic subdivision is squarely quantitative. The tests thus described are 90-minute focused tests. Kimbell et al. used two other testing methods. One involved extended project work, which was intensive, deep, and consequently not statistically reliable. The extended project work was qualitative and humanistic in nature. This was time-consuming and could only be done with a limited number of students. Another small number of students participated in modeling tests, which extended beyond the 90-minute paper-based tests. The modeling tests built on the focused tests (the students collaborated and built solution models), and were balanced between the heavy qualitative focus of the extended project work and the largely quantitative 90-minute tests (p. 35).

The marks on the 90-minute focused tests make it possible to represent the student’s process with a single holistic mark, account for depth of self-appraisal, and categorize the student’s process in terms of a reflective, active, or more balanced tendency (p. 170).
2.7.3 Discussion

Taxonomies of educational objectives like Bloom’s Revised and Marzano’s New Taxonomy provide a means to order and differentiate student behavior, but capturing evidence that objectives are being met is complicated by the “black box” of the student’s mind. Kimbell et al. developed an exemplary battery of tests that elegantly fit a theoretical model to evidence through student work. The details have been glossed over in this literature review, but the Kimbell et al. study can serve as a demonstrative case study. However, each situation requires unique solutions, so Kimbell et al. cannot serve as a completed road map.
SECTION 3 — METHODOLOGY

3.1 Approval

This research study was approved by both the North Carolina State University Institutional Review Board (IRB#1359) and the Wake County Public School System Research Review Committee. The study was conducted in the spring of 2011.
3.2 Research Design Logic

3.2.1 General Purpose

The goal of this study is to examine cognitive load effects of design strategies for visual content relationships. Though it would be desirable to directly measure germane, intrinsic, and extraneous loads in relation to design strategy, there is currently no way to do so (Schnotz and Kürschner, 2007, p. 500). We can do better with estimations of overall cognitive load, however, and make reasonable inferences for individual types of load, by monitoring the outcomes related to cognitive load effects, including comprehension (directly addressed in this study), learning (not addressed), and mental effort (indirectly addressed).

3.2.2 Research Questions and Research Design Distinctions

Each research question for this study relates to strategies for text–image integration. Instead of focusing on an isolated relationship (a certain picture–caption arrangement, say, with different image functions employed), strategies for establishing entire suites of relationships will be considered as the units of interest. Reading and using textbooks cannot be fully understood through isolated components and extrapolated results alone. Students utilizing textbooks must understand (or navigate) the myriad meaningful relationships on the page in order to achieve comprehension and learning. This study considers holistic design strategy in a bid for authenticity and ecological validity. (Studies of isolated elements can be instructive, but they must always be somewhat inauthentic to the ultimate task of designer and reader.) Differing strategies for text–image relationships are evident in familiar print media including illustrated narrative books, instructions for assembly, and reference volumes (such as field guides). This study seeks to tease out some of the implications behind available strategies.
• Research question #1 (RQ1): How does the strategy for text–image integration in instructional print media for middle school students impact their comprehension of abstract relationships in science material?

According to cognitive load theory, comprehension is achieved when a learner can successfully manipulate all requisite elements for an individual learning task in working memory. Comprehension can theoretically occur without learning. For instructional print media, comprehension must be measured while the reader is engaged with the media, and not after. Therefore, this study must include performance tests to be used in concert with variable-strategy instructional print media. As discussed in the conceptual framework, three strategies are identified along a continuum (Figure 3.2.1). Near the two extremes of text–image integration are the fully integrated and prose primary strategies, the latter being the least integrated of the three types. Somewhere on the continuum between these types is prose subsumed.

Comprehension is an issue of performance, so RQ1 requires a performance measure. While the specific levels of educational objectives are difficult to target with certainty, Bloom’s Revised (Anderson & Krathwohl, 2001) provides a promising general framework within its taxonomy. Knowledge is organized into factual and conceptual levels. Factual knowledge is at a low level of abstraction while conceptual knowledge is more complex and organized. Measuring performance with factual and conceptual knowledge might offer better insight into comprehension than one type alone.

• Research question #2 (RQ2): How does the strategy for text–image integration in instructional print media for middle school students impact their perception of task difficulty?

The effort a learner must put into a comprehension task is a significant factor in learning. The design of a textbook page may place extraneous load on a learner attempting to work with it. Perception of task difficulty is deemed a more direct measure of instructional
effects than is perceived mental effort (Brünken et al., 2003). Mental effort acts as an artificial ceiling to a learner’s cognitive load because a learner can fail to commit his or her full faculties to a task. More to the point of this study, variable print media will place variable demands on readers, resulting in variable levels of (requisite) mental effort. A learner’s assessment of mental effort is problematic because low mental effort may indicate easy materials (where it’s not necessary to commit full resources to complete a task), or conversely, especially difficult materials (so difficult that the learner disengages and gives minimal effort) (p. 56). If the learner is asked not how hard he or she worked at a task, but rather how difficult that task was itself, a more direct connection can be made to comprehension processes. Perception of task difficulty must be assessed immediately following an *individual* task, not a suite of tasks. This study must inquire into students’ perceptions of task difficulty during performance tests. For this study, the “task” entails using textbook pages to answer a set of questions.

- Research question #3 (RQ3): How does the strategy for text–image integration in instructional print media for middle school students impact their *interest level in subject matter*?

Middle school students may very well prefer one text–image integration strategy to another. Ultimately, it is curricular goals that matter in education. Of particular interest then is how text–image integration strategy might change perception of the subject matter it is used to represent. RQ3 probes into the role of text–image integration strategy as a mediating variable for interest in subject matter.

- Research question #4 (RQ4): How does the strategy for text–image integration in instructional print media for middle school students impact their *interest level in instructional print media*?
If learners find one text–image integration strategy to be inherently interesting, then they may very well approach materials designed according to that strategy with more enthusiasm and perhaps even higher investment of mental effort. While the previous research questions focus more directly on student performance with media, RQ4 is a direct inquiry into student perception of design itself.

Figure 3.2.1: Three strategies (modified from McDougal Littell, 2005a, C18–C19)
3.3  Research Design

3.3.1  Forms for Testing

This study aims to be as naturalistic as possible. According to this central tenet, students worked at their regular desks using provided textbook pages as a resource to complete multiple-choice tests. According to cognitive load theory, comprehension involves knowledge construction from working memory (Schnitz & Kürschner, 2007). By allowing subjects to work with a textbook form while attempting to answer questions about its content, performance becomes a matter of working memory as opposed to long-term memory. The textbook then serves a role as external cognition, and the textbook’s efficiency in this role should be, in part, a factor of its design. The “open book” multiple-choice test becomes the treatment task as well as the primary method of data collection.

3.3.2  Sequencing

The three described types of text–image integration strategy—prose primary (PP), prose subsumed (PS) and fully integrated (FI)—serve as the treatments of this study. Conventional textbooks are designed according to a prose primary strategy. Since it can’t be said that textbooks are not designed—that they are neutral—the prose primary strategy is not conceived of as a control but is rather one of three treatments.

Assigning subjects to groups corresponding to the three strategies was considered. This might be desirable to counteract possible familiarity effects. If students are used to working with prose primary textbooks, then that treatment might receive an artificial boost due to familiarity. Multiple treatments might show a trend of improvement over treatments with the less familiar fully integrated forms. However, access to the school was limited, and it was determined that three or four treatments alone would not be enough for subjects to gain sufficient familiarity with an unexpected design strategy.
To best control for form equivalency, each subject must receive each treatment type once. Three treatment sessions are thus required for the three treatment conditions. The subjects were available for only four class periods, which left one session open for a pre-test battery. To address possible order effects, groups were further divided to accommodate all possible sequences. This results in six order-based groups (Figure 3.3.1).

<table>
<thead>
<tr>
<th>Group</th>
<th>Treatment #1</th>
<th>Treatment #2</th>
<th>Treatment #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP-PS-FI</td>
<td>Prose Primary (PP) →</td>
<td>Prose Subsumed (PS) →</td>
<td>Fully Integrated (FI)</td>
</tr>
<tr>
<td>PP-FI-PS</td>
<td>Prose Primary (PP) →</td>
<td>Fully Integrated (FI) →</td>
<td>Prose Subsumed (PS)</td>
</tr>
<tr>
<td>PS-PP-FI</td>
<td>Prose Subsumed (PS) →</td>
<td>Prose Primary (PP) →</td>
<td>Fully Integrated (FI)</td>
</tr>
<tr>
<td>PS-FI-PP</td>
<td>Prose Subsumed (PS) →</td>
<td>Fully Integrated (FI) →</td>
<td>Prose Primary (PP)</td>
</tr>
<tr>
<td>FI-PP-PS</td>
<td>Fully Integrated (FI) →</td>
<td>Prose Primary (PP) →</td>
<td>Prose Subsumed (PS)</td>
</tr>
<tr>
<td>FI-PS-PP</td>
<td>Fully Integrated (FI) →</td>
<td>Prose Subsumed (PS) →</td>
<td>Prose Primary (PP)</td>
</tr>
</tbody>
</table>

**Figure 3.3.1: Order-based groups plan**

Hyphens in sequence group names indicate separations between treatment sessions. Both PP-PS-FI and FI-PS-PP subjects received the prose subsumed (PS) form in the second treatment session.

### 3.3.3 Sessions

For each session, students across groups worked on the same subject matter, with the three forms all based on an existing textbook spread (a spread is two facing pages as the book sits open). Thus, each session corresponds to a subject area from the 8th-grade curriculum. Eighth-grade material was chosen so that the information would be new to the 7th-grade
subjects. Each treatment session covers material from a different section of the textbook. Qualifying material had to be concrete in nature, so that imagery could have direct meaning and be easily produced. Chemistry, for instance, is too abstract and more mathematical than it is visual. The three selected spreads come from two sections of the textbook. The subjects are:

- Divides and drainage basins, covering large scale (geographic) water movement;
- Lakes and lake turnover, from the same sections as the divides subject but focused on a different scale and with other core concepts; and
- Fossil fuels and coal.

These discrete subjects differ enough that the study concerns visually rich science material in general rather than some narrow content area.

### 3.3.4 Dependent Variables

Groups were determined according to the independent variable (IV) of text–image integration strategy. Other variables were measured during the treatments or through pilot, pre- or post-tests. The dependent variables (DV) and other variables measured are:

- Interest in subject matter (DV)
- Interest in text–image integration strategy (DV)
- Task difficulty as mediated by IV (DV)
- Session duration (DV)
- Performance in science class (through a review test)
- Science self-efficacy
3.4 Sampling

3.4.1 School Selection and Description

This study includes nearly the entire 7th-grade population of a local middle school. Centennial Campus Magnet Middle School is convenient for two reasons: (a) it is local; and (b) it is affiliated with N. C. State University, with a research mandate—teachers are encouraged to participate in research studies. The students are accustomed to acting as research subjects, so a researcher is somewhat less disruptive than might otherwise be expected.

Centennial Campus Magnet Middle School (CCMMS) is in central North Carolina (the Piedmont region) in Wake County, a largely urban and suburban area. The school enrolled 601 students in the 2010/2011 school year, making it the 3rd smallest of the 32 public middle schools in Wake County, twenty of which have populations exceeding 1,000 (Wake County Public School System [WCPSS], 2010d). CCMMS is an urban school in the capital city of Raleigh with higher proportions of minority students than the national average (Table 3.4.1; WCPSS, 2010a; U. S. Department of Education [USDOE], 2009). CCMMS has limited English proficiency (LEP) students at numbers slightly above the average of Wake County public schools: 10.0% compared with 7.9% (WCPSS, 2010d). English as a Second Language Program (ESL) enrollment is 4.8%, compared with 4.3% at the county level (ibid.). 36.3% of CCMMS students qualify for the Free & Reduced Lunch program, close to the county average of 32.4% (WCPSS, 2010c). This is lower than 2008/2009 national and state averages, at 44.6% and 45.9%, respectively (USDOE, 2010). Free & Reduced Lunch is an easily accessible metric directly related to economic status.

Seventh grade classrooms at CCMMS are arranged around a central meeting space, separated from 6th and 8th grade sections. Seating arrangements vary based on teacher preference. One of the participating teachers for this study arranges desks into clusters, while the other teacher uses rows.
Table 3.4.1: Racial breakdown (WCPSS, 2010a & 2010b; USDOE, 2009)

<table>
<thead>
<tr>
<th>Racial Identification</th>
<th>School</th>
<th>County</th>
<th>National</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Indian</td>
<td>0.0%</td>
<td>0.4%</td>
<td>1.2%</td>
</tr>
<tr>
<td>Asian</td>
<td>3.8%</td>
<td>6.0%</td>
<td>5.0%</td>
</tr>
<tr>
<td>Black</td>
<td>35.0%</td>
<td>24.8%</td>
<td>17.0%</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>19.7%</td>
<td>14.6%</td>
<td>21.5%</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>3.0%</td>
<td>4.5%</td>
<td>0.5%</td>
</tr>
<tr>
<td>White</td>
<td>38.5%</td>
<td>49.5%</td>
<td>54.9%</td>
</tr>
</tbody>
</table>

3.4.2  Group Selection

In order to maximize ecological validity, cluster sampling was utilized, with classes as the natural groups. There are two science teachers at the 7th-grade level, with four classes apiece, for a total of 199 students in eight classes. Assignment of individuals to classes cannot be considered random in the strictest sense, but the school does follow “mainstreaming” practice, where students of all ability levels comingle. However, each class develops its own personality, and a few individuals’ behavior can affect all others in a given class. Furthermore, no two teachers can be assumed equivalent. For these reasons subjects were randomly assigned in roughly equal numbers to one of the three groups within each class.

3.4.3  Group Assignment

Group assignment was conducted in a manner similar to a multistage clustering procedure, in which groups are sampled first and then subject names are attained for within-group sampling (Creswell, 2003, p. 156). This was necessary as reliable class rolls were not available at the start of the study. In this case, of course, it is not a matter of sampling but
merely one of group assignment, as all 7th-grade classes at the participating school and all consenting 7th-grade students were included in the study. Forms were shuffled at eight apiece into groups of 24. Beneath these stacks, shuffled groups of three, with one of each form represented, were placed. This ensured the most random distribution of forms within the expected 23- or 24-student class size, while maintaining a good groups balance as individual classes reached higher numbers.

On the first treatment day, the shuffled forms were handed out to the subjects by the proctor walking through the classroom and giving each new subject the top form. Subjects reported their form codes as the first item of the instrument. This served to assign subjects to one of the three sub-groups (PP, PS or FI) for the first treatment. A further round of group assignment was required to determine the order in which subjects would receive the final two treatments. A subject who began with PP could either be assigned to PS and FI for the second and third treatments, respectively, or FI and PS.

Reliable class rolls were available after the first treatment, and these were used to assign subjects to one of the final six order-based groups. Within each cluster, subjects were assigned their order-based groups using an online random number generator. According to the protocol, if there were 8 subjects in the treatment #1 PP group, then numbers 1–8 were randomized, with odd numbers corresponding to the PP-PS-FI group and even numbers PP-FI-PS. With each new sorting task for odd numbered groups (5, 7 or 9 subjects in a sub-group), the meaning of odd and even randomly ordered numbers was flipped, so that final order-based groups would be as balanced as possible.

3.5 Form Development

3.5.1 Equivalency of Forms

Existing textbook spreads (McDougal Littell, 2005a) were of the prose primary strategy, as is typical of textbooks. Basic design styles—such as typeface suite, text size, title
treatments, and margins—were established and carried through for all versions of all forms. Though the original pages were prose primary, even the prose primary versions were recreations. This was done for the sake of form equivalency: alternate versions should vary based on text–image integration strategy, not aesthetics or legibility (and in this case, one designer should complete all three, so designer is not a variable).

In each case the prose primary re-creation preceded the other versions. Prose subsumed followed. For the prose-subsumed versions, the text was broken down into four discrete “chunks”, serving as extended captions. Each text passage required an image as entry point, as per the definition of the prose subsumed strategy. This meant that new imagery had to be located or created for one or more of the passages—the original designs, being prose primary, did not have ample imagery for a prose subsumed strategy. Text was shortened as possible in cases where new imagery rendered some text truly redundant. However, this was possible only on rare occasions. Table 3.5.1 shows the word count of the alternate forms for each treatment session. Text reduction for prose subsumed was minimal for the first two treatments (30 and 9 words eliminated, respectively), while the final treatment enjoyed no reduction.

Table 3.5.1: Word count by form

<table>
<thead>
<tr>
<th>Form Version</th>
<th>Session #1</th>
<th>Session #2</th>
<th>Session #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prose Primary (PP)</td>
<td>517</td>
<td>633</td>
<td>617</td>
</tr>
<tr>
<td>Prose Subsumed (PS)</td>
<td>487</td>
<td>624</td>
<td>617</td>
</tr>
<tr>
<td>Fully Integrated (FI)</td>
<td>315</td>
<td>435</td>
<td>492</td>
</tr>
</tbody>
</table>

Fully integrated forms are the most demanding of a designer and were always the last completed. The fully integrated strategy requires the deconstruction of prose, resulting in small chunks of text never longer than a single sentence. Fully integrated versions involve far more thorough conversion of text into image, wherever possible. So, for instance, while a
prose primary version may describe through prose a raft’s path down a long mountain river, that river and raft are easily converted into imagery with minimal textual chunks pointing out critical information.

The following textual passage uses 13 words: “Then the raft reaches a lake. You glide across the surface, slowing down” (McDougal Littell, 2005a). Using a sequence of images to visualize a river’s widening and calming as it works away from mountainous areas (this conversion involved changing the lake to a river for demonstration purposes), in three panels (narrative imagery structure), the requisite text decreases to five words: “The river widens, slowing down” (see Figure 8.3.1). The rest of the textual information from the original passage is amply encoded in imagery. The fact that the river widens must still be pointed out in textual format in order to make the critical aspect salient to the reader.

Fully integrated versions ultimately equate with more dramatic reductions in word count and further increases of imagery than resulted from converting prose primary into prose subsumed. Text has varying levels of specificity, and so text cannot be reduced and converted into imagery at a fixed rate. The third treatment source pages had relatively denser text than the other treatments’ pages. Fully integrated word count as compared to prose primary word count for the third treatment was only 80%, while the first and second treatments were 61% and 69%, respectively.

During item development, the alternate forms were consulted in turn to ensure that items were solvable in each case.

Kathleen Meaney, an expert designer not intimately involved in the research project, served as consultant for form quality and equivalency. A standard design critique served as a framework for analyzing the forms. Meaney “redlined” the pages, indicating where information access was in any way problematic. This critique resulted in a final round of alterations before the forms were used in the study.
3.6  Data Collection Stages

3.6.1 Plan of Sessions

The four sessions conducted with the subjects, as well as pilot tests with undergraduates, are summarized in Figure 3.6.1. The treatment test duration values are actual—not expected—lower and upper quartiles (subjects completed the tests much more quickly than expected, though the slowest subjects took roughly 40 minutes).

<table>
<thead>
<tr>
<th>PRECEDING SESSIONS</th>
<th>MONDAY #1A</th>
<th>MONDAY #1B / FRIDAY #1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item development for treatment tests</td>
<td>Small PP form (1 p.)</td>
<td>7th grade review test (after pilot test)</td>
</tr>
<tr>
<td>Brief: ~12 min</td>
<td>Small treatment test (6 factual items)</td>
<td>Science self-efficacy inventory (before pilot test)</td>
</tr>
<tr>
<td></td>
<td>Difficulty rating</td>
<td></td>
</tr>
<tr>
<td></td>
<td>~20 minutes (w/ pre-test)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MONDAY #2</th>
<th>MONDAY #3</th>
<th>MONDAY #4A</th>
<th>MONDAY #4B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects</td>
<td>Subjects</td>
<td>Subjects</td>
<td>Subjects</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PP, PS, FI forms</td>
<td>PP, PS, FI forms</td>
<td>PP, PS, FI forms</td>
<td>Picture selection: most interesting (6 items)</td>
</tr>
<tr>
<td>Treatment test (12 items)</td>
<td>Treatment test (12 items)</td>
<td>Treatment test (12 items)</td>
<td>Picture selection: least interesting (6 items)</td>
</tr>
<tr>
<td>Interest level inventory (4 items)</td>
<td>Interest level inventory (4 items)</td>
<td>Interest level inventory (4 items)</td>
<td>~5 min.</td>
</tr>
<tr>
<td>Difficulty rating</td>
<td>Difficulty rating</td>
<td>Difficulty rating</td>
<td></td>
</tr>
<tr>
<td>Difficulty explanation (open-ended)</td>
<td>Difficulty explanation (open-ended)</td>
<td>Difficulty explanation (open-ended)</td>
<td></td>
</tr>
<tr>
<td>11–18 minutes</td>
<td>10–16 minutes</td>
<td>10–16 minutes</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.6.1: Sessions plan
3.6.2 Instrumentation

Each classroom in the middle school has a full set of internet-ready laptops. Online survey instruments were created using Survey Gizmo, a robust data collection application. For each treatment session, the teacher provided a link to the day’s online survey.

Survey Gizmo allows for randomization of items. Each subject completed the same multiple-choice items, but the order of these items was random, so subjects received the items in different orders from one another. For most multiple-choice items the order of the key and distracters (the item options) was also randomly shuffled for each subject.

Items were organized into pages. To complete an item, the subject clicks on a button beside the desired answer. Only one choice may be selected for an individual item, so if the subject clicks on a second choice, this one becomes selected while the previous option is deselected. The subject cannot continue on to the next page until all items on the current page are completed.

3.6.3 Pre-Existing Data

The school maintains records on all students’ end-of-grade test scores (EOGs), but acquiring this information would have complicated matters with N. C. State University’s Institutional Review Board and Wake County Public School System’s Research Review Committee, in addition to possibly decreasing informed consent totals. Attaining these scores, desirable though they may seem, was deemed beyond the scope of this doctoral study. Instead, a brief 7th-grade review test was included in the pre-test.

Ultimately, the only pre-existing data collected for the study was sex. Sex was not expected to influence scores.
3.6.4  Pilot Tests with Undergraduates

Two types of pilot tests were conducted for this study. In the first, undergraduate students completed early versions of the instruments. Undergraduates are not proper surrogates for middle school students, but because they are so much more advanced they can serve as good indicators for items that are too difficult or constructed poorly.

During the course of the study I was teaching a course on Web design. I had my students complete early versions of the test instruments. These students were expected to attain near 100% success on the tests, so any deviation from that became a point of scrutiny in item development, indicating conceptual issues, unfair distracters, or confusing language.

3.6.5  Pilot Test with the Subjects

A second type of pilot test was built into the pre-test session. As part of the pre-test instrument, students completed six multiple-choice items using a page from the 8th grade textbook. This is the same basic design as the treatment sessions. Subjects received the page as it appeared in the textbook, in prose primary (PP) form, and completed the battery of content items. Due to extenuating circumstances at the middle school, only half of the students—students of one of the two teachers—took the pilot test portion of the pre-test (the others ultimately had less time available that week and were only administered the most critical components of the first session). This meant that scores on the pilot test could not be used as a baseline comparison for performance on the treatment tests. However, the six items in this pilot still served as an indicator to the functional difficulty of items for this population of students at the 7th grade level. Success rates on these items helped in development of the treatment test items by informal comparison. It became more reasonable to predict the difficulty level of the treatment test items during development.
3.6.6 **Pre-Test**

The pre-test was administered in the first of four sessions. Sessions were scheduled on Mondays of successive weeks. Class periods are 50 minutes. The pre-test was intended to cover around 15 minutes of class, after which subjects were to work on their own. The pre-test included: the pilot test (discussed above); a review test, meant to serve as an indicator of performance in science class; and a science self-efficacy inventory. These components were built into the online survey instrument. The form was a letter-size printout reproduction from a page in the textbook on deep ocean hydrothermal vents. Items for this pilot portion of the test were developed in the same manner as the main items in the treatment tests.

3.6.7 **Treatment Tests**

The treatment tests were expected to take around 30 minutes of class time on average, leaving an additional 20 minutes for the slower subjects. Many subjects finished the exercises much faster than expected.

The treatment tests took the form of online surveys completed on school-provided laptops. Students had the handout forms on their desks in addition to the laptops. Each form is a tabloid sheet folded in half to regular letter size. The front and back covers are blank except for a form code on the front cover: PPV, PSV or FFV—meaningless to the subjects, but indicating the text–image integration strategy (“V” stands for “version”).

The components of the survey follow, in the order in which they appeared:

1) First and last name.
2) Form code copied from cover of form, to ensure that subjects had received the correct form.
3) A message that subjects should not proceed to the next page until they have finished reading the handout (form).
4) Six factual-level test items relating to the form, in randomized order as a single page.
5) Six conceptual-level test items relating to the form, in randomized order as a single page.
6) A forced-choice Likert scale item assessing the difficulty of the task just completed.
7) A follow-up constructed response item where subjects are asked to explain why the activity was however difficult they claimed it was.
8) Four Likert scales assessing how interested subjects were in the form’s content.

3.6.8 Post-Test

The post-test was administered in the final session, immediately following the third treatment. Up until this point the subjects were never engaged to consider the design of the forms directly. The post-test was a means to get explicit responses to the designs after such time as doing so could poison the main treatment tests.

The post-test included 12 items wherein subjects selected a most interesting (first 6 items) or least interesting (remaining 6 items) page design from a selection of three apiece. The page designs were single pages from the test forms themselves, with direct comparison of the same pages across the three treatment types. This is conceived as a direct measure of interest level, whereas the treatment tests included indirect measures.
3.7 Data Collection Components

3.7.1 Performance Item Development

The main treatment test items were developed through analysis and pilot testing with undergraduate students. Multiple-choice format was selected as a quick method of data collection, allowing for numerous items in addition to other test components in the time available. Each item must require engagement with the treatment form (textbook pages).

There was an upward limit on how many choices could be developed for each item. A typical multiple-choice item includes one key (the correct or best answer) and some number of distracters (incorrect answers, meant to appear correct when a test taker does not hold the correct conceptualization). The more distracters, the less likely a random guess will be correct. Items tend to get more difficult with more distracters. High choice-count items may prove overwhelming for test takers. With each additional choice, it became dramatically more difficult in item development to maintain distracter integrity, especially since the test forms are only two textbook pages—there is only so much content to work with. The four-choice item (a key and three distracters) was quickly settled on as most reasonable.

In a few cases end-of-chapter and end-of-section questions provided readymade items for use in the test instrument. But most items had to be developed from scratch. Care was taken to ensure that the answers to items—or the information helpful in answering items—were more or less evenly distributed across the form pages. That is, one localized area (the second of four major prose sections in a prose subsumed design, for instance) should not contain a disproportionate amount of critical information. Each battery of items should in total require complete engagement with the instructional media.

Items are considered to measure factual knowledge construction when they can be answered using information found at a single location on the page, or where the answer is otherwise not problematic, but rather straightforward. A factual item may very well require reasoning rather than simple locating, but its target localization within the instructional media
represents the low level of abstraction that suggests factual information in Bloom’s Revised (Anderson & Krathwohl, 2001).

Items are considered to measure conceptual knowledge construction when the information required is found in more than one distinct location in the page design, or where the answer is problematic and in some way debatable. To complete a conceptual level item, the test taker must connect concepts. This requirement is consistent with the more complex and organized conceptual knowledge outlined in Bloom’s Revised (ibid.). Conceptual level items are expected to be inherently more difficult than factual level items, though this is a generality—individual factual items can certainly be more difficult than individual conceptual items. Committee member James Minogue, an expert on science education, was consulted to determine if the factual–conceptual distinction appeared functional.

Item development began after the forms were completed. For each item, it was confirmed that the answer was accessible across all three forms. Six to eight items apiece were initially developed for factual and conceptual sets. The 6–8 items were reduced to the final six items after pilot testing with undergraduate students. Often enough items had to be removed that altogether new items were developed for the final treatment tests. Following is an example of a conceptual level item that was altered following the pilot test, from the second treatment session, on lakes and lake turnover. The targeted forms can be found in the appendix (Figure 8.3.4 is the targeted form most similar to the original textbook design). Brackets indicate the number of responses in the pilot test (out of twelve). The asterisk identifies the key. (Note that for all multiple-choice items, the choices themselves were randomly ordered per subject.)

- When does lake turnover occur?
  - spring and fall [10] *
  - once every season [1]
  - once every year [1]
  - when a lake is frozen [0]
The optimal difficulty level for a 4-choice item is 0.74 (74% of responses correct) (Reynolds, Livingston, & Willson, 2006, p. 144). A range of difficulty across items from 0.64 to 0.84 would be ideal for the treatment tests (ibid.), but pilot tests with middle school students is beyond the scope of this study. It was necessary to estimate how undergraduate scores might translate to final test scores. The sample item above had a difficulty level of 0.84 in the pilot test, within the ideal difficulty range—but this value was attained with undergraduates, not middle school students. Any incorrect answers in the undergraduate pilot study were carefully scrutinized, given the dramatic difference in capability between 7th-grade students and undergraduates.

The answer is not directly found in the text. It is stated that turnover occurs twice a year, but only fall is ever pictured. The two pilot test–selected distracters were reconsidered before the treatment test was finalized:

1) The second choice was replaced with a more obvious distracter. Students might not be certain what counts as a season. Are summer and winter the only proper seasons with spring and fall as mere transitions? Some might consider this to be the case, and the answer isn’t easy enough to distill from the pages provided.

2) The third choice, though also selected once in the pilot, was not changed. It was determined to be a fair distracter, because it is explicitly stated that turnover occurs twice—not once—every year.

3) Because a distracter that was successful in the pilot test was retained without alteration, the replacement second choice—“in the summer”—was designed to be a more obvious incorrect answer. A test taker need only look at the summer illustration, and perhaps the fall illustration that outlines turnover, to realize that turnover does not occur in summer.

Finally, after the advice of Reynolds et al. (2006), each battery of test items was developed over a period of time allowing the developer to step back from and return to items
with a clear head (pp. 158–159). The bulk of alterations came from this simple method. It is far too easy to misjudge the logic of an item in development. By returning to it with new eyes, the developer is better able to see it as a test taker would see it, and thus judge it adequately.

Following is the final version of the sample item. Brackets indicate the percentage of responses in the actual treatment test. The asterisk identifies the key.

- When does lake turnover occur?
  - spring and fall [70%] *
  - in the summer [11%]
  - once every year [15%]
  - when a lake is frozen [4%]

The second choice, intended to be a weak distracter, still received 11% of responses. This item thus has a final difficulty level of .70, within the ideal range of .64–84.

3.7.2 Task Difficulty Assessments

The task difficulty assessment immediately follows the comprehension items (the “task”). The item is a forced-response 6-point Likert scale, scored in the range of 1–6, with 6 suggesting a very difficult task. The stem reads: “How difficult was this exercise?” Options are: very easy, easy, somewhat easy, somewhat difficult, difficult, and very difficult (Figure 3.7.1). The forced-response format was used because subjects were subsequently asked why the task was however difficult they deemed it to be. Foregoing a neutral response, such as “neither easy nor difficult”, promotes forming an opinion, which is necessary if the subsequent constructed response is to prove insightful.
The interface of the initial difficulty item requires the subject to select one option by clicking a “radio button”. If a second option is clicked—as in reconsideration—the original option is automatically deselected. Once an option is selected, the subject can continue to the next page of the survey, which presents the constructed response difficulty explanation (Figure 3.7.2). The stem text incorporates the subject’s previous answer. So a subject who had selected “somewhat difficult” for the preceding difficulty scale is prompted with: “In your own words, why did you find working with the handout to be somewhat difficult?” This function is referred to as “answer piping” in the Survey Gizmo software. An open form field is paired with the question, and subjects must type something in to proceed. There is no set limit to the number of characters that the form field will accept. A response of “!” (a single character) would be counted as complete by the system.
The open-ended difficulty response is the sole qualitative component of this study. It is meant to serve as a platform on which subjects can provide unexpected insight into their experience of working with the treatment forms.

3.7.3 Indirect Interest Level Assessments

An indirect interest level inventory followed the difficulty assessments on each treatment test. This section concluded each treatment test. The indirect interest level inventory inquires into the subject’s interest in the treatment form’s content, not the design itself. The design (the text–image integration strategy employed) mediates interest in subject matter.

The inventory is adapted from Schraw’s (1997) perceived interest questionnaire, an inventory used to gauge situational interest. Items were altered to include the name of each given treatment’s subject area (“divides and drainage basins”, etc.). The inventory was also reduced to four items, as the source inventory was too extensive for this particular use.

For each treatment the same four items were included, though again the object of each sentence reflected the treatment form’s subject matter. The first treatment session’s indirect interest level inventory items are:

1) I would read more about divides and drainage basins if I had the chance.
2) I thought the “divides and drainage basins” topic was fascinating.
3) Divides and drainage basins really grabbed my attention.
4) I got absorbed reading about divides and drainage basins without trying to.

The order of the items was randomized for each subject. The inventory was set to 5-point Likert scales, scored individually in the range of 1–5, for an aggregate score range of 4–20, with 20 representing the highest level of interest. Options are: strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree (Figure 3.7.3).
Figure 3.7.3: Indirect interest level item

In the interface, the subject clicks on one option for each item. If a second option is selected the original option is automatically deselected. Subjects could not proceed to the next page in the survey before completing the full inventory.

3.7.4 Direct Interest Level Assessments

The three versions of each page from the study are shown at thumbnail size together for the first six items. Students determined which one of the three was the most interesting by clicking on the corresponding image (Figure 3.7.4). After these six items are completed, a new page of items can be accessed. The selfsame image sets are repeated for the final six items, for which students select the least interesting of each set of three. The proctor alerted subjects to the difference between the two pages of items before the test began. Software randomized the order of the six items within each page, as well as the order of the three images for individual items.

The online test instrument tracked each individual selection. These selections also produced scores for each of the three treatment types. A selection of most interesting contributed one point to the aggregate score for the treatment type. For each item, one type must be selected. This means that a total of 6 points are awarded. It is possible for one type to receive all 6 points, though if this happens then the other two necessarily received none. A selection of least interesting on the following page deducted one point from the aggregate score for the treatment type. Due to a bug, one set of page designs—one item apiece for the most and least interesting sets—failed to display images. These items had to be removed.
from analysis, meaning that a maximum of 5 points could be awarded or deducted from one treatment type’s aggregate scores. Final aggregate scores thus start at 5 before calculations and after have a range of 0–10.

![Figure 3.7.4: Direct interest level item (randomized PP, FI, PS)](image)

### 3.7.5 Review Test

Twenty-two end-of-chapter multiple-choice items were selected from throughout the 7th-grade textbook. These items cover a range of subject matter. The participating teachers reviewed these 22 possible items, indicating which items they had covered in class that year. The list was then cut down to ten review items, all of which had been covered by both teachers.

Figure 3.7.5 is a screen capture from the test instrument. The four choices are randomly ordered. In this example and in this order, the second choice is correct. Items are randomly ordered as well, meaning this is not always item #18. (Earlier numbered items included the science self-efficacy inventory and pilot-test items.) All review items can be found in the appendix.
Each answer was tracked and correct answers (selection of the key) contributed 1 point to an aggregate review test score. Incorrect answers (selection of one of three distracters) were scored as 0. Total scores thus range from 0–10.

The review test is meant as a baseline measure of success in science class. It notably involves learning, which the performance tests of this study do not. Review test scores may provide insight into treatment test performance.

3.7.6 Science Self-Efficacy

Science self-efficacy was measured in the pre-test using Nietfeld, Cao, and Osborne’s (2006) efficacy scale, where it was shown to be internally reliable and predictive of academic performance. This is a second type of baseline measure, of particular interest in comparison with the difficulty assessments and interest level inventory results. This measure of self-efficacy might also provide insight into review test and treatment test scores. The complete inventory is:

1) I am sure that I can learn science.
2) I can get a good grade in science.
3) I am sure I could do high school level work in science.
4) I have a lot of self-confidence when it comes to science.
5) I am not the type to do well in science.
6) It takes me a long time to catch on to new topics in science.
7) Even before I begin a new topic in science, I feel confident I’ll be able to understand it.
8) I think I have good skills and strategies to learn science. (Nietfeld, Cao, & Osborne, 2006)

The inventory was set to 5-point Likert scales, scored individually in the range of 1–5, for an aggregate score range of 8–40, with 40 representing the highest level of science self-efficacy. Options are: strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree. Note that items 5 and 6 are reverse coded. The online survey randomized the order of the eight items.

Figure 3.7.6 shows the actual interface. It functions in the same manner as the indirect interest level inventory. Subjects could not proceed to the next page in the survey before completing the full inventory.

Figure 3.7.6: Science self-efficacy item

3.7.7 Test Duration

The duration of each subject’s engagement with the test instrument was carefully documented. However, duration was only measured for the entire instrument, not individual components such as reading or completing individual item sets.
3.8 Administration

3.8.1 Protocols

A graduate assistant—Rachael Huston—helped the researcher proctor sessions. For each session the researcher stayed in one teacher’s classroom while the assistant stayed in the other. Researcher and assistant switched teachers with each session. There were four sessions on consecutive Mondays.

Proctors began each session by plainly describing the day’s task to the subjects. Subjects logged on to classroom laptops and followed the day’s link to the online test instrument. While subjects signed in, proctor and teacher passed out forms. Initially forms were distributed from a randomized stack. Subsequently, specific forms were distributed to students based on a form key held by the proctor and teacher.

The proctor ensured that subjects did not open the forms until all were passed out, and noted the time as the entire class opened the forms simultaneously. (The online survey system recorded the completion time for each subject individually.) The proctor then monitored subjects to keep them from speaking with one another or navigating the Internet. The proctor also assisted any subjects who had difficulty accessing the test instrument.

Under no circumstances were subjects to be helped with the treatment test items themselves. Any inappropriate inquires were answered with “Just do the best you can.” The proctor noted any behavior that might poison a subject’s results, and in such rare cases made a descriptive note. As subjects completed the task, they returned their forms to the proctor, who checked them for marks or damage and placed them in the proper folders for redistribution in the following class.
3.9  

**Item Analysis**

### 3.9.1  

**Methods**

Item analysis was conducted according to strict protocols (Figure 3.9.1). Each of the following methods detailed in this section was employed in turn. Item difficulty and item discrimination indices were calculated for all items. Item-total correlation, a second look at item discrimination, was calculated for items and compared to earlier discrimination values. Any problematic items were finally subjected to distracter analysis.

**3.9.1.1  
Item Difficulty Index**

The item difficulty index is the proportion of subjects who correctly selected the item key. Reynolds et al. (2006, p. 144) suggest an optimal difficulty index for 4-option selected-response items of mean success across items at 74%. While extremely high or low values can indicate a bad item, variation in difficulty around 0.74 (0.64–0.84 at the least) can be desirable. The discrimination index helps to determine if more extreme values are acceptable. Reynolds et al. use *p* to indicate difficulty index.

![Figure 3.9.1: Item analysis protocols](image-url)
The item discrimination index, indicated with $D$, measures how well an individual item discriminates between high and low performers. Reynolds et al. (2006, p. 145) suggest many acceptable ways of distinguishing high and low performers. Consistent with one of their recommendations, this study defines the high performers as the top 33% and the low performers as the bottom 33%. Subjects are classified as high or low performers based on their overall scores in a test battery. It is likely that to perform the 33% distinction a number of subjects with the same test score will have to be separated into high and medium performers or low and medium performers. In the case that this must occur, the distinction can be made by random chance. For this study, statistical software (JMP) was used to subtract or add a small amount to test scores within the range of plus or minus 0.125. This randomized adjustment ensures that higher test scores are always ranked higher despite randomization. These randomly adjusted scores are only used to assign subjects to high, medium and low performing groups.

Once subjects have been assigned to high and low performing groups, individual items in that battery are then analyzed according to that distinction. The low performer (bottom) mean for an item ($p_B$) is subtracted from the high performer (top) mean ($p_T$):

$$D = p_T - p_B$$

This results in a discrimination index ($D$) from −1.0 (all low performers correct and all high performers incorrect) to 0.0 (no discrimination whatsoever) to 1.0 (perfect discrimination with all high performers correct and all low performers incorrect).

The difficulty index of an item constrains its discrimination index, however. An item that 1 of 9 subjects answer correctly, with its resulting difficulty index ($p$) of 0.11, can have at best a discrimination index ($D$) of 0.33. This would indicate that the one subject that answered the question correctly was in the high performing group. Table 3.9.1, adapted from an equivalent table in Reynolds et al. (2006, p. 147), lists the constraints on discrimination...
index based on difficulty index for this study’s 33% method of determining subject performance level. Any items outside of the approximate difficulty index range of 0.33 to 0.67 have constraints placed on the discrimination index, meaning that lower discrimination must be considered along with difficulty index. As a rule of thumb, Reynolds et al. (2006) suggest that items with $D$ values over 0.30 are “acceptable”, while those with lower discrimination should be reviewed (p. 146). Values of 0.40 and higher are said to have excellent discrimination power, while values of 0.30 to 0.39 are good, values of 0.11 to 0.29 are fair, and values at or below 0.10 are poor (ibid.). Any item that is assessed as “poor” according to these standards will be removed from the study.

Table 3.9.1: Difficulty constraints on discrimination with 33% method

<table>
<thead>
<tr>
<th>Item Difficulty Index ($p$)</th>
<th>Maximum Discr. ($D$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0.89</td>
<td>0.33</td>
</tr>
<tr>
<td>0.78</td>
<td>0.66</td>
</tr>
<tr>
<td>0.67</td>
<td>1.00</td>
</tr>
<tr>
<td>0.33</td>
<td>1.00</td>
</tr>
<tr>
<td>0.22</td>
<td>0.66</td>
</tr>
<tr>
<td>0.11</td>
<td>0.33</td>
</tr>
<tr>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

3.9.1.3 Discrimination Using Item-Total Correlation Coefficients

An alternative approach to gauging item discrimination is item-total correlation. Individual items are correlated to the full test battery. For this study item-total correlation will be unadjusted, meaning that the individual item being addressed will be embodied in the total test score (Reynolds et al., 2006, pp. 147–149). Scores are interpreted similarly to the
item discrimination index, in the sense that, for instance, values at 0.40 and above are considered excellent discriminators (p. 149). Item-total correlation coefficients are being calculated as a supplement to the favored item discrimination index only when they can provide insight into item discrimination indices that might be underestimated due to lower difficulty indices (where the item discrimination index is rated below “good” but the item difficulty index is very high or low).

Any items that are poor discriminators will be removed from the study. Furthermore, any items that rate as “fair” according to both measures of item discrimination are considered suspect and will be removed as well. This is done because both methods of calculating item discrimination may inflate values slightly: the 33% high–low determination in the item discrimination index removes the middle group from calculations, rendering differences more dramatic, and item-total correlation coefficients being unadjusted allows each individual item to affect its own item-total calculation.

3.9.1.4 Distracter Analysis

Distracter analysis is a quantitative method for determining how well an item is functioning (Reynolds et al., 2006, pp. 151–152). Distracters are the incorrect choices in a selected-response item, and they are so named because they are meant to appear as viable options for those who don’t have the requisite knowledge, and as incorrect answers to those that are sufficiently knowledgeable. Individual distracters should have negative discrimination, meaning that low performers should select them more often than high performers. Furthermore, all distracters should seem like viable answers to the uninformed, so distracters that garner no selections are poor and should be replaced.

Distracter analysis is primarily a method for test development. After one use of a test, items can be rewritten so that distracters perform better in subsequent administration. However, this is not the goal of the current study, as its tests are for one-time use. Distracter analysis would be most useful during pilot testing, but there was no opportunity to pilot test
the instruments in this study with truly equivalent surrogates—undergraduates were used for this purpose.

Ultimately, distracter analysis will only be conducted here for items that appear suspect after the previous item analyses, for insight into performance and as justification for exclusion.

3.9.2 Review Test Items

The sample for the review test is \( N=148 \) subjects.

3.9.2.1 Item Difficulty Index for Review Test

The mean \( p \) value (difficulty index) across the pre-test review items was 0.67, near the 4-choice ideal of 0.74 (Reynolds et al., 2006, p. 144). Individual \( p \) values ranged wildly from 0.11 to 0.92 (see Table 3.9.2). Review item #6 was the extreme low value. Six of ten items fell in the range of 0.60 to 0.90. As such the review items were overall slightly more difficult than is ideal.

Review item #6 begs for scrutiny, as its \( p \) value is lower than one would expect even from random guessing (any difficulty index under 0.25 stands out especially for this reason). The stem of this item had been changed from the textbook’s original language to make it clearer. The original stem was “Genes are sequences of DNA, which are made up of” (McDougal Littell, 2005b, p. C157). “Which” in the stem should refer to “genes,” as the parenthetical builds on a plural. However, this distinction was expected to be far too subtle for middle school students (and perhaps everyone else). The final item, with selection rates in brackets and key (correct answer) indicated with an asterisk, follows. Note that a misreading of the original stem would indicate the incorrect answer “chromosomes” as an equal choice to the key “nucleotides.” (Options were randomly ordered in test.)
• Genes (sequences of DNA) are made up of:
  o nucleotides [11.6%] *
  o chromosomes [85.6%]
  o phosphates [2.1%; two responses]
  o ribosomes [0.7%; one response]

This is the only item where a distracter was selected more often than the key. It is also extraordinary in that two of the distracters together account for only 3 responses out of 148. “Chromosomes” was probably selected at such a high rate in part because that is the most recognizable of the four terms. Furthermore, all a subject had to do to believe he or she had found the right answer in “chromosomes” was to misread the directionality of the stem. Chromosomes are made up of genes, or “sequences of DNA”, but careful reading reveals that the item concerns that which is smaller and a part of genes, not that which genes are smaller than and a part of.

3.9.2.2 Item Discrimination Index for Review Test

Review test items ranged in item discrimination index \((D)\) from 0.24 (in 3 items) to 0.51 (Table 3.9.2). Five of the 10 items have excellent discrimination power. Two more have good discrimination power. Only 3 of the 10 items fall below the “good” threshold and beg some level of review (Reynolds et al., 2006, p. 246). These are all considered to have fair discrimination power, all at 0.24. These “fair” items have the two highest and the lowest difficulty indices in the test battery \((p = 0.11, 0.89 \text{ and } 0.92)\). This means that each of these items was constrained to an approximate maximum \(D\) value of 0.33, putting into context the designation of “fair discrimination power”. Under these circumstances, each of these items appears to discriminate acceptably. Review item #6, for instance, which was previously discussed due to its remarkably high difficulty \((p = 0.11)\), while only answered correctly by 27% of high performers, was all the more dramatically answered correctly by only 1 of the
low performers (2%). The other two items with fair discrimination power were answered correctly by all of the high performers (100%), and by 76% of the low performers.

There is no evidence in the item discrimination indices suggesting that any of the review test items should be removed from the study.

### 3.9.2.3 Discrimination Using Item-Total Correlation Coefficients for Review Test

Table 3.9.3 shows item-total correlation coefficients for the review test battery. Eight of the 10 items are considered to have excellent discrimination power, being above 0.39, while the remaining items are in the upper range of “good” discrimination. The three items that were rated “fair” for the item discrimination index are either “good” (two of them) or “excellent” here. This helps to quantify the confidence in these items despite the “fair” rating for the item discrimination index. According to the standards of this study, no review items will be removed due to discrimination measures.
Table 3.9.3: Review test item-total correlation

<table>
<thead>
<tr>
<th>Review Test</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review #1</td>
<td>0.5208</td>
</tr>
<tr>
<td>Review #2</td>
<td>0.4173</td>
</tr>
<tr>
<td>Review #3</td>
<td>0.3902</td>
</tr>
<tr>
<td>Review #4</td>
<td>0.4004</td>
</tr>
<tr>
<td>Review #5</td>
<td>0.4117</td>
</tr>
<tr>
<td>Review #6</td>
<td>0.3515</td>
</tr>
<tr>
<td>Review #7</td>
<td>0.5430</td>
</tr>
<tr>
<td>Review #8</td>
<td>0.5105</td>
</tr>
<tr>
<td>Review #9</td>
<td>0.4499</td>
</tr>
<tr>
<td>Review #10</td>
<td>0.4632</td>
</tr>
</tbody>
</table>

3.9.2.4  
**Distracter Analysis for Review Test**

All items in the review test appear to discriminate well, so distracter analysis need not be performed as a rule. Review item #6 is curious, though, as previously noted, and is addressed here.

Table 3.9.4 shows the differential selections of the high and low performers. While this item is certainly abnormal, nothing suggests that it fails to measure the construct. The most common selection is a distracter, but more low performers selected it than did high performers. More high performers selected the key than low performers. Indeed only one of the latter successfully selected the key. Only low performers (3) selected the second distracter. The third distracter appears to have had no effect. In normal item development, this would be a problem to be remedied, but since the first distracter is so strong, and the item still discriminates well, it is fine here.
3.9.3 Pilot Performance Test Items

The pilot performance test was intended for all subjects, but unforeseen circumstances at the middle school during the first week of the study made it necessary to remove this portion (and this portion only) of the pre-test for all students of teacher B.

The pilot performance test consists of 6 factual items addressing a single page handout on deep ocean hydrothermal vents. The treatment tests each address a two-page handout, which in practice is far more than double the complexity. On a two-page handout, requisite information is not nearly as easy to find, and it becomes possible to model higher-level interaction. Conceptual items would be far more difficult to develop for such an abbreviated text.

Because only half of the subjects participated in this component of the pre-test session, the pilot performance test could only serve to aid in item development by helping to predict item difficulty.

With only 6 items in the test, neither measure of discrimination can be very meaningful. In both cases the total number of items is directly tied to accuracy. Besides, there is no purpose in interrogating the pilot test items given their limited efficacy. Only the item difficulty indices will be addressed for the pilot treatment test.

The sample for the pilot treatment test is \( N=76 \) subjects.

<table>
<thead>
<tr>
<th>Review Item #6</th>
<th>Key</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>High performers (top 33%)</td>
<td>13</td>
<td>35</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Low performers (bottom 33%)</td>
<td>1</td>
<td>45</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
3.9.3.1 Item Difficulty Index for Pilot Performance Test

The pilot test items have a mean $p$ value of 0.78, just slightly easier than the ideal 0.74. The items themselves have a narrow range of 0.75 to 0.83. No item stands out in terms of difficulty.

3.9.4 Treatment #1 Performance Test Items

The sample for treatment session #1 is $N=146$ subjects. The treatment performance tests measure comprehension, where the review test measured content mastery, or learning.

3.9.4.1 Item Difficulty Index for Treatment #1 Performance Test

The mean $p$ value (difficulty index) across the treatment #1 performance items is 0.54, below the 4-choice ideal of 0.74 (Reynolds et al., 2006, p. 144). Individual $p$ values range from 0.36 to 0.75 (Table 3.9.5). The highest $p$ value in the set is near the target mean, illustrating how difficult this battery was in practice.

<table>
<thead>
<tr>
<th>Treatment #1</th>
<th>Difficulty ($p$)</th>
<th>Treatment #1</th>
<th>Difficulty ($p$)</th>
<th>Treatment #1</th>
<th>Difficulty ($p$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual #1.1</td>
<td>0.6575</td>
<td>Conceptual #1.1</td>
<td>0.3699</td>
<td>Factual #1 Set</td>
<td>0.6416</td>
</tr>
<tr>
<td>Factual #1.2</td>
<td>0.6507</td>
<td>Conceptual #1.2</td>
<td>0.3973</td>
<td>Conceptual #1 Set</td>
<td>0.4475</td>
</tr>
<tr>
<td>Factual #1.3</td>
<td>0.6301</td>
<td>Conceptual #1.3</td>
<td>0.3562</td>
<td>Full #1 Set</td>
<td>0.5445</td>
</tr>
<tr>
<td>Factual #1.4</td>
<td>0.5411</td>
<td>Conceptual #1.4</td>
<td>0.5274</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factual #1.5</td>
<td>0.6164</td>
<td>Conceptual #1.5</td>
<td>0.4795</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factual #1.6</td>
<td>0.7534</td>
<td>Conceptual #1.6</td>
<td>0.5548</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.9.5: Treatment #1 difficulty
The factual item set has a mean $p$ value of 0.64, with items ranging from 0.54 to 0.75. The conceptual item set was more difficult, with a mean $p$ value of 0.45 and items ranging from 0.36 to 0.55. It is prudent to inspect the more difficult items.

Conceptual item #1.1 was answered correctly by 37% of subjects:

- How do bowl-shaped drainage basins differ from other drainage basins?
  - Bowl-shaped drainage basins are surrounded on all sides by divides. [0.37] *
  - Bowl-shaped drainage basins drain water at their centers. [0.27]
  - Only bowl-shaped drainage basins contain lakes, while other drainage basins contain rivers. [0.27]
  - Bowl-shaped drainage basins are flatter than other drainage basins. [0.09]

No distracter is more powerful than the key, though two of them are only 10% lower in response rate. Conceptual-level items require that the reader put together pieces of information found in more than one location on the page. These items also involve reasoning beyond simply finding the correct information.

Upon review, this item is suspect. The PP version only mentions bowl-shaped drainage basins once: “In a bowl-shaped basin, the water may collect at the bottom of the basin or evaporate” (McDougal Littell, 2005a, p. C17). This follows a sentence: “In most places, the water eventually flows to the sea.” Thus, the bowl-shaped basin is held in contrast to other basins that flow out. The illustration on the page does not show any bowl-shaped basins, but does represent the flow out of regular basins, which always have one side “open”. Finally, a sentence that precedes the others describes what happens in a bathroom sink, though it is not explicitly referred to as being “bowl-shaped”. The FI version includes an image of a sink, above the sink sentence, with the bowl-shaped sentence placed nearby and pointing to it.

The first distracter is wily, as nowhere is it stated that water doesn’t “drain down” geographically—it must be reasoned through. To avoid this distracter the subject must
recognize the limits of the metaphor. The second distracter, selected at the same rate as the first (27%), is certainly fair and can be eliminated without too much difficulty by a subject, as can the third and less popular distracter. Conceptual item #1.1 is a candidate for exclusion if it fails to discriminate.

Conceptual item #1.2 was answered correctly by only 40% of subjects:

- What is NOT a possible path for a drop of water to take?
  - flowing by river out of a bowl-shaped drainage basin and into the ocean [0.40] *
  - flowing swiftly down a mountain side and then slowing near the ocean [0.29]
  - falling from a cloud directly into a lake [0.12]
  - falling from a cloud onto a steep ridge and then down into a valley [0.18]

Strong emphasis (“NOT”) was added to the stem in development because this is an elimination question—much of its difficulty is likely a product of this. Hasty item inspection followed by failure to scrutinize all choices would lead to a subject selecting the first distracter found (as with all items they were randomly ordered per subject), as it would mistakenly meet the conditions of a misread stem asking for a possible path of water.

This item is the second to consider bowl-shaped drainage basins in the stem–key relationship. The first and third distracters are both described in some manner in the forms, through text, image, or both. The second distracter is the only one never directly addressed, but it should be clear that rain can fall on a lake (the imagery only shows rain falling on a mountain ridge). Since this proved to be the weakest distracter, despite seeming less obvious than the other distracters, it might suggest that the difficulty level does indeed largely derive from the negative wording in the stem. This item does not appear to be a candidate for exclusion.

Conceptual item #1.3 was answered correctly by only 36% of subjects:
- You are lost in the wild and need water. It hasn’t rained in weeks. Where would you expect to find water, and why?
  - Further away from a divide, because gravity pulls it down. [0.36] *
  - Nearest a divide, because rainwater falls there first. [0.27]
  - Nearest a divide, because divides trap water. [0.32]
  - Further away from a divide, though it will be salt water and undrinkable. [0.05]

This item is a scenario that requires some effort of imagination as well as problem solving. It was expected to be among the most difficult items. All forms include the statement: “the force of gravity pulls water downhill.” The subject must fit this fact into a model that includes the passage of time since the last rain. However, upon inspection, this item appears unfair even before the distractors are addressed. The scenario aims to be deep—as scenarios do—yet the answer is overly simplified. Water might be near a divide for many reasons. Though a great deal of scrutiny may reveal the key to be the best answer of the lot, the distinction may be too subtle. The third distracter is easily avoided (though 8 subjects selected it), but the other two distracters might have passable arguments for their selection. This item is a strong candidate for exclusion, pending discrimination measures.

Conceptual item #1.5 was answered correctly by only 48% of subjects:

- What is true of the relationship between river current and divides?
  - Rivers might flow quickly near divides because the ground is steeper there. [0.48] *
  - Rivers might flow slowly near divides because divides contain lakes. [0.09]
  - Rivers might flow slowly near divides because rivers get wider there and the current changes. [0.23]
  - Rivers might flow quickly near divides because that is where rain falls. [0.20]
Here the strongest distracter is only half as successful as the key. The first distracter is just factually wrong. Divides are separators, not containers. The second distracter is the opposite of the river narrative found on the form (in text or text and image). This might have been selected so frequently because the language is so thick for this item. The third distracter is half true: the expectation tends to be that rivers flow quickly near divides. This expectation pairs the third distracter with the key. However, the distracter includes the false statement that the primary factor in river current is where rain falls, and not the slope of the land. This distinction certainly makes this a more difficult item, but there is little reason to consider exclusion here.

Scrutiny here has been reserved for items with difficulty indices below 0.50. Upon inspection, none of the less difficult items warrant protracted discussion here. All treatment instruments are reproduced in the appendix.

3.9.4.2 Item Discrimination Index for Treatment #1 Performance Test

Session #1 factual test items range in item discrimination index \((D)\) from 0.22 to 0.49, with the second-lowest value at 0.41 (Table 3.9.6). Five of the six items thus have excellent discrimination power, while factual item #1.5 has only fair discrimination power.

Session #1 conceptual test items range in item discrimination index from 0.41 to 0.59. All of the conceptual items have excellent discrimination power. Note that high and low performer groups were determined from the combined scores of the two sets. Calculating the performer groups for the sets individually would produce inflated scores, in part because the sets only number six items apiece.
Table 3.9.6: Treatment #1 discrimination

<table>
<thead>
<tr>
<th>Treatment #1</th>
<th>Mean ($p$)</th>
<th>High ($p_h$)</th>
<th>Low ($p_l$)</th>
<th>Discr. ($D$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual #1.1</td>
<td>0.6575</td>
<td>0.8367</td>
<td>0.4082</td>
<td>0.4286</td>
</tr>
<tr>
<td>Factual #1.2</td>
<td>0.6507</td>
<td>0.8367</td>
<td>0.4082</td>
<td>0.4286</td>
</tr>
<tr>
<td>Factual #1.3</td>
<td>0.6301</td>
<td>0.8980</td>
<td>0.4286</td>
<td>0.4694</td>
</tr>
<tr>
<td>Factual #1.4</td>
<td>0.5411</td>
<td>0.7347</td>
<td>0.3265</td>
<td>0.4082</td>
</tr>
<tr>
<td>Factual #1.5</td>
<td>0.6164</td>
<td>0.6939</td>
<td>0.4694</td>
<td>0.2245</td>
</tr>
<tr>
<td>Factual #1.6</td>
<td>0.7534</td>
<td>0.9796</td>
<td>0.4898</td>
<td>0.4898</td>
</tr>
<tr>
<td>Conceptual #1.1</td>
<td>0.3699</td>
<td>0.5714</td>
<td>0.1429</td>
<td>0.4286</td>
</tr>
<tr>
<td>Conceptual #1.2</td>
<td>0.3973</td>
<td>0.5918</td>
<td>0.1837</td>
<td>0.4082</td>
</tr>
<tr>
<td>Conceptual #1.3</td>
<td>0.3562</td>
<td>0.6939</td>
<td>0.1429</td>
<td>0.5510</td>
</tr>
<tr>
<td>Conceptual #1.4</td>
<td>0.5274</td>
<td>0.7959</td>
<td>0.2245</td>
<td>0.5714</td>
</tr>
<tr>
<td>Conceptual #1.5</td>
<td>0.4795</td>
<td>0.7551</td>
<td>0.2449</td>
<td>0.5102</td>
</tr>
<tr>
<td>Conceptual #1.6</td>
<td>0.5548</td>
<td>0.8980</td>
<td>0.3061</td>
<td>0.5918</td>
</tr>
</tbody>
</table>

High performers selected the key for factual item #1.5 69% of the time while low performers succeeded at 47%:

- Every divide:
  - pushes water in more than one direction. [0.62] *
  - pushes water in one direction. [0.20]
  - is in a flatter region. [0.04]
  - is a high mountain range. [0.14]

The most successful distracter is indirectly refuted in the text, which describes water as flowing in “different directions” from divides. The third distracter is likewise refuted: “Divides exist in mountain ranges as well as in flatter regions.” The item appears logical, but its fair discrimination power begs further review.
3.9.4.3 Discrimination Using Item-Total Correlation Coefficients for Treatment #1 Performance Test

Table 3.9.7 shows item-total correlation coefficients for the treatment #1 items. Eight of the 12 items are considered to have excellent discrimination power, being above 0.39, while 3 of the remaining items are in the range of “good” discrimination. Factual item #1.5 is again determined to have only fair discrimination power. As such, it is removed from further calculations, meaning that the factual item set for session #1 is reduced to 5 items, while the full battery for session #1 is reduced to 11 items.

<table>
<thead>
<tr>
<th>Treatment #1</th>
<th>R</th>
<th>Treatment #1</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual #1.1</td>
<td>0.3937</td>
<td>Conceptual #1.1</td>
<td>0.4177</td>
</tr>
<tr>
<td>Factual #1.2</td>
<td>0.4235</td>
<td>Conceptual #1.2</td>
<td>0.3780</td>
</tr>
<tr>
<td>Factual #1.3</td>
<td>0.4948</td>
<td>Conceptual #1.3</td>
<td>0.4726</td>
</tr>
<tr>
<td>Factual #1.4</td>
<td>0.3645</td>
<td>Conceptual #1.4</td>
<td>0.4840</td>
</tr>
<tr>
<td>Factual #1.5</td>
<td>0.2680</td>
<td>Conceptual #1.5</td>
<td>0.4659</td>
</tr>
<tr>
<td>Factual #1.6</td>
<td>0.5062</td>
<td>Conceptual #1.6</td>
<td>0.5511</td>
</tr>
</tbody>
</table>

3.9.4.4 Distracter Analysis for Treatment #1 Performance Test

Factual item #1.5 appears to function adequately according to distracter analysis (Table 3.9.8). More high performers selected the key, and more low performers selected each of the distracters. However, its discrimination power precipitated its removal.
Table 3.9.8: Factual #1.5 distracter analysis

<table>
<thead>
<tr>
<th>Factual Item #1.5</th>
<th>Key</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>High performers (top 33%)</td>
<td>34</td>
<td>10</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Low performers (bottom 33%)</td>
<td>23</td>
<td>13</td>
<td>4</td>
<td>9</td>
</tr>
</tbody>
</table>

Conceptual items #1.1 (Table 3.9.9) and #1.3 (Table 3.9.10), both flagged during difficulty analysis, share the same basic distracter profile, with high performers selecting the key more often and low performers selecting each distracter more often. There is thus no indication in distracter analysis to suggest removing either item.

Table 3.9.9: Conceptual #1.1 distracter analysis

<table>
<thead>
<tr>
<th>Conceptual Item #1.1</th>
<th>Key</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>High performers (top 33%)</td>
<td>28</td>
<td>11</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Low performers (bottom 33%)</td>
<td>7</td>
<td>19</td>
<td>13</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 3.9.10: Conceptual #1.3 distracter analysis

<table>
<thead>
<tr>
<th>Conceptual Item #1.3</th>
<th>Key</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>High performers (top 33%)</td>
<td>34</td>
<td>7</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Low performers (bottom 33%)</td>
<td>7</td>
<td>16</td>
<td>21</td>
<td>5</td>
</tr>
</tbody>
</table>
3.9.5  Treatment #2 Performance Test Items

The sample for treatment session #2 is \( N=143 \) subjects.

3.9.5.1  Item Difficulty Index for Treatment #2 Performance Test

The mean \( p \) value (difficulty index) across the treatment #2 performance items is 0.74: the ideal value for 4-choice selected-response tests (Reynolds et al., 2006, p. 144). Individual \( p \) values range from 0.45 to 0.95 (Table 3.9.11). Ten of the 12 items fall within the range of 0.66 to 0.86.

Table 3.9.11: Treatment #2 difficulty

<table>
<thead>
<tr>
<th>Treatment #2</th>
<th>Difficulty (( p ))</th>
<th>Treatment #2</th>
<th>Difficulty (( p ))</th>
<th>Treatment #2</th>
<th>Difficulty (( p ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual #2.1</td>
<td>0.6783</td>
<td>Conceptual #2.1</td>
<td>0.6923</td>
<td>Factual #2 Set</td>
<td>0.8135</td>
</tr>
<tr>
<td>Factual #2.2</td>
<td>0.9510</td>
<td>Conceptual #2.2</td>
<td>0.6643</td>
<td>Conceptual #2 Set</td>
<td>0.6562</td>
</tr>
<tr>
<td>Factual #2.3</td>
<td>0.8392</td>
<td>Conceptual #2.3</td>
<td>0.4545</td>
<td>Full #2 Set</td>
<td>0.7354</td>
</tr>
<tr>
<td>Factual #2.4</td>
<td>0.8531</td>
<td>Conceptual #2.4</td>
<td>0.7203</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factual #2.5</td>
<td>0.8601</td>
<td>Conceptual #2.5</td>
<td>0.7133</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factual #2.6</td>
<td>0.7063</td>
<td>Conceptual #2.6</td>
<td>0.6923</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The factual item set has a mean \( p \) value of 0.81, and includes the item with the 0.95 difficulty index. The conceptual item set is more difficult, with a mean \( p \) value of 0.66. Only one item has a difficulty index below the set’s mean, at 0.45. The items with difficulty indices outside of the range of 0.66 to 0.86 stand out and will be addressed here.

Factual item #2.2 was answered correctly by 95% of subjects:

- How are lakes and ponds different?
  - Lakes are larger than ponds. [0.95] *
  - Ponds are larger than lakes. [0.00]
- Lakes provide homes for insects; ponds do not. [0.01]
- Ponds provide homes for insects; lakes do not. [0.03]

Out of 143 subjects, only 7 failed to locate the key. No subjects selected the first distracter. An item with a $p$ value of 0.95 cannot accomplish much. Upon close inspection, it is not surprising that this item proved to be problematically easy. All form versions explicitly state that the “main difference” between lakes and ponds is size, with lakes being larger (McDougal Littell, 2005a, p. C18). It is also quite clear in all versions that insects inhabit both lakes and ponds. This item thus requires only the most basic location of explicitly provided information, which was not the goal for any items, even those at the lower “factual” level.

Conceptual item #2.3 was answered correctly by only 45% of subjects:

- What about a lake is the SAME in the summer and winter seasons?
  - it is not experiencing turnover [0.45] *
  - its water temperature [0.14]
  - location of the fish [0.10]
  - its nutrients are being stirred up [0.31]

This item requires the subject to compare information from at least two locations on the page for each option assessment. In all versions information about lakes in summer is available in diagrammatic form. In the prose primary and prose subsumed versions, summer information is redundant in prose format. In the fully integrated version, only some of the summer information is redundant in text chunks, which are ancillary to the major diagram (see Figure 8.3.3 in the appendix). The fully integrated version is the only one that provides winter information in diagrammatic format. The other versions rely fully on prose.

All distracters in this item are strong. The key and third distracter both address aspects that the summer and winter seasons share. The third distracter—the most powerful of the three—just happens to be wrong for both seasons. That this is the strongest distracter
might suggest some confusion as to what the stem of the item requires, despite the emphasis ("SAME"). After all, it is especially clear in the summer diagram that the nutrients have collected at the bottom of the lake. Text makes this explicit: “nutrients settle to the bottom of the lake” (McDougal Littell, 2005a, p. C19).

The first and second distracter both address aspects that differ between the summer and winter seasons. The first distracter is arguably obvious. It would be assumed that subjects would expect temperature to vary between seasons. Perhaps the relative strength of this distracter is further evidence that the item’s stem was difficult to understand.

### 3.9.5.2 Item Discrimination Index for Treatment #2 Performance Test

Session #2 factual test items range in item discrimination index ($D$) from 0.10 to 0.69 (Table 3.9.12).

<table>
<thead>
<tr>
<th>Treatment #2</th>
<th>Mean ($p$)</th>
<th>High ($p_{h}$)</th>
<th>Low ($p_{l}$)</th>
<th>Discr. ($D$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual #2.1</td>
<td>0.6783</td>
<td>1.0000</td>
<td>0.3125</td>
<td>0.6875</td>
</tr>
<tr>
<td>Factual #2.2</td>
<td>0.9510</td>
<td>0.9792</td>
<td>0.8750</td>
<td>0.1042</td>
</tr>
<tr>
<td>Factual #2.3</td>
<td>0.8392</td>
<td>0.9792</td>
<td>0.5833</td>
<td>0.3958</td>
</tr>
<tr>
<td>Factual #2.4</td>
<td>0.8531</td>
<td>1.0000</td>
<td>0.6667</td>
<td>0.3333</td>
</tr>
<tr>
<td>Factual #2.5</td>
<td>0.8601</td>
<td>1.0000</td>
<td>0.6458</td>
<td>0.3542</td>
</tr>
<tr>
<td>Factual #2.6</td>
<td>0.7063</td>
<td>0.9375</td>
<td>0.4167</td>
<td>0.5208</td>
</tr>
<tr>
<td>Conceptual #2.1</td>
<td>0.6923</td>
<td>1.0000</td>
<td>0.4167</td>
<td>0.5833</td>
</tr>
<tr>
<td>Conceptual #2.2</td>
<td>0.6643</td>
<td>1.0000</td>
<td>0.3125</td>
<td>0.6875</td>
</tr>
<tr>
<td>Conceptual #2.3</td>
<td>0.4545</td>
<td>0.8125</td>
<td>0.1875</td>
<td>0.6250</td>
</tr>
<tr>
<td>Conceptual #2.4</td>
<td>0.7203</td>
<td>1.0000</td>
<td>0.4583</td>
<td>0.5417</td>
</tr>
<tr>
<td>Conceptual #2.5</td>
<td>0.7133</td>
<td>0.9583</td>
<td>0.4167</td>
<td>0.5417</td>
</tr>
<tr>
<td>Conceptual #2.6</td>
<td>0.6923</td>
<td>0.9583</td>
<td>0.2708</td>
<td>0.6875</td>
</tr>
</tbody>
</table>
Two of the six items have excellent discrimination power, while a further 3 have good discrimination power. Factual item #2.2 has poor discrimination power \((D=0.10)\). According to the standards of this study, the “poor” rating immediately disqualifies this item.

Session #2 conceptual test items range in item discrimination index from 0.54 to 0.69, meaning they all have excellent discrimination power.

Aside from factual item #2.2, which was previously discussed because of its high difficulty index (the item was especially easy) and failure to discriminate, no session #2 items require close inspection here. Conceptual item #2.3 discriminates extremely well despite its difficulty \((p=0.45)\).

### 3.9.5.3 Discrimination Using Item-Total Correlation Coefficients for Treatment #2 Performance Test

After removing factual item #2.2, all items receive an “excellent” discrimination rating using item-total correlation (Table 3.9.13).

<table>
<thead>
<tr>
<th>Treatment #2</th>
<th>(R)</th>
<th>Treatment #2</th>
<th>(R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual #2.1</td>
<td>0.5977</td>
<td>Conceptual #2.1</td>
<td>0.5573</td>
</tr>
<tr>
<td>Factual #2.2</td>
<td>0.2798</td>
<td>Conceptual #2.2</td>
<td>0.6324</td>
</tr>
<tr>
<td>Factual #2.3</td>
<td>0.5182</td>
<td>Conceptual #2.3</td>
<td>0.4965</td>
</tr>
<tr>
<td>Factual #2.4</td>
<td>0.5404</td>
<td>Conceptual #2.4</td>
<td>0.5374</td>
</tr>
<tr>
<td>Factual #2.5</td>
<td>0.5381</td>
<td>Conceptual #2.5</td>
<td>0.5942</td>
</tr>
<tr>
<td>Factual #2.6</td>
<td>0.5500</td>
<td>Conceptual #2.6</td>
<td>0.6124</td>
</tr>
</tbody>
</table>

Table 3.9.13: Treatment #2 item-total correlation
Distracter Analysis for Treatment #2 Performance Test

Conceptual item #2.3 was the sole item in session #2 with an item difficulty index below 0.50. Distracter analysis supports its favorable discrimination rating (Table 3.9.14). The key was selected at a higher rate by high performers, while all distracters were selected at a higher rate by low performers. The third distracter was selected by 7 of the high performers, while the remaining distracters were selected once apiece. Conceptual item #2.3 fares well in distracter analysis.

Table 3.9.14: Conceptual #2.3 distracter analysis

<table>
<thead>
<tr>
<th>Conceptual Item #2.3</th>
<th>Key</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>High performers (top 33%)</td>
<td>39</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Low performers (bottom 33%)</td>
<td>9</td>
<td>12</td>
<td>5</td>
<td>22</td>
</tr>
</tbody>
</table>

No remaining session #2 items require distracter analysis. Factual item #2.2 was removed during discrimination analysis, resulting in a total treatment item count of 11.

3.9.6 Treatment #3 Performance Test Items

The sample for treatment session #3 is N=148 subjects.

3.9.6.1 Item Difficulty Index for Treatment #3 Performance Test

The mean $p$ value (difficulty index) across the treatment #3 performance items is 0.67. Individual $p$ values range from 0.35 to 0.85 (Table 3.9.15). Only one item has an item difficulty index lower than 0.50.
The factual item set has a mean $p$ value of 0.76, with items that range from 0.72 to 0.85. The conceptual item set has a mean $p$ value of 0.59. Aside from the low individual $p$ value of 0.35, conceptual items range from 0.50 to 0.84.

<table>
<thead>
<tr>
<th>Treatment #3</th>
<th>Difficulty ($p$)</th>
<th>Treatment #3</th>
<th>Difficulty ($p$)</th>
<th>Treatment #3</th>
<th>Difficulty ($p$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual #3.1</td>
<td>0.8041</td>
<td>Conceptual #3.1</td>
<td>0.5203</td>
<td>Factual #3 Set</td>
<td>0.7601</td>
</tr>
<tr>
<td>Factual #3.2</td>
<td>0.7568</td>
<td>Conceptual #3.2</td>
<td>0.5405</td>
<td>Conceptual #3 Set</td>
<td>0.5878</td>
</tr>
<tr>
<td>Factual #3.3</td>
<td>0.6757</td>
<td>Conceptual #3.3</td>
<td>0.8378</td>
<td>Full #3 Set</td>
<td>0.6740</td>
</tr>
<tr>
<td>Factual #3.4</td>
<td>0.8514</td>
<td>Conceptual #3.4</td>
<td>0.5000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factual #3.5</td>
<td>0.7162</td>
<td>Conceptual #3.5</td>
<td>0.3514</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factual #3.6</td>
<td>0.7568</td>
<td>Conceptual #3.6</td>
<td>0.7770</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conceptual item #3.5 was answered correctly by only 35% of subjects:

- Why might a local government choose to approve deep coal mining but not surface coal mining?
  - because the community especially loves its natural landscape [0.35] *
  - because the community is especially worried about worker health [0.09]
  - so that there is no pollution in the creation of electricity [0.30]
  - because only surface mining has unwanted costs associated with it [0.26]

This item requires speculative reasoning. The subject must recognize critical differences between the two types of coal mining and imagine how others might reason through those differences. The distinction between the best answer and the others is a subtle one. The options tend to be fairly complex. It is no surprise that this item proves to be among the more difficult of the treatment tests.
The second distracter was selected almost as frequently as the key. The third distracter is not far behind. This item is suspect and requires special attention in discrimination and distracter analysis. It is a possible candidate for exclusion.

**3.9.6.2 Item Discrimination Index for Treatment #3 Performance Test**

Session #3 factual test items range in item discrimination index ($D$) from 0.27 to 0.59 (Table 3.9.16). Three of the six items have excellent discrimination power. Two items have good discrimination power. Factual item #3.4 has only fair discrimination power ($D=0.27$).

Session #3 conceptual test items range in item discrimination index from 0.24 to 0.73. All save conceptual item #3.3 have excellent discrimination power. Conceptual item #3.3 has only fair discrimination power ($D=0.24$). Conceptual item #3.5, flagged before as an especially difficult item, has excellent discrimination power at 0.65.

**Table 3.9.16: Treatment #3 discrimination**

<table>
<thead>
<tr>
<th>Treatment #3</th>
<th>Mean ($\mu$)</th>
<th>High ($\mu_t$)</th>
<th>Low ($\mu_l$)</th>
<th>Discr. ($D$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual #3.1</td>
<td>0.8041</td>
<td>1.0000</td>
<td>0.5510</td>
<td>0.4490</td>
</tr>
<tr>
<td>Factual #3.2</td>
<td>0.7568</td>
<td>0.8980</td>
<td>0.5102</td>
<td>0.3878</td>
</tr>
<tr>
<td>Factual #3.3</td>
<td>0.6757</td>
<td>0.9388</td>
<td>0.3469</td>
<td>0.5918</td>
</tr>
<tr>
<td>Factual #3.4</td>
<td>0.8514</td>
<td>0.9796</td>
<td>0.7143</td>
<td>0.2653</td>
</tr>
<tr>
<td>Factual #3.5</td>
<td>0.7162</td>
<td>0.8980</td>
<td>0.4694</td>
<td>0.4286</td>
</tr>
<tr>
<td>Factual #3.6</td>
<td>0.7568</td>
<td>0.9592</td>
<td>0.5714</td>
<td>0.3878</td>
</tr>
<tr>
<td>Conceptual #3.1</td>
<td>0.5203</td>
<td>0.9184</td>
<td>0.2245</td>
<td>0.6939</td>
</tr>
<tr>
<td>Conceptual #3.2</td>
<td>0.5405</td>
<td>0.8367</td>
<td>0.2449</td>
<td>0.5918</td>
</tr>
<tr>
<td>Conceptual #3.3</td>
<td>0.8378</td>
<td>0.9388</td>
<td>0.6939</td>
<td>0.2449</td>
</tr>
<tr>
<td>Conceptual #3.4</td>
<td>0.5000</td>
<td>0.8980</td>
<td>0.1633</td>
<td>0.7347</td>
</tr>
<tr>
<td>Conceptual #3.5</td>
<td>0.3514</td>
<td>0.7959</td>
<td>0.1429</td>
<td>0.6531</td>
</tr>
<tr>
<td>Conceptual #3.6</td>
<td>0.7770</td>
<td>0.9796</td>
<td>0.5714</td>
<td>0.4082</td>
</tr>
</tbody>
</table>
Factual item #3.4 was answered correctly by fully 85% of subjects, which in part explains its fair discrimination power:

- What is the main use of coal?
  - as an easy-burning source of energy [0.85] *
  - as a source of acid [0.02]
  - to cool the machinery in a power plant [0.07]
  - to produce pressure underground [0.05]

The first distracter requires the subject to differentiate between purpose and byproduct, as burning coal does produce acid. The second distracter seems to be the most obvious of the three, but it proved the opposite. The third distracter creates a basic causality problem: coal is produced by pressure underground; it doesn’t cause pressure. All distracters were selected, just by very few in each case.

Conceptual item #3.3 was answered correctly by fully 84% of subjects, which limits its discrimination index (as with the previously discussed item):

- What is NOT run by electricity?
  - fossil fuel power plants [0.84] *
  - an air conditioner [0.06]
  - giant factories [0.07]
  - a television [0.03]

This is a process-of-elimination item, in terms of structure, though a confident test-taker can select the key with immediacy. Ultimately, the key is rooted in a fundamental understanding of the use of fossil fuels: they can produce electricity cheaply. Thus fossil fuel power plants must produce electricity, not be run with it.

The first and third distracters are seemingly both obvious, though the first proved as strong as the more compelling distracter that remains. The second distracter relates to a
sentence that appears in each form version: “Electricity runs nearly everything in modern life, from giant factories to the smallest light in your home” (McDougal Littell, 2005a, p. A87).

Ultimately, conceptual item #3.3 does not fare well under basic scrutiny. It does not appear to test conceptual knowledge. Two distracters are far too obvious, and the third distracter relies on a statement found in a single location. The answer is a fundamental matter. It might be argued that as such it is actually a promising item—it addresses a core concept of the handout—but it may be too simple to be of use. Factual item #3.4 has virtually the same item difficulty index and item discrimination index, but it is apparent that its solution requires some degree of reasoning. Conceptual item #3.3 seems to require far less, despite its status as a conceptual-level item.

3.9.6.3 Discrimination Using Item-Total Correlation Coefficients for Treatment #3 Performance Test

Ten of the twelve items from session #3 have excellent discrimination power according to their item-total correlation coefficients (Table 3.9.17).

<table>
<thead>
<tr>
<th>Treatment #3</th>
<th>R</th>
<th>Treatment #3</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual #3.1</td>
<td>0.4933</td>
<td>Conceptual #3.1</td>
<td>0.5875</td>
</tr>
<tr>
<td>Factual #3.2</td>
<td>0.4540</td>
<td>Conceptual #3.2</td>
<td>0.4991</td>
</tr>
<tr>
<td>Factual #3.3</td>
<td>0.5438</td>
<td>Conceptual #3.3</td>
<td>0.2961</td>
</tr>
<tr>
<td>Factual #3.4</td>
<td>0.3856</td>
<td>Conceptual #3.4</td>
<td>0.6246</td>
</tr>
<tr>
<td>Factual #3.5</td>
<td>0.4121</td>
<td>Conceptual #3.5</td>
<td>0.5995</td>
</tr>
<tr>
<td>Factual #3.6</td>
<td>0.4118</td>
<td>Conceptual #3.6</td>
<td>0.5100</td>
</tr>
</tbody>
</table>
Factual item #3.4, which received a fair rating for its item discrimination index, is in the high “good” range here, allaying concerns. Conceptual item #3.3, the other item with an earlier fair rating, does not do so well here. Its item-total correlation coefficient is 0.30, at the very bottom of the “fair” range. Its tenuous status here appears near enough to a double fair rating through the dual discrimination analyses to justify exclusion from the study, especially considering its low face validity—the item appears to be bad upon inspection.

3.9.6.4 Distracter Analysis for Treatment #3 Performance Test

Three session #3 items were flagged during difficulty and discrimination analyses, none of which has an eccentric distracter profile. Conceptual item #3.5 was notable for its item difficulty index. It discriminated well, however, and its distracter profile (Table 3.9.18) reiterates this. The distracters are especially strong for low performers, and relatively weak for high performers. The third distracter did prove compelling for some high performers.

Table 3.9.18: Conceptual #3.5 distracter analysis

<table>
<thead>
<tr>
<th>Conceptual Item #3.5</th>
<th>Key</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>High performers (top 33%)</td>
<td>39</td>
<td>1</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Low performers (bottom 33%)</td>
<td>7</td>
<td>8</td>
<td>19</td>
<td>15</td>
</tr>
</tbody>
</table>

Factual item #3.4 and conceptual item #3.3 were flagged during discrimination analysis. These items have nearly identical distracter profiles, though the latter has been removed from the study and the former retained.
3.9.7 Item Analysis in Review

3.9.7.1 Optimizing Difficulty

While it was possible to establish strict guidelines for the discrimination analyses, item difficulty was the result of an item development process that could not include pilot testing with true surrogate subjects. Though an ideal item difficulty index across a single battery is 0.74, there was no means by which to aim for this target with any accuracy. As such, item difficulty could only serve as a starting point in item analysis.

Item development was conducted before and amidst the four sessions. As treatment session #1 difficulty indices became available immediately following administration, item development for session #2 was in its mature stages while session #3 forms were being finalized. It became apparent immediately following treatment #1 that its items were too difficult. This represented a threat to the entire study. It was determined that difficulty should be adjusted for the second and third treatment tests, despite the resultant threat to validity and potential imbalance across treatment sessions. This manipulation makes it necessary to consider the second and third treatments both together and apart from treatment #1 in subsequent analyses.

Table 3.9.19 presents combined item difficulty indices ($p$) for each set of items. Each mean value is the sum of item difficulty indices for the set. Since each set contains 6 items (this data includes all items), the possible mean range is 0.00 to 6.00. Each set’s value is compared to the overall mean of all sets combined. The resulting values are a testament to the varying difficulty level of the sets. Each conceptual set is more difficult than its corresponding factual set, which is desirable.
Table 3.9.19: Difficulty indices compared

<table>
<thead>
<tr>
<th>Full Item Sets</th>
<th>Total Mean</th>
<th>Mean Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual #1 Set</td>
<td>3.8493</td>
<td>-0.0586</td>
</tr>
<tr>
<td>Factual #2 Set</td>
<td>4.8881</td>
<td>+0.9802</td>
</tr>
<tr>
<td>Factual #3 Set</td>
<td>4.5608</td>
<td>+0.6529</td>
</tr>
<tr>
<td>Conceptual #1 Set</td>
<td>2.6849</td>
<td>-1.2229</td>
</tr>
<tr>
<td>Conceptual #2 Set</td>
<td>3.9371</td>
<td>+0.0292</td>
</tr>
<tr>
<td>Conceptual #3 Set</td>
<td>3.5270</td>
<td>-0.3808</td>
</tr>
</tbody>
</table>

3.9.7.2 Excluded Items

Items were interrogated based on item difficulty index, item discrimination index, and item-total correlations. Distracter analysis was conducted with flagged items only. Table 3.9.20 lists all items discussed in detail, along with the initial condition that precipitated review. (An item flagged for both item difficulty index and item-total correlation would count item difficulty index as the initial condition, as it occurs earlier in the analysis sequence.)

Three items were excluded from the study, one from each treatment session: factual items #1.5 and #2.2, and conceptual item #3.3. These exclusions are reviewed in Table 8.5.2 in the appendix.

Removal of the three excluded items results in reduced item sets (see Table 8.5.3 in the appendix). Half of the sets retain 6 items total, while the other half are reduced to 5 items total. All further calculations for this study will utilize the reduced item sets. Excluded items receive no further consideration.
Table 3.9.20: Item interrogation summary

<table>
<thead>
<tr>
<th>Item Interrogated</th>
<th>Initial Flag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review Item #6</td>
<td>Item difficulty index ($p=0.11$)</td>
</tr>
<tr>
<td>Factual Item #1.5</td>
<td>Item discrimination index ($D=0.22$)</td>
</tr>
<tr>
<td>Conceptual Item #1.1</td>
<td>Item difficulty index ($p=0.37$)</td>
</tr>
<tr>
<td>Conceptual Item #1.2</td>
<td>Item difficulty index ($p=0.40$)</td>
</tr>
<tr>
<td>Conceptual Item #1.3</td>
<td>Item difficulty index ($p=0.36$)</td>
</tr>
<tr>
<td>Conceptual Item #1.5</td>
<td>Item difficulty index ($p=0.48$)</td>
</tr>
<tr>
<td>Factual Item #2.2</td>
<td>Item difficulty index ($p=0.95$)</td>
</tr>
<tr>
<td>Conceptual Item #2.3</td>
<td>Item difficulty index ($p=0.45$)</td>
</tr>
<tr>
<td>Factual Item #3.4</td>
<td>Item discrimination index ($D=0.27$)</td>
</tr>
<tr>
<td>Conceptual Item #3.3</td>
<td>Item discrimination index ($D=0.24$)</td>
</tr>
<tr>
<td>Conceptual Item #3.5</td>
<td>Item difficulty index ($p=0.35$)</td>
</tr>
</tbody>
</table>
SECTION 4 — GROUPS

4.1 Final Participants

4.1.1 Sample Size and Makeup

The 7th grade enrollment at the participating school is nearly 200 students. Informed consent was adequately completed for 167 subjects: 89 female and 78 male. Nine subjects were removed from the study. Each case of exclusion is detailed here:

1) Open-ended difficulty assessment: explicitly admitted to guessing for second and third sessions.
2) Open-ended difficulty assessment: claimed to not have had or used handout. (It is inconceivable that a subject did not receive a handout for one session, but it might have been overlooked or ignored.)
3) Assigned to a repeat subgroup: received FI version twice and no PP version. Subjects copied a code given on the form. Likely form distribution error.
4) Open-ended difficulty assessment: declared poor English language knowledge, sufficient to render text-based testing meaningless. Session #1: “Because I don’t well english”. Session #2: “because”. Session #3: “Because english is difficult.” (There are many students at the school for whom English is a second language. The task difficulty assessment items helped to isolate any whose proficiency was especially low.)

5) Took extended break for eye problem during third session.

6) Open-ended difficulty assessment: does not seem to grasp the language of the forms, or at least the test instrument. Session #1: “Sorry, I do not understand”. Session #2: “sorry, I can’t explain that.” Session #3: “because I am not america people, so I do not understand many word.”

7) Assigned to repeat subgroup: received PS version twice and no FI version. Form distribution error.

8) Open-ended difficulty assessment: self-reported use of web search to answer the questions. This was of course not allowed. Test proctor monitored class but cheating is always possible. Session #3: “I really didn’t think that the handout has as much information as needed! To be truthful I thought google was a source for one of the only question that i needed help to answer, so I decided to use google.com for backup. And then again I thought the article was somewhat easy because, I found most answers in the Article, about half of them for more immidiate help.”

9) Open-ended difficulty assessment: claimed not to use the handout for the first session, and did not “understand” the handout for the remaining sessions. Session #1: “i could not use the hand out i did not know how to use the ahnd outs so i used what i already know”

Excluding these 9 subjects resulted in a reduced sample of 158 subjects: 84 female and 74 male. Many subjects missed one or more of the treatment sessions. Only 118 subjects attended all four sessions.
Subjects were found in eight classes, four apiece for the two participating teachers. Subject numbers per class for full-session participants can be found in Table 4.1.1.

**Table 4.1.1: Subject sex by classroom**

<table>
<thead>
<tr>
<th>Classroom</th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher A, Class 1</td>
<td>11</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>Teacher A, Class 2</td>
<td>12</td>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>Teacher A, Class 3</td>
<td>9</td>
<td>13</td>
<td>22</td>
</tr>
<tr>
<td>Teacher A, Class 4</td>
<td>11</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>Teacher B, Class 1</td>
<td>9</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>Teacher B, Class 2</td>
<td>12</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>Teacher B, Class 3</td>
<td>12</td>
<td>10</td>
<td>22</td>
</tr>
<tr>
<td>Teacher B, Class 4</td>
<td>8</td>
<td>9</td>
<td>17</td>
</tr>
<tr>
<td>Teacher A Total</td>
<td>43</td>
<td>38</td>
<td>81</td>
</tr>
<tr>
<td>Teacher B Total</td>
<td>41</td>
<td>36</td>
<td>77</td>
</tr>
<tr>
<td>Grand Total</td>
<td>84</td>
<td>74</td>
<td>158</td>
</tr>
</tbody>
</table>

**4.1.2 Group Size**

Subjects were assigned to one of the six order-based groups within each class. Due to unstable rolls and attendance, the order-based groups are not as balanced as was hoped (see Table 4.1.2). However, for each given session the number of subjects assigned to the three versions is far more balanced (see Table 4.1.3).
### Table 4.1.2: Order-based groups per classroom

<table>
<thead>
<tr>
<th>Classroom</th>
<th>PP-PS-FI</th>
<th>PP-FI-PS</th>
<th>PS-PP-FI</th>
<th>PS-FI-PP</th>
<th>FI-PP-PS</th>
<th>FI-PS-PP</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher A, Class 1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>Teacher A, Class 2</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>21</td>
</tr>
<tr>
<td>Teacher A, Class 3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>Teacher A, Class 4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>Teacher B, Class 1</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>18</td>
</tr>
<tr>
<td>Teacher B, Class 2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Teacher B, Class 3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>Teacher B, Class 4</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>Teacher A Total</td>
<td>13</td>
<td>12</td>
<td>12</td>
<td>16</td>
<td>16</td>
<td>12</td>
<td>81</td>
</tr>
<tr>
<td>Teacher B Total</td>
<td>13</td>
<td>12</td>
<td>15</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>77</td>
</tr>
<tr>
<td>Grand Total</td>
<td>26</td>
<td>24</td>
<td>27</td>
<td>28</td>
<td>31</td>
<td>22</td>
<td>158</td>
</tr>
</tbody>
</table>

### Table 4.1.3: Group assignment per treatment

<table>
<thead>
<tr>
<th>Full-Session Participants</th>
<th>PP</th>
<th>PS</th>
<th>FI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment Session #1</td>
<td>38</td>
<td>39</td>
<td>41</td>
</tr>
<tr>
<td>Treatment Session #2</td>
<td>45</td>
<td>38</td>
<td>35</td>
</tr>
<tr>
<td>Treatment Session #3</td>
<td>35</td>
<td>41</td>
<td>42</td>
</tr>
</tbody>
</table>
4.2 Equivalence in Categorical Data

4.2.1 Preliminary Scrutiny into Categorical Data

A better understanding of the sample should precede inquiry into the independent variable. Are there differences between males and females? Are scores on the varying tests and inventories influenced by teacher and class assignment? Observation by the researcher and research assistant suggests a real difference between the participating teachers’ classrooms. One teacher’s classes were consistently chaotic and loud while the other teacher’s were ordered and relatively quiet. In one classroom desks were organized into clusters where students faced one another; in the other desks were organized into rows, facing the teacher and allowing the researcher to observe all subjects’ laptop screens simultaneously.

Analysis of variance (ANOVA) was run for the following variables with each of sex, class assignment, and teacher:

- Review test scores
- Combined treatment test scores
- Science self-efficacy scores
- Combined indirect interest level ratings
- Combined difficulty assessment ratings
- Adjusted direct interest level ratings
- Combined session duration measures

In the interest of brevity, only those ANOVA tests that found significant differences at $P \leq 0.05$ will be discussed here. ANOVA tables for this study can be found in the appendix.
4.2.2 The Teacher Variable

The teacher variable shows no significant differences in any of the ANOVA tests.

4.2.3 The Class Variable

Class assignment shows significant differences in only two cases: combined treatment test scores and session duration.

4.2.3.1 Class Assignment Differences in Combined Treatment Tests

Combining treatment test scores produces total scores in the range of 0 to 33. The lowest class mean is over 7 points lower than the highest class mean (Table 4.2.1). Not surprisingly given this difference, an ANOVA test found significant differences in the groups ($P=0.0334$).

<table>
<thead>
<tr>
<th>Combined Treatment Tests</th>
<th>$N$</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher A, Class 1</td>
<td>16</td>
<td>23.19</td>
<td>1.57</td>
<td>20.07</td>
<td>26.30</td>
</tr>
<tr>
<td>Teacher A, Class 2</td>
<td>18</td>
<td>19.78</td>
<td>1.48</td>
<td>16.84</td>
<td>22.71</td>
</tr>
<tr>
<td>Teacher A, Class 3</td>
<td>18</td>
<td>24.89</td>
<td>1.48</td>
<td>21.95</td>
<td>27.82</td>
</tr>
<tr>
<td>Teacher A, Class 4</td>
<td>14</td>
<td>20.71</td>
<td>1.68</td>
<td>17.38</td>
<td>24.04</td>
</tr>
<tr>
<td>Teacher B, Class 1</td>
<td>16</td>
<td>23.69</td>
<td>1.57</td>
<td>20.57</td>
<td>26.80</td>
</tr>
<tr>
<td>Teacher B, Class 2</td>
<td>18</td>
<td>19.94</td>
<td>1.48</td>
<td>17.01</td>
<td>22.85</td>
</tr>
<tr>
<td>Teacher B, Class 3</td>
<td>14</td>
<td>17.78</td>
<td>1.68</td>
<td>14.45</td>
<td>21.11</td>
</tr>
<tr>
<td>Teacher B, Class 4</td>
<td>11</td>
<td>21.82</td>
<td>1.89</td>
<td>18.06</td>
<td>25.57</td>
</tr>
</tbody>
</table>

A follow up Student’s $t$ test isolates 5 pairwise differences (Table 8.8.1 in the appendix). Class assignment at the middle school is non-random, with certain higher-level
classes being offered at particular times. Furthermore, classes are loci of sustained interactions over a long period of time; it should be expected that some classes, as a group, develop and perpetuate habits that can lead to a more or less distracting environment. This was certainly the observation of the study proctors. Of course, subjects were randomly assigned to groups, and attempts were made to balance group assignment between classes, as it was the starting assumption of this study that classes would bear differences.

4.2.3.2 Class Assignment Differences in Session Duration

Analysis of variance found significant differences is session duration based on class assignment ($P=0.0093$). Per-class means for duration, in seconds, are reported in Table 4.2.2.

Table 4.2.2: Session duration means by classroom

<table>
<thead>
<tr>
<th>Session Duration</th>
<th>N</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher A, Class 1</td>
<td>16</td>
<td>2212.19</td>
<td>210.22</td>
<td>1795.9</td>
<td>2628.5</td>
</tr>
<tr>
<td>Teacher A, Class 2</td>
<td>18</td>
<td>2431.00</td>
<td>198.20</td>
<td>2038.5</td>
<td>2823.5</td>
</tr>
<tr>
<td>Teacher A, Class 3</td>
<td>18</td>
<td>2158.94</td>
<td>198.20</td>
<td>1766.4</td>
<td>2551.5</td>
</tr>
<tr>
<td>Teacher A, Class 4</td>
<td>14</td>
<td>3273.86</td>
<td>224.74</td>
<td>2828.8</td>
<td>3718.9</td>
</tr>
<tr>
<td>Teacher B, Class 1</td>
<td>16</td>
<td>2688.12</td>
<td>210.22</td>
<td>2271.8</td>
<td>3104.5</td>
</tr>
<tr>
<td>Teacher B, Class 2</td>
<td>18</td>
<td>2241.78</td>
<td>198.20</td>
<td>1849.3</td>
<td>2634.3</td>
</tr>
<tr>
<td>Teacher B, Class 3</td>
<td>14</td>
<td>2565.93</td>
<td>224.74</td>
<td>2120.9</td>
<td>3011.0</td>
</tr>
<tr>
<td>Teacher B, Class 4</td>
<td>11</td>
<td>2657.18</td>
<td>253.54</td>
<td>2155.1</td>
<td>3159.3</td>
</tr>
</tbody>
</table>

A Student’s $t$ test finds significant differences in 5 pairwise comparisons (Table 8.8.2 in the appendix). Both teachers individually notified the researcher that the final period was more difficult to manage, and these classes were two of the three slowest classes completing the treatment sessions. In the case of teacher A’s final period, the classroom was certainly chaotic, with many disruptions. It is no surprise that session duration proved excessive in that
That same teacher’s previous period took approximately $\frac{2}{3}$ the time to complete the same work.

### 4.2.4 The Sex Variable

Sex shows significant differences for the review test, the treatment tests, the science self-efficacy inventory, and session duration.

#### 4.2.4.1 Sex Differences in the Review Test

Males scored higher than females on the review battery (Table 4.2.3). An ANOVA test found this to be a significant difference ($P=0.0010$).

**Table 4.2.3: Review test means by sex**

<table>
<thead>
<tr>
<th>Review Battery</th>
<th>N</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female Subjects</td>
<td>80</td>
<td>6.2125</td>
<td>0.1953</td>
<td>5.8266</td>
<td>6.5984</td>
</tr>
<tr>
<td>Male Subjects</td>
<td>68</td>
<td>7.1765</td>
<td>0.2118</td>
<td>6.7579</td>
<td>7.5950</td>
</tr>
</tbody>
</table>

#### 4.2.4.2 Sex Differences in Combined Treatment Tests

As might be expected following the findings of the review battery, males outperformed females on the treatment batteries when scored in aggregate ($P=0.0130$), which ranges in score from 0 to 33 (Table 4.2.4). Note that while the review test measures science knowledge (on recently learned material) the treatment batteries measure performance with science information. On average, males answered nearly 1 more performance item correctly per treatment session.
### 4.2.4.3 Sex Differences in the Science Self-Efficacy Inventory

Males in the study not only received higher test scores than females, but they felt more confident in their abilities in the domain of science. Males have higher science self-efficacy than females (Table 4.2.5, $P=0.0077$).

#### Table 4.2.5: Science self-efficacy means by sex

<table>
<thead>
<tr>
<th>Science Self-Efficacy</th>
<th>N</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female Subjects</td>
<td>80</td>
<td>29.5875</td>
<td>0.5829</td>
<td>28.435</td>
<td>30.740</td>
</tr>
<tr>
<td>Male Subjects</td>
<td>68</td>
<td>31.9118</td>
<td>0.6323</td>
<td>30.662</td>
<td>33.161</td>
</tr>
</tbody>
</table>

### 4.2.4.4 Sex Differences in Session Duration

Finally, females spent significantly longer than males completing the treatment sessions (Table 4.2.6, $P=0.0316$).
Table 4.2.6: Session duration means by sex

<table>
<thead>
<tr>
<th>Session Duration</th>
<th>N</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female Subjects</td>
<td>68</td>
<td>2653.85</td>
<td>105.52</td>
<td>2445.0</td>
<td>2862.7</td>
</tr>
<tr>
<td>Male Subjects</td>
<td>57</td>
<td>2314.04</td>
<td>115.26</td>
<td>2085.9</td>
<td>2542.2</td>
</tr>
</tbody>
</table>

4.3  

Equivalence of Groups

4.3.1  

Preliminary Scrutiny into Group Performance

Scores for batteries and inventories can be used to check the equivalence of the treatment sequence groups for equivalency. Ideally, of course, the sequence groups are equivalent. The hypothesis for the following tests is that at least one of the groups is not equivalent to the others. Three item sets will be used to gauge equivalence: the review test, the treatment test batteries combined into a total score, and the science self-efficacy inventory.

4.3.2  

Sequence Group Differences in the Review Test

There is no indication that the sequence groups are dissimilar in terms of science knowledge. Means for sequence group performance on the review test are reported in Table 4.3.1. An ANOVA test finds no significant differences ($P=0.7523$).
4.3.3 Sequence Group Differences in the Treatment Tests

Group performance on individual treatment tests is of course a function of the independent variable. Sequence groups assign each subject to each of the treatment types once. Aggregate treatment test scores should thus not vary between groups if the groups are equivalent. When treatment tests are combined into aggregate scores, which fall in the range of 0 to 33, there is no significant difference in the sequence groups ($P=0.3932$), meaning there is no reason to think that the groups are not equivalent (Table 4.3.2).

<table>
<thead>
<tr>
<th>Treatment Batteries</th>
<th>N</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq. Group PP-PS-FI</td>
<td>24</td>
<td>21.1250</td>
<td>1.3278</td>
<td>18.496</td>
<td>23.754</td>
</tr>
<tr>
<td>Seq. Group PP-FI-PS</td>
<td>17</td>
<td>22.8235</td>
<td>1.5777</td>
<td>19.700</td>
<td>25.948</td>
</tr>
<tr>
<td>Seq. Group PS-PP-FI</td>
<td>19</td>
<td>21.2105</td>
<td>1.4923</td>
<td>18.256</td>
<td>24.166</td>
</tr>
<tr>
<td>Seq. Group PS-FI-PP</td>
<td>22</td>
<td>22.8182</td>
<td>1.3869</td>
<td>20.072</td>
<td>25.564</td>
</tr>
</tbody>
</table>
4.3.4 Sequence Group Differences in Science Self-Efficacy

Finally, the sequence groups exhibit no significant differences ($P=0.2223$) in science self-efficacy ratings (Table 4.3.3).

**Table 4.3.3: Science self-efficacy means by group**

<table>
<thead>
<tr>
<th>Science Self-Efficacy</th>
<th>$N$</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq. Group PP-PS-FI</td>
<td>26</td>
<td>32.6538</td>
<td>1.0369</td>
<td>30.604</td>
<td>34.704</td>
</tr>
<tr>
<td>Seq. Group PP-FI-PS</td>
<td>20</td>
<td>29.4000</td>
<td>1.1823</td>
<td>27.063</td>
<td>31.737</td>
</tr>
<tr>
<td>Seq. Group PS-PP-FI</td>
<td>26</td>
<td>29.5769</td>
<td>1.0369</td>
<td>27.527</td>
<td>31.627</td>
</tr>
<tr>
<td>Seq. Group PS-FI-PP</td>
<td>26</td>
<td>31.1154</td>
<td>1.0369</td>
<td>29.066</td>
<td>33.165</td>
</tr>
<tr>
<td>Seq. Group FI-PP-PS</td>
<td>29</td>
<td>31.0345</td>
<td>0.9818</td>
<td>29.094</td>
<td>32.975</td>
</tr>
<tr>
<td>Seq. Group FI-PS-PP</td>
<td>21</td>
<td>29.6190</td>
<td>1.1538</td>
<td>27.338</td>
<td>31.900</td>
</tr>
</tbody>
</table>
5.1 Reliability

Reliability estimates were calculated using Cronbach’s alpha ($\alpha$). Alpha estimates the degree to which items in a set measure a shared construct. It is a measure of internal reliability (Reynolds et al., 2006, pp. 93–97). Conventional interpretation of alpha is that a value at or above 0.70 is acceptable and needs no further interpretation (Cortina, 1993, p. 101). However, Cortina (1993) has shown that the number of items in a set greatly impacts alpha, where alpha becomes artificially inflated as the number of items increases. Cortina compares two scales with alpha of 0.80: one with 3 items and another with 10 items. A 3-item set with $\alpha=0.80$ has an average pairwise correlation of 0.57. A 10-item set with the same alpha has an average pairwise correlation of only 0.28. Cortina demonstrates that a 6-item set with an average pairwise correlation of 0.30 has an alpha 16 points lower than an 18-item set with the same average pairwise correlation—0.72 compared to 0.88 (p. 102). This study has test sets of 11, 10, 6 and 5 items total, and Likert inventories of 8 and 4 items total.
Particularly difficult to gauge here is the difference between test items and Likert scales. The performance test items for this study are organized into sets that are indeed intended to measure a single construct. For instance, treatment session #2 factual items are meant to measure subject performance with factual information for the “lakes and lake turnover” handout. However, each individual item must be answered by attending to information found in particular locations on the page in alternating codes (text, image, and text–image complexes). So though the items are unified in that each requires the subject to manipulate factual-level information about lakes from a single handout, they do so through performance with varied elements within the handout. Contrast this with a 4-item Likert inventory that asks how interested a subject is in the handout’s content four times over with only slightly varied wording. Ultimately, even at best, the test item sets should be expected to yield lower reliability estimates than the more direct unitary Likert inventories. A precise quantification of this expected difference in this particular case is elusive. Reynolds et al. (2006) acknowledge that reliability estimates of 0.60 and above are often deemed acceptable for performance assessments, but express discomfiture with any values below 0.70 (p. 103). Reliability estimates of 0.80 and above are generally expected for personality tests (ibid.).

The previous section on item analysis is the most grounded means by which to interrogate the test items, while this section on item reliability is the sole means by which to interrogate the Likert inventories.

5.2 Inquiry Into the Secondary Variables

5.2.1 Secondary Performance Test Batteries

5.2.1.1 Reliability Estimate for Review Test

Figure 5.2.1 is the histogram for the review test. Seven of ten items correct is the modal value.
The correlation matrix for the review test can be found in the appendix (Table 8.6.2) along with all other correlation matrices. Less than a third of the pairwise correlations (13/45) are significant ($P \leq 0.05$), all of which are positive correlations. Review item #9 does not correlate significantly with any other items.

The mean of all pairwise correlations is 0.11. The review test correlations are thus low overall. The items covered varying topics in the 7th grade curriculum, which helps to explain the individual correlations. Alpha for the set is 0.53 (Table 5.2.1). The review test’s lower alpha is expected because it covers subjects across the 7th grade curriculum.

Though the review test is not a direct performance test, one would expect a strong positive correlation between it and the performance batteries. Due to its low alpha rating, aggregate review test scores were compared with the aggregate treatment test scores (Table 8.6.14). The review test correlates positively ($r$) and significantly ($P$) with all of the treatment performance tests.
Table 5.2.1: Review test reliability estimate

<table>
<thead>
<tr>
<th>Review Test (N=148)</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Review Set</td>
<td>0.5284</td>
</tr>
<tr>
<td>Exclude: Item #1</td>
<td>0.4728</td>
</tr>
<tr>
<td>Exclude: Item #2</td>
<td>0.5288</td>
</tr>
<tr>
<td>Exclude: Item #3</td>
<td>0.5039</td>
</tr>
<tr>
<td>Exclude: Item #4</td>
<td>0.5356</td>
</tr>
<tr>
<td>Exclude: Item #5</td>
<td>0.5153</td>
</tr>
<tr>
<td>Exclude: Item #6</td>
<td>0.5133</td>
</tr>
<tr>
<td>Exclude: Item #7</td>
<td>0.4628</td>
</tr>
<tr>
<td>Exclude: Item #8</td>
<td>0.4754</td>
</tr>
<tr>
<td>Exclude: Item #9</td>
<td>0.5033</td>
</tr>
<tr>
<td>Exclude: Item #10</td>
<td>0.5065</td>
</tr>
</tbody>
</table>

5.2.1.2 Reliability Estimate for Pilot Performance Test

Most of the pilot test’s pairwise correlations ($\rho_{15}$) are significant ($P \leq 0.05$) (Table 8.6.8). All significant correlations are positive. The mean pairwise correlation is 0.20 (Table 5.3.1). The battery has an alpha of 0.60, just at the lower threshold. The pilot performance test serves no further purpose.

5.2.2 Science Self-Efficacy Inventory

5.2.2.1 Response Values for the Science Self-Efficacy Inventory

The science self-efficacy inventory was administered to 148 subjects in the pre-test session preceding the performance pilot test. It is an 8-item Likert inventory with 5 options per item. Items #5 and #6 are written in the negative and reverse coded. A value of 5 on an
item indicates the highest self-efficacy rating, a 1 the lowest. The full inventory is thus in the range of 8–40. A histogram of the full inventory is presented as Figure 5.2.2. There is a slight left skew with a notable outlier at 12, the lowest individual value by a full 5 points. Mean values per item range from 3.18 to 4.55 (see Table 8.6.9 in the appendix).

![Figure 5.2.2: Science self-efficacy scores](image)

5.2.2.2 Reliability Estimate for the Science Self-Efficacy Inventory

All pairwise correlations for the science self-efficacy inventory are significant (Table 8.6.1). The mean pairwise correlation is 0.43. The resulting alpha for the set is 0.85 (Table 5.2.2), successfully meeting the more demanding expectations of $\alpha=0.80$ for personality scales (Reynolds et al., 2006, p. 103). Alpha would not increase with the removal of any item. The science-self efficacy inventory survives as a reliable measure.
Table 5.2.2: Science self-efficacy reliability estimate

<table>
<thead>
<tr>
<th>Science Self-Efficacy (N=148)</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Set</td>
<td>0.8507</td>
</tr>
<tr>
<td>Exclude: Item #1</td>
<td>0.8441</td>
</tr>
<tr>
<td>Exclude: Item #2</td>
<td>0.8338</td>
</tr>
<tr>
<td>Exclude: Item #3</td>
<td>0.8318</td>
</tr>
<tr>
<td>Exclude: Item #4</td>
<td>0.8191</td>
</tr>
<tr>
<td>Exclude: Item #5</td>
<td>0.8320</td>
</tr>
<tr>
<td>Exclude: Item #6</td>
<td>0.8449</td>
</tr>
<tr>
<td>Exclude: Item #7</td>
<td>0.8282</td>
</tr>
<tr>
<td>Exclude: Item #8</td>
<td>0.8282</td>
</tr>
</tbody>
</table>

5.2.3 Session Duration

Because the complete treatment sessions—and not the individual components—were timed, validity of session duration measures is a concern. This is especially true because each session included a constructed response item regarding task difficulty that was not time-restricted. There is but a single duration measure for each treatment session. So in order to estimate reliability duration must be compared across sessions.

5.2.3.1 Recorded Values for Session Duration

The mean session duration falls within the range of 13 to 15 minutes for each treatment (see Table 8.6.13 in the appendix). Some subjects finished remarkably quickly, as is evidenced in the minimum values. The histograms (Figure 5.2.3, Figure 5.2.4, and Figure 5.2.5) reveal a number of outliers.
Figure 5.2.3: Session #1 duration record

Figure 5.2.4: Session #2 duration record

Figure 5.2.5: Session #3 duration record
5.2.3.2 Cronbach’s Alpha for Session Duration

The mean pairwise correlation for the duration of the three sessions is 0.74. The corresponding alpha is 0.89 (Table 5.2.3). Session duration, due to questions of validity, must be considered with some hesitation despite the good reliability estimate. Engagement with the form through the treatment test is the only component desirable to measure, but interest level ratings, the difficulty rating, and most notably the open-ended difficulty explanation, are all included in the duration measures.

<table>
<thead>
<tr>
<th>Session Duration</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration as a Set</td>
<td>0.8908</td>
</tr>
<tr>
<td>Excl.: Session #1 Duration</td>
<td>0.7477</td>
</tr>
<tr>
<td>Excl.: Session #2 Duration</td>
<td>0.7520</td>
</tr>
<tr>
<td>Excl.: Session #3 Duration</td>
<td>0.7221</td>
</tr>
</tbody>
</table>

5.3 Text–Image Integration Strategy’s Impact on Comprehension

5.3.1 Reliability Estimates for Treatment Tests

All alpha values for all performance and review test batteries are reported in Table 5.3.1.
Table 5.3.1: Performance test reliability estimates

<table>
<thead>
<tr>
<th>Item Set</th>
<th>Items</th>
<th>Subjects</th>
<th>Pair. Mean</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review Set</td>
<td>10</td>
<td>148</td>
<td>0.1132</td>
<td>0.5284</td>
</tr>
<tr>
<td>Treatment Pilot Set</td>
<td>6</td>
<td>76</td>
<td>0.2020</td>
<td>0.6036</td>
</tr>
<tr>
<td>Factual #1 Set</td>
<td>5</td>
<td>146</td>
<td>0.1181</td>
<td>0.3989</td>
</tr>
<tr>
<td>Conceptual #1 Set</td>
<td>6</td>
<td>146</td>
<td>0.1588</td>
<td>0.5317</td>
</tr>
<tr>
<td>Full #1 Set</td>
<td>11</td>
<td>146</td>
<td>0.1313</td>
<td>0.6234</td>
</tr>
<tr>
<td>Factual #2 Set</td>
<td>5</td>
<td>143</td>
<td>0.2780</td>
<td>0.6456</td>
</tr>
<tr>
<td>Conceptual #2 Set</td>
<td>6</td>
<td>143</td>
<td>0.2594</td>
<td>0.6757</td>
</tr>
<tr>
<td>Full #2 Set</td>
<td>11</td>
<td>143</td>
<td>0.2503</td>
<td>0.7821</td>
</tr>
<tr>
<td>Factual #3 Set</td>
<td>6</td>
<td>148</td>
<td>0.1361</td>
<td>0.4837</td>
</tr>
<tr>
<td>Conceptual #3 Set</td>
<td>5</td>
<td>148</td>
<td>0.2587</td>
<td>0.6386</td>
</tr>
<tr>
<td>Full #3 Set</td>
<td>11</td>
<td>148</td>
<td>0.1838</td>
<td>0.7156</td>
</tr>
</tbody>
</table>

5.3.1.1 Correlation Coefficients and Cronbach’s Alpha for Treatment #1 Performance Test

Histograms for the factual and conceptual item sets from treatment #1 are presented in Figure 5.3.1. Modal values are at 3 and 2 correct items, respectively. Skew is slight for both sets.

Few of the treatment #1 pairwise correlations (17/55) are significant ($P \leq 0.05$, Table 8.6.3). All significant correlations are positive. Factual item #1 does not correlate significantly with any other item. The mean pairwise correlation is 0.12 for the factual set and 0.16 for the conceptual set (Table 5.3.1). The mean pairwise correlation for the combined treatment #1 sets is 0.13.

Alpha for the combined sets is 0.62. This qualifies the full #1 set for the lower acceptable threshold of 0.60. The factual set is $\alpha = 0.40$ and the conceptual set is $\alpha = 0.53$. 
5.3.1.2 Correlation Coefficients and Cronbach’s Alpha for Treatment #2 Performance Test

Histograms for the factual and conceptual item sets from treatment #2 are presented as Figure 5.3.2. Modal values are at 5 and 6 correct items, respectively—both maximum values. Both sets exhibit a strong left skew.

Almost all of the pairwise correlations (48/55) in treatment #2 achieve significance (Table 8.6.4). All significant correlations are positive.
The mean pairwise correlation is 0.28 for the factual set, 0.26 for the conceptual set, and 0.25 for the sets in combination (Table 5.3.1). Alpha for the full set is 0.78, well above the more demanding threshold of 0.70. The individual sets, being lower in count at 5 and 6 items, have correspondingly lower alphas of 0.65 (factual) and 0.68 (conceptual). Despite the low item count, these numbers are in the middle to upper acceptable range.

5.3.1.3 Correlation Coefficients and Cronbach’s Alpha for Treatment #3 Performance Test

Histograms for the factual and conceptual item sets from treatment #3 are presented as Figure 5.3.3. Modal values are at 6 and 2 correct items, respectively. The factual set is strongly skewed left.

![Histograms](image)

Figure 5.3.3: Treatment #3 scores

Just over half of the treatment #3 performance test pairwise correlations \((30/55)\) are significant (Table 8.6.5). All significant correlations are positive. The mean pairwise correlation is 0.17 for the factual set, 0.26 for the conceptual set, and 0.18 for the full set (Table 5.3.1).

The reliability estimate for the full set reaches the higher threshold at \(\alpha=0.72\). The conceptual set reaches the acceptable threshold at \(\alpha=0.64\). Alpha for the factual set is 0.48.
This places the third treatment somewhere between the first and second in terms of estimated reliability.

5.3.1.4 Reliability Summary for the Treatment Performance Tests

The reliability estimate for the entire treatment #2 set is very high. The individual factual and conceptual sets are lower, but acceptable, falling below $\alpha=0.70$ but well above the lower limit of $\alpha=0.60$. Of the other factual and conceptual sets, only conceptual #3 set reaches the lower acceptable threshold.

The full set for treatment #3 reaches the higher threshold, while the full set for treatment #1 is acceptable.

In review, the following combined sets have very favorable reliability estimates:

- Full #2 set
- Full #3 set

The following individual and combined sets have acceptable reliability estimates, meeting the lower threshold requirement:

- Full #1 set
- Factual #2 set
- Conceptual #2 set
- Conceptual #3 set

The following individual sets have poor reliability estimates:

- Factual #1 set
- Conceptual #1 set
- Factual #3 set
The conceptual item sets were consistently more difficult than the corresponding factual item sets (see Table 8.5.1 in the appendix). But the factual–conceptual distinction is meant to be more than a matter of difficulty. The meaningful distinction should be a subtle one in practice. Subjects who are good at manipulating conceptual information are probably good at manipulating factual information as well. Likewise, subjects who are poor with one are probably poor with another. The expected difference would come with subjects who can only work successfully at the lower level of factual information but tend to go astray when asked to think at a deeper level. There may also be subjects who are good at reasoning but less inclined to seek and find more straightforward information.

In any case, the hypothesis is that factual and conceptual items function differently. The null hypothesis is that they only differ in terms of difficulty level. To make a more specific hypothesis, it would be expected that factual items should correlate with other factual items more so than with conceptual items. Conceptual items should correlate with conceptual items more so than with factual items. We can define pairwise correlations within one treatment as either set-internal or set-external. Both factual–factual and conceptual–conceptual pairings are set-internal. Any mixed factual–conceptual pairing is set-external. There are more set-external than set-internal pairs for each treatment. Are correlations in fact higher for set-internal pairs?

Table 5.3.2 shows that in all cases, including all treatments in combination, set-internal pairwise correlations are stronger than set-external pairwise correlations.
5.3.2 Review of Comprehension

Recall the first research question:

- How does the strategy for text–image integration in instructional print media for middle school students impact their comprehension of abstract relationships in science material?

Comprehension is measured through the three treatment performance tests. Treatment session #1 proved less reliable than the others, but still at an acceptable level.

Abstract relationships in science material were targeted at two levels: factual relationships and conceptual relationships. However, only treatment session #2 offers the possibility that a difference between these levels of abstraction might be established, due to reliability estimates.
5.3.3  

*Mean Differences in Comprehension*

Mean group differences for the treatment performance tests are provided in Table 5.3.3 (and visualized in Figure 5.3.4). In all cases, the FI group produced the highest scores. In the first two treatments the PP group produced the lowest scores, while in treatment #3 it was the PS group that scored lowest.

Each treatment test contains 11 items, scored at 1 point apiece. The mean differences for treatment #1 are all less than 0.2 points. The means for treatment #3 vary nearly three times as much, with values within 0.6 of each other. Treatment #2 has by far the most varied group means, with a difference of 1.4 points between the PP and FI groups. This translates to nearly $1\frac{1}{2}$ more correct answers out of only 11 items for those using a fully integrated form over those with a prose primary form.

<table>
<thead>
<tr>
<th>Combined Treatment Tests</th>
<th>N</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form #1: PP Version</td>
<td>47</td>
<td>5.8085</td>
<td>0.3586</td>
<td>5.0997</td>
<td>6.5173</td>
</tr>
<tr>
<td>Form #1: PS Version</td>
<td>51</td>
<td>5.9608</td>
<td>0.3442</td>
<td>5.2803</td>
<td>6.6412</td>
</tr>
<tr>
<td>Form #1: FI Version</td>
<td>48</td>
<td>5.9792</td>
<td>0.3548</td>
<td>5.2778</td>
<td>6.6805</td>
</tr>
<tr>
<td>All #1 Forms</td>
<td>146</td>
<td>5.9178</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form #2: PP Version</td>
<td>54</td>
<td>7.0926</td>
<td>0.3614</td>
<td>6.3782</td>
<td>7.8070</td>
</tr>
<tr>
<td>Form #2: PS Version</td>
<td>43</td>
<td>8.1395</td>
<td>0.4049</td>
<td>7.3390</td>
<td>8.9401</td>
</tr>
<tr>
<td>Form #2: FI Version</td>
<td>46</td>
<td>8.5435</td>
<td>0.3915</td>
<td>7.7694</td>
<td>9.3175</td>
</tr>
<tr>
<td>All #2 Forms</td>
<td>143</td>
<td>7.8741</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form #3: PP Version</td>
<td>44</td>
<td>7.3636</td>
<td>0.3823</td>
<td>6.6080</td>
<td>8.1193</td>
</tr>
<tr>
<td>Form #3: PS Version</td>
<td>53</td>
<td>6.9245</td>
<td>0.3484</td>
<td>6.2360</td>
<td>7.6131</td>
</tr>
<tr>
<td>Form #3: FI Version</td>
<td>51</td>
<td>7.4902</td>
<td>0.3551</td>
<td>6.7883</td>
<td>8.1921</td>
</tr>
<tr>
<td>All #3 Forms</td>
<td>148</td>
<td>7.2500</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.3.4 Hypothesis Tests for Comprehension

Analysis of variance was run for each treatment battery to follow up on the mean differences per assigned form. Treatments #1 ($P=0.9332$) and #3 ($P=0.4937$) show no significant differences (significance standards for this study are at $P\leq0.05$).

However, treatment #2 does show significant differences in performance based on form assigned, at $P=0.0202$. Pairwise comparisons using Student’s $t$ reveal that the significant difference exists between the PP and FI versions at $P=0.0073$ (Table 8.8.3 in the appendix). The difference in the PP and PS versions does not quite achieve significance ($P=0.0557$).

As treatment #2 is the only treatment cleared for separate analysis of factual and comprehension item sets, there is hope that the differences observed might be further illuminated. ANOVA tests reveal significant differences for the 6-item factual set ($P=0.0289$), but not quite the 5-item conceptual set ($P=0.0586$).

For the factual set, Student’s $t$ reveals that both pairs of PP comparisons are significant: both FI and PS groups outperformed the PP group significantly (Table 8.8.4). There is no significant difference between PS and FI.
A student’s $t$ test for the treatment #2 conceptual set, which nearly achieved significance according to ANOVA, finds that FI outperformed PP significantly (Table 8.8.5).

5.3.5  
*Duration as a Factor in Comprehension*

It could be that performance differences are attributable to time spent with the form, which is represented by the duration variable. If subjects spent longer with FI forms, for example, then this might explain higher scores.

The duration means for the treatment #2 forms vary as follows: subjects spent the most time with the PP form and the least time with the FI form (Table 5.3.4). An ANOVA test failed to find any significant differences ($P=0.0674$).

<table>
<thead>
<tr>
<th>Treatment Session #2</th>
<th>N</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form #2: PP Version</td>
<td>54</td>
<td>853.67</td>
<td>39.32</td>
<td>775.92</td>
<td>931.41</td>
</tr>
<tr>
<td>Form #2: PS Version</td>
<td>43</td>
<td>722.26</td>
<td>44.07</td>
<td>635.13</td>
<td>809.38</td>
</tr>
<tr>
<td>Form #2: FI Version</td>
<td>46</td>
<td>833.78</td>
<td>42.61</td>
<td>749.55</td>
<td>918.02</td>
</tr>
</tbody>
</table>

5.3.6  
*Comprehension: Conclusion*

In each treatment session the mean score for FI exceeded the mean score for PP. PS was the lowest scoring strategy in treatment #3, and had the middle scores in the first two treatments. The treatment #1 test is not considered reliable.

An ANOVA test suggested significant differences in the treatment #2 test. FI mean scores were significantly higher than PP scores. ANOVA tests found significant differences in only the individual factual #2 set, though the conceptual #2 set was very near the
significance level. For the factual set, PP scores were significantly lower than both PS and FI scores. For the conceptual set, PP scores were significantly lower than FI scores only.

5.4 **Text–Image Integration Strategy’s Impact on Task Difficulty**

5.4.1 **Reliability Estimates for Task Difficulty Scales**

A task difficulty scale immediately followed each treatment test. Subjects assessed the difficulty level of the task. Each subject’s answer was “piped” into a subsequent constructed response item that inquired into his or her selection.

5.4.1.1 **Response Values for the Task Difficulty Scales**

Each task difficulty assessment consisted of a single task difficulty scale in forced-response Likert format with 6 options. Items are coded such that a 6 indicates the highest difficulty. Histograms for the scales are presented as Figure 5.4.1.

![Difficulty Scale #1](image1)
![Difficulty Scale #2](image2)
![Difficulty Scale #3](image3)

**Figure 5.4.1: Difficulty scores**

In each case multiple subjects selected the highest and lowest levels of difficulty (see Table 8.6.12 in the appendix). The first treatment was rated as the most difficult of the three,
at 3.55 (where a mean rating of 3.50 falls exactly in the middle of the scale). The second treatment was rated as the easiest of the three, at 2.78. The third treatment is in between at 3.14.

5.4.1.2 Correlation Coefficients and Cronbach’s Alpha for the Task Difficulty Scales

It is not possible to produce a reliability estimate for each individual treatment session as this would include only one item. If the three difficulty scales are considered a set, across treatments, then each pairwise correlation is significant and they fall in the range of 0.38 to 0.55 (Table 5.4.1). Furthermore, alpha in this scenario is 0.73 (Table 5.4.2).

Table 5.4.1: Comparing difficulty scores across treatments

<table>
<thead>
<tr>
<th>Likert Scale by Likert Scale</th>
<th>$r$</th>
<th>Subjects</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty Item #2 by Difficulty Item #1</td>
<td>0.5492</td>
<td>133</td>
<td>0.4181</td>
<td>0.6579</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Difficulty Item #3 by Difficulty Item #1</td>
<td>0.3830</td>
<td>136</td>
<td>0.2294</td>
<td>0.5179</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Difficulty Item #3 by Difficulty Item #2</td>
<td>0.4372</td>
<td>135</td>
<td>0.2896</td>
<td>0.5644</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Table 5.4.2: Difficulty-across-treatments reliability estimate

<table>
<thead>
<tr>
<th>Difficulty Assessment</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined Difficulty Scales</td>
<td>0.7339</td>
</tr>
<tr>
<td>Exclude: Diff. Item #1</td>
<td>0.6112</td>
</tr>
<tr>
<td>Exclude: Diff. Item #2</td>
<td>0.6005</td>
</tr>
<tr>
<td>Exclude: Diff. Item #3</td>
<td>0.7263</td>
</tr>
</tbody>
</table>
5.4.2 Review of Task Difficulty

Recall the second research question:

- How does the strategy for text–image integration in instructional print media for middle school students impact their perception of task difficulty?

Perception of task difficulty is measured through the three treatment performance tests. Treatment session #1 proved less reliable than the others, but still at an acceptable level.

Whatever the impact of text–image integration strategy on science comprehension, its effect on the perception of task difficulty is important. If one strategy makes the task of learning appear more or less difficult, then the selection of text–image integration strategy would impact student confidence and perhaps student engagement.

5.4.3 Mean Differences in Task Difficulty

Mean sequence groups differences for the task difficulty ratings are provided in Table 5.4.3. The means appear to be close. The PP group rated their task the most difficult in two of the treatments. The PS group rated their task the most difficult in one treatment, and the least difficult in the other two.

Each treatment’s task difficulty rating is a single Likert scale with 6 levels, scored from 1 to 6. The higher the value, the more difficult the task was deemed. The greatest mean difference observed occurs in treatments #2 and #3, with a difference of 0.36 between the highest and lowest ratings.
5.4.4 **Hypothesis Tests for Task Difficulty**

Analysis of variance was run for each treatment’s task difficulty scale. In no case are significant differences observed (significance standards for this study are at $P \leq 0.05$).

5.4.5 **Performance as a Factor in Task Difficulty**

How well did the subjects do with their task difficulty ratings? Task difficulty should be a function of the task (test items, information and handout design) and the abilities of the individual completing the task. The latter is hopefully captured in part by differences in the performance test scores. Did subjects who performed poorly recognize this and rate the task more difficult than more successful peers?

Linear regression was performed for combined difficulty rating ($CD$) and combined performance rating ($CP$). Data is plotted in Figure 8.9.1 in the appendix. Summary of fit (Table 8.9.1), ANOVA (Table 8.9.2), and parameter estimate (Table 8.9.3) are also provided.

### Table 5.4.3: Difficulty assessment means by form

<table>
<thead>
<tr>
<th>Combined Treatment Tests</th>
<th>N</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form #1: PP Version</td>
<td>47</td>
<td>3.6596</td>
<td>0.1480</td>
<td>3.3671</td>
<td>3.9521</td>
</tr>
<tr>
<td>Form #1: PS Version</td>
<td>51</td>
<td>3.4706</td>
<td>0.1421</td>
<td>3.1898</td>
<td>3.7514</td>
</tr>
<tr>
<td>Form #1: FI Version</td>
<td>48</td>
<td>3.5417</td>
<td>0.1464</td>
<td>3.2522</td>
<td>3.8311</td>
</tr>
<tr>
<td>All #1 Forms</td>
<td>146</td>
<td>3.5548</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form #2: PP Version</td>
<td>54</td>
<td>2.9630</td>
<td>0.1495</td>
<td>2.6673</td>
<td>3.2586</td>
</tr>
<tr>
<td>Form #2: PS Version</td>
<td>43</td>
<td>2.6047</td>
<td>0.1676</td>
<td>2.2733</td>
<td>2.9360</td>
</tr>
<tr>
<td>Form #2: FI Version</td>
<td>46</td>
<td>2.7391</td>
<td>0.1620</td>
<td>2.4188</td>
<td>3.0595</td>
</tr>
<tr>
<td>All #2 Forms</td>
<td>143</td>
<td>2.7832</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form #3: PP Version</td>
<td>44</td>
<td>3.1591</td>
<td>0.1651</td>
<td>2.8328</td>
<td>3.4854</td>
</tr>
<tr>
<td>Form #3: PS Version</td>
<td>53</td>
<td>3.3019</td>
<td>0.1504</td>
<td>3.0046</td>
<td>3.5992</td>
</tr>
<tr>
<td>Form #3: FI Version</td>
<td>51</td>
<td>2.9412</td>
<td>0.1533</td>
<td>2.6381</td>
<td>3.2442</td>
</tr>
<tr>
<td>All #3 Forms</td>
<td>148</td>
<td>3.1351</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
in the appendix. A significant relationship was found ($P<0.0001$), with the following equation:

$$CD = 12.1648 - (0.1321 \times CP)$$

As expected, the better a subject does on a performance test, the less difficult he or she deems it to be.

5.4.6 Task Difficulty: Conclusion

No significant differences were found for the task difficulty variable. There is nothing to suggest that text–image integration strategy affects perception of task difficulty.

5.5 Text–Image Integration Strategy’s Impact on Interest Level

5.5.1 Reliability Estimates for Interest Level Inventories

The indirect interest level inventories explicitly address a subject’s interest level in the subject matter of a given form. (Each indirect interest level inventory is thus part of a treatment session, encountered after the performance test and difficulty assessment.) Text–image integration strategy is conceived of as a mediating variable in this case. Differences across forms are theorized to represent an indirect measure of interest in the design strategy.

5.5.1.1 Response Values for the Indirect Interest Level Inventories

Each interest level inventory is a set of 4 Likert scales with 5 options per item. A value of 5 on an item indicates the highest interest level, while a 1 indicates the lowest interest level. Each inventory has a score range of 4–20. Histograms for the inventories are presented as Figure 5.5.1.
The highest and lowest possible values were selected for each individual scale (see Table 8.6.10 in the appendix for basic statistics). In each treatment at least one subject gave the highest possible interest rating and at least 8 subjects gave the lowest possible rating. Mean item values range from 2.34 to 2.85, all below the middle “neither agree nor disagree” rating, which is scored as 3.

5.5.1.2 Correlation Coefficients and Cronbach’s Alpha for the Indirect Interest Level Inventories

All pairwise correlations for the individual inventories are significant (Table 8.6.6). The mean pairwise correlations for the treatments are 0.59, 0.61 and 0.70. When all indirect interest items are considered together, the mean pairwise correlation is sensibly lower, at 0.41. Alpha is strong for each indirect interest inventory (Table 5.5.1). Treatment #1 is $\alpha=0.85$, treatment #2 is $\alpha=0.86$, and treatment #3 is $\alpha=0.90$, all easily meeting the requirements for personality scales.
Table 5.5.1: Indirect interest level reliability estimates

<table>
<thead>
<tr>
<th>Interest Level (N=146)</th>
<th>α</th>
<th>Interest Level (N=143)</th>
<th>α</th>
<th>Interest Level (N=148)</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest #1 Inventory</td>
<td>0.8510</td>
<td>Interest #2 Inventory</td>
<td>0.8638</td>
<td>Interest #3 Inventory</td>
<td>0.9013</td>
</tr>
<tr>
<td>Exclude: Item #1.1</td>
<td>0.8295</td>
<td>Exclude: Item #2.1</td>
<td>0.8212</td>
<td>Exclude: Item #3.1</td>
<td>0.8825</td>
</tr>
<tr>
<td>Exclude: Item #1.2</td>
<td>0.7790</td>
<td>Exclude: Item #2.2</td>
<td>0.8210</td>
<td>Exclude: Item #3.2</td>
<td>0.8555</td>
</tr>
<tr>
<td>Exclude: Item #1.3</td>
<td>0.7959</td>
<td>Exclude: Item #2.3</td>
<td>0.8288</td>
<td>Exclude: Item #3.3</td>
<td>0.8629</td>
</tr>
<tr>
<td>Exclude: Item #1.4</td>
<td>0.8362</td>
<td>Exclude: Item #2.4</td>
<td>0.8339</td>
<td>Exclude: Item #3.4</td>
<td>0.8882</td>
</tr>
</tbody>
</table>

5.5.2 Review of Interest in Subject Matter

Recall the third research question:

- How does the strategy for text–image integration in instructional print media for middle school students impact their interest level in subject matter?

Interest level in subject matter was assessed in each of the three treatment sessions. Engagement is an important pedagogical issue, and many students don’t want to be in school. Any design strategy that would increase student interest level in course content would be of great interest. This is considered an indirect measure of interest level in the text–image integration strategy employed.

The individual indirect interest level inventories have good to excellent reliability estimates.

5.5.3 Mean Differences in Indirect Interest Level

Mean differences for indirect interest ratings are presented in Table 5.5.2 (and visualized in ). Subjects using PP forms gave the highest interest ratings twice and the lowest interest rating once. There is no regular pattern in the means. The overall ratings suggest that
the lakes and fossil fuels content was more interesting to the subjects than treatment #1’s divides and drainage basins.

Table 5.5.2: Indirect interest rating means by form

<table>
<thead>
<tr>
<th>Combined Treatment Tests</th>
<th>N</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form #1: PP Version</td>
<td>47</td>
<td>10.1064</td>
<td>0.4987</td>
<td>9.121</td>
<td>11.092</td>
</tr>
<tr>
<td>Form #1: PS Version</td>
<td>51</td>
<td>8.9020</td>
<td>0.4788</td>
<td>7.956</td>
<td>9.848</td>
</tr>
<tr>
<td>Form #1: FI Version</td>
<td>48</td>
<td>10.0625</td>
<td>0.4935</td>
<td>9.087</td>
<td>11.038</td>
</tr>
<tr>
<td>All #1 Forms</td>
<td>146</td>
<td>9.6712</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form #2: PP Version</td>
<td>54</td>
<td>11.0556</td>
<td>0.4537</td>
<td>10.159</td>
<td>11.953</td>
</tr>
<tr>
<td>Form #2: PS Version</td>
<td>43</td>
<td>10.9535</td>
<td>0.5085</td>
<td>9.948</td>
<td>11.959</td>
</tr>
<tr>
<td>Form #2: FI Version</td>
<td>46</td>
<td>10.6957</td>
<td>0.4916</td>
<td>9.724</td>
<td>11.668</td>
</tr>
<tr>
<td>All #2 Forms</td>
<td>143</td>
<td>10.9091</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form #3: PP Version</td>
<td>44</td>
<td>9.5000</td>
<td>0.5721</td>
<td>8.369</td>
<td>10.631</td>
</tr>
<tr>
<td>Form #3: PS Version</td>
<td>53</td>
<td>11.4151</td>
<td>0.5212</td>
<td>10.385</td>
<td>12.445</td>
</tr>
<tr>
<td>Form #3: FI Version</td>
<td>51</td>
<td>10.8627</td>
<td>0.5314</td>
<td>9.813</td>
<td>11.913</td>
</tr>
<tr>
<td>All #3 Forms</td>
<td>148</td>
<td>10.6554</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.5.2: Indirect interest rating means by form visualized
5.5.4  

**Hypothesis Tests for Indirect Interest Level**

No significant differences were found for the first two treatments \((P=0.1411, P=0.8607)\). Treatment #3 did show significant differences for interest level ratings \((P=0.0446)\), and a follow-up Student’s \(t\) test found PS ratings to be significantly higher than PP ratings, with FI not significantly different from either of the others (Table 8.8.6).

5.5.5  

**Indirect Interest Level: Conclusion**

Significant differences were found in treatment #3, where PS was found to produce higher interest ratings than PP. No other differences were found.

5.6  

**Interest in Text–Image Integration Strategy**

5.6.1  

**Reliability Estimate for Direct Interest Inventory**

The direct interest level inventory was presented in a separate instrument that was administered to subjects immediately after they finished the final treatment session. A link in the final treatment instrument, appearing after responses were submitted through a mouse click, took subjects to the direct interest level inventory. For whatever reason, a number of subjects finished the final treatment but failed to complete the direct interest level inventory. In no instance did this appear to be a function of time limit. It is expected that these subjects started and perhaps even completed the inventory, but failed to click “submit.” The resulting sample size for the direct interest level inventory is 137 subjects.

The inventory was meant to include all individual pages (as images) from the treatment sessions. Each session’s form consisted of two pages in a spread. This results in a total of 6 pages. However, the left-hand page of the treatment #1 forms did not display due to
a software error (the missing page corresponds to what would have been item #1.1). Thus there are only 5 pages in consideration.

For each item, the three versions of a given page are displayed, and the subject must either click to identify the most interesting of the three (the “positive” items), or the least interesting of the three (the “negative” items), depending upon the prompt. The 5 positive items appeared in random order on one page. After this page was completed, the 5 negative items appeared in random order on the next page. This means that subjects did not have their positive answers available for review when determining their negative answers.

5.6.1.1 Raw Response Values for the Direct Interest Level Inventory

Response values for the direct interest level inventory are reported in Table 5.6.1. Recall that the positive items identify the “most interesting” of the design alternates, and the negative items identify the “least interesting.” These are raw values, so the strong showing for fully integrated (FI) in the positive set actually corresponds to the “weak” showing for FI in the negative set. That is, low values for FI in the negative set indicate a higher interest quotient.

The trend is towards fully integrated being more interesting than prose subsumed, which in turn is more interesting than prose primary, or FF > PS > PP. The careful observer will note that negative item #3.1 seems odd. It is the only negative item where PP is not the top selection. In fact, only 9 subjects selected PP here (well off the average of 70.6). This oddity precipitated a review of the test instrument as well as the coding process.

The item was properly coded. Where this item is exceptional is in the design itself. Prose primary designs are not prose exclusive designs: they do include imagery and diagrams, the latter of which represent fully integrated zones on the page.
Table 5.6.1: Raw direct interest level ratings

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dir. Int. Pos. #1.2</td>
<td>(20) 0.1460</td>
<td></td>
<td>(28) 0.2044</td>
<td></td>
<td>(89) 0.6496</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>Dir. Int. Pos. #2.1</td>
<td>(25) 0.1825</td>
<td></td>
<td>(45) 0.3285</td>
<td></td>
<td>(67) 0.4891</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>Dir. Int. Pos. #2.2</td>
<td>(9) 0.0657</td>
<td></td>
<td>(26) 0.1898</td>
<td></td>
<td>(102) 0.7445</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>Dir. Int. Pos. #3.1</td>
<td>(39) 0.2847</td>
<td></td>
<td>(17) 0.1241</td>
<td></td>
<td>(81) 0.5912</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>Dir. Int. Pos. #3.2</td>
<td>(12) 0.0876</td>
<td></td>
<td>(23) 0.1679</td>
<td></td>
<td>(102) 0.7445</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>Positive Item Mean</td>
<td>(21.0) 0.1533</td>
<td></td>
<td>(27.8) 0.2029</td>
<td></td>
<td>(88.2) 0.6438</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>Dir. Int. Neg. #1.2</td>
<td>(76) 0.5547</td>
<td></td>
<td>(32) 0.2336</td>
<td></td>
<td>(29) 0.2117</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>Dir. Int. Neg. #2.1</td>
<td>(70) 0.5110</td>
<td></td>
<td>(37) 0.2701</td>
<td></td>
<td>(30) 0.2190</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>Dir. Int. Neg. #2.2</td>
<td>(93) 0.6788</td>
<td></td>
<td>(22) 0.1606</td>
<td></td>
<td>(22) 0.1606</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>Dir. Int. Neg. #3.1</td>
<td>(9) 0.0657</td>
<td></td>
<td>(100) 0.7300</td>
<td></td>
<td>(28) 0.2044</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>Dir. Int. Neg. #3.2</td>
<td>(105) 0.7664</td>
<td></td>
<td>(7) 0.0511</td>
<td></td>
<td>(25) 0.1825</td>
<td></td>
<td>137</td>
</tr>
<tr>
<td>Negative Item Mean</td>
<td>(70.6) 0.5153</td>
<td>(39.6) 0.2891</td>
<td>(26.8) 0.1956</td>
<td></td>
<td>137</td>
<td></td>
<td>137</td>
</tr>
</tbody>
</table>

The PP design displayed for this item includes a particularly large diagram, seeming to overpower the text, counter to the trend in prose primary designs (Figure 5.6.1, arranged such that the PP design is at left and the FI design is at right). When viewed as a full spread, the prose appears more dominant (Figure 8.3.7 in the appendix), but as an isolated page it shares much with a fully integrated design and by comparison has the largest discernable single graphic. This then is the proverbial exception that proves the rule, especially given that the PS design—the most prose-driven of this particular set—takes on the regular profile of the PP designs, with its 100 selections \( \frac{100}{137} \) as least interesting.
Surely then removing negative item #3.1 would increase reliability (yet to be calculated), but doing so might be disingenuous. Prose primary designs do include diagrams, sometimes quite large, and it was determined that this should remain in the data.

Since the direct interest level inventory utilizes categorical data, the data must be adjusted to allow for a reliability estimate akin to that which the other batteries and inventories were subjected. This adjustment is based on the study’s assumption that prose subsumed falls between prose primary and fully integrated in a linear relationship.

5.6.1.2 Adjusted Response Values for the Direct Interest Level Inventory

The adjusted scores for the direct interest level inventory are oriented towards the interest level in the fully integrated strategy. For positive items, selections are scored as follows:
• FI = +1
• PS = 0
• PP = −1

For negative items, the scores are reversed, with FI being −1 and PP being +1. Response values are reported in Table 8.6.11 in the appendix. Negative item #3.1, discussed above, is the only item with a negative mean. The next lowest mean is negative item #2.1 at +0.29. The histogram for the adjusted direct interest level inventory reveals a strong left skew, implying greater interest in the fully integrated strategy (Figure 5.6.2). Subjects are certainly not unanimous in favoring the fully integrated strategy, as 2 total scores of −9 (heavily favoring PP) attest. It must be noted that while a score of 0 on an individual item indicates a selection of PS, a total score for the inventory at or near a mean of 0 does not necessarily indicate a strong preference for PS. While a strong preference for PS should result in an aggregate value near 0, so too would a lack of preference for any strategy. A mixture of indiscriminant selections “for” and “against” PP and FI can yield an individual rating of 0. This adjustment method is best as a comparison of PP and FI strategies.

Figure 5.6.2: Adjusted direct interest level scores
5.6.1.3 Correlation Coefficients and Cronbach’s Alpha for the Adjusted Direct Interest Level Inventory

All correlations for the adjusted direct interest inventory are positive and significant at $P \leq 0.05$ (Table 8.6.7). The mean pairwise correlation for all items is 0.40. The reliability estimate for the adjusted inventory is $\alpha = 0.86$ (Table 5.6.2). Only removal of positive item #3.1 (the same page that stood out earlier in the negative set) would increase the inventory’s reliability estimate. The threshold for personality tests—the appropriate standard for this inventory—is $\alpha = 0.80$.

Table 5.6.2: Adjusted direct interest level reliability estimate

<table>
<thead>
<tr>
<th>Direct Interest Level</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adj. Direct Interest Inventory</td>
<td>0.8649</td>
</tr>
<tr>
<td>Exclude: Pos. Item #1.2</td>
<td>0.8573</td>
</tr>
<tr>
<td>Exclude: Pos. Item #2.1</td>
<td>0.8556</td>
</tr>
<tr>
<td>Exclude: Pos. Item #2.2</td>
<td>0.8488</td>
</tr>
<tr>
<td>Exclude: Pos. Item #3.1</td>
<td>0.8768</td>
</tr>
<tr>
<td>Exclude: Pos. Item #3.2</td>
<td>0.8477</td>
</tr>
<tr>
<td>Exclude: Neg. Item #1.2</td>
<td>0.8436</td>
</tr>
<tr>
<td>Exclude: Neg. Item #2.1</td>
<td>0.8428</td>
</tr>
<tr>
<td>Exclude: Neg. Item #2.2</td>
<td>0.8452</td>
</tr>
<tr>
<td>Exclude: Neg. Item #3.1</td>
<td>0.8585</td>
</tr>
<tr>
<td>Exclude: Neg. Item #3.2</td>
<td>0.8414</td>
</tr>
</tbody>
</table>

As a last inquiry into the direct interest level inventory, the positive and negative items were tallied as individual sets, and these sets were compared. The pair of set scores correlate at $r = 0.6477$. 

179
In summary, the direct interest level inventory appears to represent an internally consistent measure.

5.6.2 Review of Interest in Media

Recall the fourth and final research question:

- How does the strategy for text–image integration in instructional print media for middle school students impact their interest level in instructional print media?

Interest level in design strategy was measured in the post-test. Doing so any earlier would have poisoned the results of the treatment sessions, in that the post-test basically revealed what the study was about in some detail. While during treatment sessions it was certainly obvious to subjects that there were different forms distributed, they weren’t asked to consider those forms. Rather, they had a task to complete and hopefully took the task at face value. If the indirect interest level variable addresses how students actively engage with subject matter, then the direct interest level variable addresses how students approach the material through which subject matter is presented (or more accurately, constructed).

Direct interest level is separate from the treatment sessions. It was measured immediately following treatment session #3, in its own instrument.

5.6.3 Mean Differences in Direct Interest Level

The direct interest level inventory is composed of 10 items. For each item, the subject was shown the three versions of a single form page at thumbnail size. For the first 5 items, the subject selected the most interesting of the three; for the remaining 5 items, the subject selected the least interesting of the three (the same image sets as with the first 5 items).
Each strategy type received its own total score. Scores started at 5. Each “most interesting” selection counts as +1 for the strategy selected. Each “least interesting” selection counts as −1. This results in bounded scores such that the total for the three scores must always equal 15. Individual scores range from 0 to 10. Histograms for the three strategies are presented as Figure 5.6.3. Fully integrated has the strongest showing with its left skew. The other two strategies are skewed right. Simple statistics are presented in Table 5.6.3. The mean value across strategies, due to the bounded scoring, is 5. Fully integrated is above this center point at 7.24. Prose subsumed is just below the center point at 4.57. Prose primary is below at 3.19.

![Histograms for the three strategies](image)

**Figure 5.6.3: Total strategy interest scores**

<table>
<thead>
<tr>
<th>Direct Interest (N=137)</th>
<th>Mean</th>
<th>S.D.</th>
<th>Range</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prose Primary (PP)</td>
<td>3.1898</td>
<td>2.0988</td>
<td>0...10</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Prose Subsumed (PS)</td>
<td>4.5693</td>
<td>1.7690</td>
<td>0...10</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Fully Integrated (FI)</td>
<td>7.2409</td>
<td>3.0619</td>
<td>0...10</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

181
5.6.4 *Direct Interest Level Significance*

Full pairwise correlations for the direct interest level exercise are presented in Table 5.6.4. FI is highly negatively correlated with PP and PS. PP and PS are positively correlated, with a much lower $r$. Each pairwise correlation is significant.

<table>
<thead>
<tr>
<th>Selection</th>
<th>by</th>
<th>Selection</th>
<th>$r$</th>
<th>Subjects</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prose Subsumed</td>
<td>by Prose Primary</td>
<td>0.2479</td>
<td>137</td>
<td>0.0837</td>
<td>0.3991</td>
<td>0.0035</td>
<td></td>
</tr>
<tr>
<td>Fully Integrated</td>
<td>by Prose Primary</td>
<td>-0.8287</td>
<td>137</td>
<td>-0.8748</td>
<td>-0.7677</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>Fully Integrated</td>
<td>by Prose Subsumed</td>
<td>-0.7477</td>
<td>137</td>
<td>-0.8134</td>
<td>-0.6631</td>
<td>&lt;.0001</td>
<td></td>
</tr>
</tbody>
</table>

Confidence intervals (at 95% confidence level) for the three strategies are presented in Table 5.6.5.

<table>
<thead>
<tr>
<th>Forms</th>
<th>Parameter</th>
<th>Estimate</th>
<th>Lower CI</th>
<th>Upper CI</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prose Primary (PP)</td>
<td>Mean</td>
<td>3.1898</td>
<td>2.8352</td>
<td>3.5444</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Std Dev</td>
<td>2.0988</td>
<td>1.8762</td>
<td>2.3817</td>
<td>0.95</td>
</tr>
<tr>
<td>Prose Subsumed (PS)</td>
<td>Mean</td>
<td>4.5693</td>
<td>4.2705</td>
<td>4.8682</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Std Dev</td>
<td>1.7690</td>
<td>1.5814</td>
<td>2.0075</td>
<td>0.95</td>
</tr>
<tr>
<td>Fully Integrated (FI)</td>
<td>Mean</td>
<td>7.2409</td>
<td>6.7236</td>
<td>7.7582</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Std Dev</td>
<td>3.0619</td>
<td>2.7373</td>
<td>3.4747</td>
<td>0.95</td>
</tr>
</tbody>
</table>

None of the intervals overlap, which means that with 95% confidence, the true means are separated. This separation is visualized in Figure 5.6.4. The true mean estimates are quite
far apart. Fully integrated is more interesting than prose subsumed to this population of students, which in turn is more interesting than prose primary.

95% Confidence Intervals

<table>
<thead>
<tr>
<th>PP</th>
<th>PS</th>
<th>FI</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

Figure 5.6.4: Direct interest confidence intervals visualized

5.6.5 Interest in Strategy: Conclusion

There is strong evidence that fully integrated is the most interesting and prose primary is the least interesting of the strategies outlined and addressed in this study (at 95% confidence level). It should also be noted that these results are consistent with the assumption that prose subsumed falls between the other two strategies in a linear relationship.
SECTION 6 — DISCUSSION

6.1 Notes on Comprehension

6.1.1 Format

The measure of comprehension undertaken in this study is based on cognitive load theory (CLT) but cannot elucidate a particular kind of load with accuracy. This is a limitation of CLT—at least currently—that is perhaps compounded by the holistic nature of this study. The implication of comprehension measures here is variability in extraneous cognitive load, however, even if this load cannot be measured directly. This is logical. The forms were carefully constructed so that basic information was equal across them, while the encoding of that information shifts between text and image and the relationships therein shift in correspondence.

It is particularly encouraging that the treatment session with the best reliability estimate—a “good” rating, and not far from excellent—is the one in which comprehension differences were found. This is a suggestive finding, and a call for future work. The first
treatment did not prove reliable, so the finding that the fully integrated strategy (FI) was more comprehensible than the prose primary strategy (PP) occurred for one of only two good opportunities.

Why was the finding of treatment #2 not repeated in treatment #3? Two FI-assigned subjects’ difficulty explanations offer a possible answer regarding the treatment #3 FI form:

All the information was scattered and not organised that well. Yes the pictures did help but it was very confusing and crazy! [All difficulty explanations are direct quotes with no editorial corrections.]

And:

I got distracted by all the pictures and the diagrams were a little too much.

There are actually no equivalent complaints levied at the treatment #2 FI form. While that cannot be conclusive, it’s easy to compare the FI versions for the second and third treatments and agree that the third is especially complicated. This is a function of the inherent complexity of the information embedded on the original textbook source pages.

The fully integrated strategy produces information structures in a parallel format, meaning that all information can be accessed at any time. In contrast, the prose primary strategy produces serial information structures, where an extended passage must be read through to access critical information. Comprehension here is a measure of how well subjects worked with or through the handouts, so superior performance by FI is sensible.

Prose subsumed represents a balance between PP and FI, as one especially thoughtful subject explains:

I thought that working with this handout to answer the survey questions was somewhat difficult because I did not know much at all about
fossil fuels when I first started to read the handout. Also, I find this subject (fossil fuels) to be a little confusing even with the handout to help me out. However, I thought that the handout was very well constructed, because it had a good balance between the plain text and the diagrams/pictures. This handout suited my learning style better than the one we did last Monday about water turnover, lakes, and ponds, because that one had too many pictures and not enough plain text. The pictures and diagrams kept this handout interesting and the words provided large amounts of information as well. This handout’s format made it more simple for me to answer the questions on the survey. Thanks. [Treatment #3, PS form, “somewhat difficult” rating—note the reference to learning styles.]

6.1.2 Factual vs. Conceptual Knowledge

The factual–conceptual knowledge distinction is a compelling one. It seems likely that the text–image integration strategy effect on comprehension is complicated. Measuring factual and conceptual knowledge manipulation separately provides an opportunity to qualify any form-based differences. A parallel structure should make it easier to locate facts than a serial structure by virtue of its multitude of entry points. When constructing a performance test of factual knowledge, however, it is important that test completion not simply hinge on finding answers in the text. Even factual knowledge must be measured through knowledge manipulation, meaning that subjects must have to think and not just seek. One subject lamented this fact:

I think all the questions should have been based on the topic and content in the selection. I also recommend that you make the question easier so you can spot the answer in the text easier.
Conceptual knowledge requires that students discern principles operating beneath otherwise factual information. Some subjects were aware enough of their thought processes (in other words, metacognitive) to recognize this requirement in the performance tests: “The questions did relate to the handout though they were I believe inference questions”; “But some of the questions you did have to think about, as they were not directly in the handout”; and simply, “There was clues that i could find” (clues but not answers per se).

The potential difference in comprehension of factual versus conceptual knowledge is especially interesting if it can be compared to learning, as defined by CLT. Perhaps a subject that does well with factual but not conceptual knowledge performs adequately on a performance test but retains very little.

The factual–conceptual distinction may be heavily impacted by the integration of text and image. The scope of this study makes it difficult to isolate especially subtle differences in cognitive processing as only 5 or 6 items for each knowledge type were included per session. Future work would do well to extend factual and conceptual batteries.

6.1.3 Interactions

Where does increased performance come from? The text–image integration strategy employed is theorized to have a major direct impact. Comprehension may very well interact with other factors, though. Interest in one strategy might promote greater mental effort, resulting in greater working memory resources for germane load on top of whatever extraneous and intrinsic load is necessarily in effect. Text–image integration strategy may impact engagement duration in such a way that one strategy extends engagement, and by virtue of this extension, increases functional comprehension. This would be another example of manipulation of mental effort.

Preference may also play a role. There is evidence in the difficulty explanations (section 8.10) that the subjects have participated in and remain cognizant of some form of learning styles rating. For instance:
I found working with this handout somewhat difficult because I don't think of myself as a visual learner, so looking at all of the diagrams and pictures was a bit challenging for me. However, some people might have been more successful with the pictures and diagrams instead of just plain text on a page. I think that it was a good idea to create a textbook with all of these pictures and diagrams, I had to adjust to all of the pictures and diagrams on the pages but it was nothing major and wasn’t much of a problem. [Treatment #2, FI form, “somewhat difficult” rating.]

This subject rated treatment #2 as “somewhat difficult” despite receiving a perfect score on the test (as well as the other treatment tests). How might other students who are less capable unwittingly retard their own performance because they believe they are visual learners (or otherwise) and are faced with materials they deem an ill fit to their style? This is a potential harm that learning styles—a likely baseless theory in practice (Coffield et al., 2004)—might do to student engagement with instructional media.

Familiarity could play a role in comprehension. This study’s intervention was too brief to address familiarity effects. Neither FI nor PS is used in textbooks systematically (though both appear in isolation periodically). At least one subject equated “textbook style” with prose:

I think that I can get more questions right and understand easier if I read something textbook style. I struggle with diagrams because they are hard for me to take in.

Might a more integrated strategy seem disorienting to students at first but prove more effective with increasing familiarity? One subject did not initially think to seek out answers in diagrams though they were provided:
Some of the information was really hard to find in the selection. Example: what “peat” was. It wasn’t in the text which is where I usually look for information first, and I read the text about 5 times just to make sure I didn’t skip it. And some other information was in the pictures too, which took awhile to make sure it was right. It wasn’t all that difficult though, because the other information that was in the text, I got right away, because I somewhat remembered what it said. The questions were pretty easy to get, and weren’t something that a 7th grader couldn’t understand. I didn’t get some of the words, and had to read over, and use context clues to figure out what they actually meant. [Treatment #3, PS form, “somewhat difficult” rating.]

If this subject were to receive another PS form the following day or week (this particular subject only attended the one treatment session), wouldn’t he or she consider images and diagrams as more viable information sources that second time around? This is a question of familiarity, and in order to answer it, a much more extensive study would be needed.

6.2 *Notes on Task Difficulty*

Subjects were fairly accurate with their difficulty assessments when compared to their test scores. The more difficult a session was in practice (that is, the lower the overall test scores), the more difficult the subjects rated it.

Perception of task difficulty may place limitations on comprehension. PP was associated with more difficult ratings on average than FI in all treatments, though never was this difference found to be significant. This study found no evidence that text–image integration strategy impacts perception of task difficulty.
Each treatment provided a total of 11 test items and 6 levels on the single difficulty rating. This offered too little distinction to perform deep analysis of the relationship between perception of task difficulty and comprehension. What would be possible with finer measures? Of particular interest would be any relationship between a comprehension–difficulty complex and science mastery (here measured with the review test). Does one strategy tend to empower low-mastery students? If one text–image integration strategy renders science content less intimidating to less successful students, then it would have considerable pedagogical implications.

There were clearly dramatic differences in reading ability among the subjects. There is evidence in the difficulty explanations: “I didn't get what the questions were asking me”; “because I am not america people, so I do not understand many word”; “Because english is difficult.” Compare those responses to this pair by subjects who rated the corresponding sessions as “very easy”:

It was very straightforward with the information it was telling me and the answers were very simple ones where you got most of them from the text. The pictures also helped to add another way to get the information.

And:

The questions were easy and the information was organized well. The pictures also helped to get a better feel about the information.

This study provided no means by which to measure relative English language skills. It stands to reason that a fully integrated strategy would provide greater assistance to low English proficiency students than would a more language-driven prose primary strategy by virtue of appearing less difficult (as well as actually being less difficult in practice). But this can only be determined with language proficiency scores.
Indirect interest level represents the impact that text–image integration strategy has on however interesting the subject matter appears. Format, then, can alter in a fundamental way what content is—they are not truly separable. One subject connects design to interest in subject matter:

because the information was clear and it was spread out. The handout had bolded vocabulary words and pictures to keep you interested.

The basic profile of interest level ratings by form was unstable across treatments (see Figure 5.5.2). The prose subsumed strategy produced each of the lowest, middle and highest ratings across the treatments. Prose primary produced the highest ratings twice and the lowest ratings once. For the third treatment, PS produced significantly higher ratings than PP. The inconsistency of the means across treatments suggests that this solitary finding be considered with some hesitation. These findings are not as strong as the comprehension findings.

One might think otherwise when looking at the results and speculate that prose primary produces the highest interest level. The direct interest level ratings, where subjects looked at form page thumbnails and gauged interest in their formal qualities, revealed that the treatment #3 PP form was eccentric in one particular way: it included a large diagram that was actually reduced in the PS and FI versions (recall Figure 5.6.1 and preceding explanation). The PS and FI versions included more images, but there might be something to the size of graphical zones on the page. If in some functional way PS is performing as the text-heavy form within the third set, then PS producing the highest indirect interest ratings in that session is actually consistent with the higher ratings of PP in the earlier treatments (though it’s important to emphasize that those differences were not significant).

The expectation, based in part on the literature review, was that a heavier image load would equate with higher interest in subject matter. This implicit hypothesis is has not been answered definitively either way, but it begs further exploration.
The direct interest ratings, which gauge subject interest in the forms themselves, produced the strongest results of this study. The basic profile of interest ratings in media is one of more positive ratings for higher levels of text–image integration: FI is more interesting than PS is more interesting than PP. This was a strong finding at 95% confidence (and remains so at 99% confidence). Modal values were very close to perfectly spaced extremes within the range of possible values. The behavior of one eccentric item with an atypical PP design appears to be responsible for keeping the results from being perfectly regular, with the modal value for FI at 10, PS at 5, and PP at 0 (these values were actually 10, 4, and 1, respectively, with PP and PS switching for that one item).

One subject, faced with a prose subsumed version, helps to interpret the findings:

…the text wasn’t easy to apeal to my senses. I say this because I do not like to stare at stuff I am not interested in for even a short time.

This is eerily similar to one of Nietzsche’s (2003) maxims:

The more abstract the truth you want to teach the more you must seduce the senses to it. (p. 99)

The least integrated design strategy possible, assuming one wants to embed very specific information, is a prose exclusive strategy. Prose exclusive was not tested here, as it is not a form of text–image integration at all. A prose exclusive design presents information in one of two possible formats alone—propositional, linguistic, text—leaving it up to the reader to produce any kind of mental imagery. Imagery is often included with the purpose of “seducing the senses.”

The direct interest measure is one of seduction. Subjects would typically recognize one of the images as the form they utilized in a treatment session. The strong and consistent
rating of FI over PP suggests that the subjects were comparing the images at hand and not favoring (or penalizing) whatever forms they were assigned.

The findings suggest that as text–image integration (and consequently image load) increases, so too does interest quotient. A fully integrated design is more time-consuming to produce than a prose primary design, but an increase in student interest may justify the extra work.

6.5 Notes on Duration

Duration could not be used as a valid measure in this study as it included the highly variable time spent on the open-ended difficulty explanations. However, duration as a possible measure remains an important consideration. According to cognitive load theory, the issue of text–image integration should be one of efficiency. Does one strategy promote comprehension more efficiently by creating equal or even greater gains with a shorter period of engagement? One strategy may not increase comprehension over another, but may require less time to do so. In other words, a shorter duration may be desirable considering other factors. On the other hand, a longer duration may be desirable as it suggests students becoming more engrossed in their instruction. In future work, should duration be measured in a reliable fashion, it will become important to probe into the relationships between duration and other variables such as comprehension, learning, and interest level. After all, either lesser or greater engagement duration may possibly be desirable.

Duration would have different meaning if a research design presented subjects with too much work to do in a restricted time frame. The question of efficiency and the relationships with comprehension, learning and interest level remain interesting in this scenario, but duration becomes a limiting factor rather than an outcome.
6.6 Conclusion

This study is based in part on cognitive load theory (CLT), though it does not seek to measure any kind of load directly. Comprehension as a measure concerns the portion of mental load functioning as germane load even if germane load itself cannot be measured. The research design seeks to fix intrinsic load across treatment groups, leaving variation as a factor of germane load and extraneous load—assuming equal mental effort. According to CLT, it is theoretically possible to successfully comprehend something but fail to create long term memory structures from it, meaning that comprehension does not directly imply learning, and does not necessarily equate with germane load activity. This is the difficulty of cognitive load studies. Ultimately, comprehension in this study helps in the estimation of extraneous load through comparison of treatment groups.

The findings of this study can be summarized in plain language as such:

1) The fully integrated design strategy (FI) produces more visually interesting design than does the prose subsumed strategy (PS).
2) FI also produces more visually interesting design than does PP.
3) PS produces more visually interesting design than does PP, filling out the simple direct interest relationship: FI > PS > PP.
4) PS appears to represent a middle point between PP and FI in a linear relationship.
5) FI may promote comprehension over PP (there were significant differences in one of three treatments and a consistent profile throughout), though how frequently this occurs remains to be seen.
6) PS may render subject matter more interesting than PP (there were significant differences in one of three treatments), though this was not a consistent relationship across treatments.
7) There is no indication from this study that text–image integration strategy affects perception of task difficulty.
The fully integrated strategy—certainly the most interesting to subjects—may promote comprehension and thus represent a lower extraneous load–producing option for textbook designers. The problem remains that fully integrated designs are labor-intensive. Instead of dismissing them as problematic according to textbook publishing industry standards (and FI may not even be prohibitive in practice, anyway, considering the time investing in textbook development), we might consider a different model for giving teachers access to instructional media. With the advent of dynamic and social Web-based media, another model emerges where discrete materials can be accessible on a per-county, per-school, and even per-teacher basis.

This is a holistic study whose primary construct is design strategy, not some distinct and isolated design-based relationship. It must be recognized as such. This study outlines a way to address instructional media that is consistent with how designers actually produce design, sensitive to design methodology. Every designer—either implicitly or explicitly—creates a meaningful product according to some strategy. We can thus manipulate design products through articulation and dissemination of a coherent, more productive design strategy.

6.7 Limitations

This study has certain practical and situational limitations. These can be summarized as follows:

1) The findings of this study must be limited to middle school students. There is no way to predict how younger or older subject might respond to text–image integration strategies. The subjects in this study are growing up in a time of heavy media saturation. They are accustomed to receiving information in many small bursts. This experience should color their responses to text–image
integration strategy, possibly in ways inconsistent with other people whose developmental media consumption is in some way dissimilar.

2) The participating middle school is a magnet school. Admission is not based on ability, so there is no expectation that this sample of students is a misrepresentation of North Carolina or American middle school students in general, but certainly there are types of schools and this is but one type.

3) There was no measure of English language proficiency. Random assignment to groups suggests that this should not be a major problem. Rather, this might be characterized as a missed opportunity.

4) The review test measured science learning through content covered in the first three quarters of the school year. Since there was no control over the conditions of the original instruction on which item completion was predicated, an item count of 10 might be considered low.

5) The factual–conceptual item distinction was relatively weak (per reliability estimates) largely due to the low item count of 5 or 6 per treatment. Differential performance with these types of knowledge is a valuable line of inquiry and deserves greater depth.

6) Overall there were too few test items to isolate more subtle relationships. Increasing the number of items would only be possible if the treatment forms were also expanded. As such this limitation is simply a factor of the scope of this study.

7) The first treatment did not prove reliable. This hurt generalizability, as there were only two chances to find differences in the dependent variables.

8) Even if the first treatment proved reliable, a count of three treatment sessions offers minimal opportunities to isolate differences. Increasing the session count was impossible for this study, both due to availability at the participating school and the workload involved in preparing a full treatment session.
6.8 Strengths

Despite any limitations of this study, its strengths are legion. Some points of interest are:

1) Subjects completed the tasks, which were designed to be similar to typical classroom activity, in their regular science classrooms, at their own desks. The result is high ecological validity.

2) This study had a large sample size of 158 subjects. This decreases the chances of sampling error.

3) The interest-in-instructional-media findings are definitive, providing stark contrast between the three text–image integration strategies.

4) The interest-in-instructional-media findings support the basic framework of separate strategies in a linear relationship, with the prose subsumed strategy falling somewhere between fully integrated and prose primary at the extremes.

5) This study establishes a repeatable school-integrated research model. It is difficult to gain access to young subjects. This research model makes it easier to convince a school to participate: treatment sessions can double as regular science instruction. (Though in the case of this particular study, 7th grade subjects were being tested with material from the 8th grade curriculum.)

6.9 Future Work

As with all research, this study suggests future directions for inquiry into the variables of interest.

1) Engagement duration with media represents a variable of great interest. Duration could not be measured accurately in this study. Future work would
do well to measure the time subjects spend completing a battery of test items. This would be possible using the same instrumentation, programming time stamps into the opening of the first screen of test items and the closing of the final screen of test items. Measuring engagement with individual items, though interesting, would be far more difficult.

2) Future studies might increase test item count in order to capture more subtle relationships. Doing so would involve a significant increase in workload during development.

3) Along the same lines, an increase in test duration—and especially an increase in treatment sessions—would provide better hope of isolating differential comprehension with instructional media.

4) Future studies might include a measure of English language proficiency. The conversion of English-based information into imagery, when producing PS or FI forms, should render media more approachable for low English proficiency students. This assertion should be tested.

5) This research design exposed each subject to each design strategy once in rotation. Future work might assign subjects a single strategy in order to test for any familiarity effects. One strategy may be disorienting at first but prove more effective with increased familiarity.

6) If testing occurred some time after engagement with the instructional media, then learning could be measured. A comparison of comprehension and learning with one set of media would serve to better appreciate the actual function of text–image integration strategy.

7) The researcher alone developed all the forms used in this study. The employ of multiple designers would make for a stronger argument that any effects are the result of the underlying text–image integration strategy and not the product of any kind of designer bias.

8) None of the issues addressed in this study are confined to print-based media. Everything should apply to explicitly interactive screen-based media as well.
Future work could utilize screen-based media. Doing so would surely expose new and important issues.

The issues raised by this study are important for educational practice. If science is worthwhile teaching—and it most certainly is—then the text and images with which students construct science knowledge are worth the considerable efforts required for closer inspection.
SECTION 7 — REFERENCES


Mayer, R. E. (2003). The promise of multimedia learning: Using the same instructional
design methods across different media. *Learning and Instruction, 13*, 125–139.

design: Using annotated illustrations to foster meaningful learning of science text.
*Educational Technology and Research Development, 43*(1), 31–43.


McConkie, G. W. (1997). Eye movement contingent display control: Personal reflections and

Houghton Mifflin Company.

Houghton Mifflin Company.


Moreno, R. (2006). When worked examples don’t work: Is cognitive load theory at an
impasse?. *Learning and Instruction, 16*, 170–181.

*Educational Technology Research and Development, 27*(1), 45–53.


210


Figure 8.1.1: Bloom’s Revised schematic
Figure 8.1.2: The New Taxonomy schematic
8.2  

**Text–Image Integration Strategies**

![Three strategies](image)

**Prose Primary**

**Prose Subsumed**

**Fully Integrated**

**Figure 8.2.1:** Three strategies (modified from McDougal Littell, 2005a, C18–C19)
8.3 Final Test Forms

8.3.1 Treatment #1 Forms

Figure 8.3.1: PP form #1 (modified from McDougal Littell, 2005a, pp. C16–C17)
1.2 Fresh water flows and freezes on Earth.

**Before, you learned...**
- Water moves in parts of Earth’s ocean.
- Water continually cycles from Earth’s surface to the atmosphere.

**Now, you will learn...**
- How fresh water flows and collects on land.

How does water flow and collect on land?

Imagine you are in a kayak in a river, paddling through white-water rapids. Your kayak splashes around hazards, carving its way downstream. Then the water gets wider and its current decreases. You glide across the surface, slowing down. When before you had to fight against the strong current, now you can paddle to move swiftly.

In your kayak you are following the path a water drop might take on its way to the oceans. All over the planet, the flow of gravity makes water downhill. Fresh water flows downhill in a series of streams and rivers, collects in lakes and ponds, and eventually flows into the oceans. All of this water flows between high points called divides, in areas called drainage basins.

**Figure 8.3.2: PS form #1 (modified from McDougal Littell, 2005a, pp. C16–C17)**
Figure 8.3.3: FI form #1 (modified from McDougal Littell, 2005a, pp. C16–C17)
8.3.2 Treatment #2 Forms

Surface water collects in ponds and lakes.

Lakes and ponds form where water naturally collects in low parts of land. Some lakes were formed during the last ice age. For example, the Great Lakes were formed when huge sheets of ice scraped out large areas of glacial depressions. Other lakes, such as Cramer Lake in Oregon, were formed when water collected inside the crater of a meteorite impact.

Water can fill a lake in several ways. Where the land surface dips below the level of underground water, the low land area fills with water. Rainfall and other precipitation contribute water to all lakes. Water may flow from a lake to a stream or river. Water may also flow away from a lake through a stream running downhill from the lake. Many lakes experience only steady levels because of the balance of flow in and flow out.

The main difference between a pond and a lake is in their overall size. A pond is smaller and shallower than a lake, and there are many plants, such as water lilies and water milfoil, rooted in its muddy bottom. A lake may have water so deep that sunlight can’t reach the bottom. In the deeper part of the lake, plants can’t take root, so they grow only around the lake’s edges. Ponds and lakes provide homes for many kinds of fish, insects, and other wildlife. They also provide resting places for migrating birds.

Lake trout

The water in a lake is not as deep as it might appear. The changing temperatures of the seasons affect the water and cause it to move within the lake in a yearly cycle.

In a place with cold winters, ice may form on a lake, so that the wind cannot ruffle the surface. The water temperature in the lake remains steady, and the water stays mixed. The water just below the surface ice is near freezing, so the fish move to the bottom, where the water is a bit warmer.

In many lakes the water temperature at different levels vary as the seasons change. In the spring and summer, heat from the horizon warms a layer of water at the top of the lake. Because the colder water beneath the top layer is denser than the warmer water above it, the water levels do not rise much. The warmer water contains more oxygen, so fish may be more plentiful in the upper part of the lake.

In the fall, deep cool and the surface waters mix too. The upper layer becomes heavy and sinks, so that the lake “closes.” Nutrients from animals and from dead plants and organisms are stirred up from the bottom. These nutrients are used by many life forms in the lake. The rising and sinking of cold and warm water layers in a lake is called meromixis. Temperature changes below each year as the seasons change.

Figure 8.3.4: PP form #2 (modified from McDougal Littell, 2005a, pp. C18–C19)
Figure 8.3.5: PS form #2 (modified from McDougal Littell, 2005a, pp. C18–C19)
Figure 8.3.6: FI form #2 (modified from McDougal Littell, 2005a, pp. C18–C19)
### 8.3.3 Treatment #3 Forms

**Fossil fuels supply most of society’s energy.**

When you turn on the air conditioner, a computer, or a microwave oven, you may use energy from fossil fuels. Billions of people depend on these fuels—coal, oil, and natural gas—for electricity, heat, and fuel.

A **fossil fuel** is an energy source formed from ancient plants and animals buried in Earth’s crust millennia of years. The energy in each fuel represents a form of stored sunlight, since ancient organisms depended on the sun. The buried organisms form layers at the bottom of oceans, ponds, and swamps. Over a long time, this material is compressed and pushed deeper into Earth’s crust. High heat and pressure change it chemically into coal, oil, and natural gas.

**Coal** is a solid fossil fuel formed underground from heated and compressed plant material. As shown below, heat and pressure determine the type of coal formed. The hardest coal makes the best energy source if burned hotter and much cleaner than sulfur coal.

At one time, coal was the main source of energy in the United States.

The world’s largest coal deposits are in the United States, Russia, and China. People use surface mining and deep mining to obtain coal. In surface mining, overlying soil is stripped away to expose the coal. In deep mines, miners must go underground to dig out the coal. Most of the world’s coal is used to fuel power plants and to run factories that produce steel and cement.

When burned as a fuel, however, coal produces hazardous substances that pollute air and water. Also, surface mining can destroy entire landscapes. Coal dust in deep mines damages miners’ lungs. Reducing pollution, renewing landscapes, and protecting miners cost millions of dollars. Society faces a difficult choice—keep the cost of energy low or raise the price to protect the environment and human health.

**Figure 8.3.7: PP form #3 (modified from McDougal Littell, 2005a, pp. A86–A87)**

- **S** "Fossil fuels supply most of society’s energy.**
- **S** "When you turn on the air conditioner, a computer, or a microwave oven, you may use energy from fossil fuels. Billions of people depend on these fuels—coal, oil, and natural gas—for electricity, heat, and fuel.
- **S** "A **fossil fuel** is an energy source formed from ancient plants and animals buried in Earth’s crust millennia of years. The energy in each fuel represents a form of stored sunlight, since ancient organisms depended on the sun. The buried organisms form layers at the bottom of oceans, ponds, and swamps. Over a long time, this material is compressed and pushed deeper into Earth’s crust. High heat and pressure change it chemically into coal, oil, and natural gas.
- **S** "**Coal** is a solid fossil fuel formed underground from heated and compressed plant material. As shown below, heat and pressure determine the type of coal formed. The hardest coal makes the best energy source if burned hotter and much cleaner than sulfur coal.
- **S** "At one time, coal was the main source of energy in the United States.
- **S** "The world’s largest coal deposits are in the United States, Russia, and China. People use surface mining and deep mining to obtain coal. In surface mining, overlying soil is stripped away to expose the coal. In deep mines, miners must go underground to dig out the coal. Most of the world’s coal is used to fuel power plants and to run factories that produce steel and cement.
- **S** "When burned as a fuel, however, coal produces hazardous substances that pollute air and water. Also, surface mining can destroy entire landscapes. Coal dust in deep mines damages miners’ lungs. Reducing pollution, renewing landscapes, and protecting miners cost millions of dollars. Society faces a difficult choice—keep the cost of energy low or raise the price to protect the environment and human health.

**Figure 8.3.7: PP form #3 (modified from McDougal Littell, 2005a, pp. A86–A87)**
Fossil fuels supply most of society’s energy.

When you turn on the air conditioner, a computer or a microwave oven, you use energy from fossil fuels. Billions of people depend on these fuels—coal, oil, and natural gas—for electricity, heat, and fuel.

A fossil fuel is an energy source formed from ancient plants and animals buried in Earth’s crust. For millions of years, the energy in such a fuel represents a form of stored sunlight, since ancient organisms depended on the sun. The buried organisms form layers at the bottom of oceans, ponds, and swamps. Over a long time, the material is compressed and pushed deeper into Earth’s crust. High heat and pressure change it chemically into coal, oil, and natural gas.

**How Are Fossil Fuels Formed?**

Fossil fuels form easily and produce a lot of heat. They use sunlight to grow most of the power plants that generate electricity that fuels a burning fuel is used to change water into steam. The steam turns a turbine. The turbine drives a generator to produce electricity, which is carried through power lines to homes and cities. Electricity runs nearly everything in modern life, from giant ovens to the smallest jukebox in your home.

As these resources also have a serious drawback. fossil fuels produce massive carbon dioxide, harmful acids, and other forms of pollution. Most of this pollution comes from power plants and fossil fuel burned by cars and other vehicles.

**Coal Formation**

Coal is a solid fossil fuel formed underground from buried and decayed plant material. Heat and pressure determine the type of coal formed. The region is due to the heat and pressure. The solid fuels released during the coal formation process. Most of this pollution comes from power plants and fossil fuel burned by cars and other vehicles.

Figure 8.3.8: PS form #3 (modified from McDougal Littell, 2005a, pp. A86–A87)
Figure 8.3.9: FI form #3 (modified from McDougal Littell, 2005a, pp. A86–A87)
8.4 Instruments

8.4.1 Pre-Test Instrument

Figure 8.4.1: Pre-test interface screen capture
Figure 8.4.2: Further pre-test interface screens
8.4.2  

Treatment #1 Test Instrument

Figure 8.4.3: Treatment #1 test interface screen capture
Figure 8.4.4: Further treatment #1 test interface screens
8.4.3 Treatment #2 Test Instrument

Figure 8.4.5: Treatment #2 test interface screen capture
Figure 8.4.6: Further treatment #2 test interface screens
8.4.4 Treatment #3 Test Instrument

Figure 8.4.7: Treatment #3 test interface screen capture
Figure 8.4.8: Further treatment #3 test interface screens
8.4.5 Post-Test Instrument

MP#5 / April 18 (post)
Opinions (part 1)

Quickly react to the design of these textbook pages. Select the one you find to be the most interesting. There are 6 items on this page and an additional 6 items on the next page. This is page 1 of 2.

1. Which textbook design is the MOST interesting? *

2. Which textbook design is the MOST interesting? *

Figure 8.4.9: Post-test interface screen capture
Figure 8.4.10: Further post-test interface screens
### 8.5 Item Analysis Summary

#### Table 8.5.1: Factual vs. conceptual set difficulty

<table>
<thead>
<tr>
<th>Full Item Sets</th>
<th>Total Mean</th>
<th>Mean Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual #1 Set</td>
<td>3.8493</td>
<td>−0.0586</td>
</tr>
<tr>
<td>Factual #2 Set</td>
<td>4.8881</td>
<td>+0.9802</td>
</tr>
<tr>
<td>Factual #3 Set</td>
<td>4.5608</td>
<td>+0.6529</td>
</tr>
<tr>
<td>Conceptual #1 Set</td>
<td>2.6849</td>
<td>−1.2229</td>
</tr>
<tr>
<td>Conceptual #2 Set</td>
<td>3.9371</td>
<td>+0.0292</td>
</tr>
<tr>
<td>Conceptual #3 Set</td>
<td>3.5270</td>
<td>−0.3808</td>
</tr>
</tbody>
</table>

#### Table 8.5.2: Item interrogation exclusions

<table>
<thead>
<tr>
<th>Item Excluded</th>
<th>Condition for Removal from Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual Item #1.5</td>
<td>Fair rating ((D&lt;0.30)) in both discrimination assessments</td>
</tr>
<tr>
<td>Factual Item #2.2</td>
<td>Poor rating ((D&lt;0.11)) in item discrimination index</td>
</tr>
<tr>
<td>Conceptual Item #3.3</td>
<td>Fair rating ((D&lt;0.30)) in both discrimination assessments; low face validity</td>
</tr>
</tbody>
</table>


8.6 Item Reliability

8.6.1 Correlation Matrices

Table 8.6.1: Science self-efficacy correlations (all significant at $P \leq .05$)

<table>
<thead>
<tr>
<th></th>
<th>SSE #1</th>
<th>SSE #2</th>
<th>SSE #3</th>
<th>SSE #4</th>
<th>SSE #5</th>
<th>SSE #6</th>
<th>SSE #7</th>
<th>SSE #8</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE #1</td>
<td>0.4383</td>
<td>0.3059</td>
<td>0.5197</td>
<td>0.4181</td>
<td>0.1955</td>
<td>0.3485</td>
<td>0.4433</td>
<td></td>
</tr>
<tr>
<td>SSE #2</td>
<td>0.4383</td>
<td>0.4464</td>
<td>0.4058</td>
<td>0.4936</td>
<td>0.3426</td>
<td>0.3926</td>
<td>0.4639</td>
<td></td>
</tr>
<tr>
<td>SSE #3</td>
<td>0.3059</td>
<td>0.4464</td>
<td>0.5641</td>
<td>0.4146</td>
<td>0.3378</td>
<td>0.5629</td>
<td>0.4217</td>
<td></td>
</tr>
<tr>
<td>SSE #4</td>
<td>0.5197</td>
<td>0.4058</td>
<td>0.5641</td>
<td>0.4964</td>
<td>0.4030</td>
<td>0.4985</td>
<td>0.6467</td>
<td></td>
</tr>
<tr>
<td>SSE #5</td>
<td>0.4181</td>
<td>0.4936</td>
<td>0.4146</td>
<td>0.4964</td>
<td>0.4204</td>
<td>0.3775</td>
<td>0.4376</td>
<td></td>
</tr>
<tr>
<td>SSE #6</td>
<td>0.1955</td>
<td>0.3426</td>
<td>0.3378</td>
<td>0.4030</td>
<td>0.4204</td>
<td>0.4831</td>
<td>0.4000</td>
<td></td>
</tr>
<tr>
<td>SSE #7</td>
<td>0.3485</td>
<td>0.3926</td>
<td>0.5629</td>
<td>0.4985</td>
<td>0.3775</td>
<td>0.4831</td>
<td>0.4207</td>
<td></td>
</tr>
<tr>
<td>SSE #8</td>
<td>0.4433</td>
<td>0.4639</td>
<td>0.4217</td>
<td>0.6467</td>
<td>0.4376</td>
<td>0.4000</td>
<td>0.4207</td>
<td></td>
</tr>
</tbody>
</table>
Table 8.6.2: Review test correlations (bold is significant at $P \leq 0.05$)

<table>
<thead>
<tr>
<th></th>
<th>Rev. 1</th>
<th>Rev. 2</th>
<th>Rev. 3</th>
<th>Rev. 4</th>
<th>Rev. 5</th>
<th>Rev. 6</th>
<th>Rev. 7</th>
<th>Rev. 8</th>
<th>Rev. 9</th>
<th>Rev. 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rev. 1</td>
<td>0.1319</td>
<td>0.3588</td>
<td>0.1210</td>
<td>0.1087</td>
<td>0.0294</td>
<td>0.2532</td>
<td>0.2197</td>
<td>0.1537</td>
<td>0.2376</td>
<td></td>
</tr>
<tr>
<td>Rev. 2</td>
<td>0.1319</td>
<td>0.1884</td>
<td>0.2425</td>
<td>-0.0294</td>
<td>0.1100</td>
<td>0.1216</td>
<td>-0.1142</td>
<td>0.1274</td>
<td>-0.0870</td>
<td></td>
</tr>
<tr>
<td>Rev. 3</td>
<td>0.3588</td>
<td>0.1884</td>
<td>-0.0555</td>
<td>0.0793</td>
<td>0.1298</td>
<td>0.1121</td>
<td>0.1873</td>
<td>0.0193</td>
<td>0.0351</td>
<td></td>
</tr>
<tr>
<td>Rev. 4</td>
<td>0.1210</td>
<td>0.2425</td>
<td>-0.0555</td>
<td>-0.1180</td>
<td>0.0555</td>
<td>0.1345</td>
<td>0.0534</td>
<td>0.0950</td>
<td>0.0126</td>
<td></td>
</tr>
<tr>
<td>Rev. 5</td>
<td>0.1087</td>
<td>-0.0294</td>
<td>0.0793</td>
<td>-0.1180</td>
<td>0.0662</td>
<td>0.2571</td>
<td>0.2808</td>
<td>0.1299</td>
<td>0.0686</td>
<td></td>
</tr>
<tr>
<td>Rev. 6</td>
<td>0.0294</td>
<td>0.1100</td>
<td>0.1298</td>
<td>0.0555</td>
<td>0.0662</td>
<td>-0.0008</td>
<td>0.0762</td>
<td>0.0761</td>
<td>0.1850</td>
<td></td>
</tr>
<tr>
<td>Rev. 7</td>
<td>0.2532</td>
<td>0.1216</td>
<td>0.1121</td>
<td>0.1345</td>
<td>0.2571</td>
<td>-0.0008</td>
<td>0.2531</td>
<td>0.1189</td>
<td>0.2034</td>
<td></td>
</tr>
<tr>
<td>Rev. 8</td>
<td>0.2197</td>
<td>-0.1142</td>
<td>0.1873</td>
<td>0.0534</td>
<td>0.2808</td>
<td>0.0762</td>
<td>0.2531</td>
<td>0.1094</td>
<td>0.2812</td>
<td></td>
</tr>
<tr>
<td>Rev. 9</td>
<td>0.1537</td>
<td>0.1274</td>
<td>0.0193</td>
<td>0.0950</td>
<td>0.1299</td>
<td>0.0761</td>
<td>0.1189</td>
<td>0.1094</td>
<td>0.0760</td>
<td></td>
</tr>
<tr>
<td>Rev. 10</td>
<td>0.2376</td>
<td>-0.0870</td>
<td>0.0351</td>
<td>0.0126</td>
<td>0.0686</td>
<td>0.1850</td>
<td>0.2034</td>
<td>0.2812</td>
<td>0.0760</td>
<td></td>
</tr>
</tbody>
</table>

Table 8.6.3: Treatment #1 correlations ($P \leq 0.05$)

<table>
<thead>
<tr>
<th></th>
<th>Fact. 1.1</th>
<th>Fact. 1.2</th>
<th>Fact. 1.3</th>
<th>Fact. 1.4</th>
<th>Fact. 1.6</th>
<th>Con. 1.1</th>
<th>Con. 1.2</th>
<th>Con. 1.3</th>
<th>Con. 1.4</th>
<th>Con. 1.5</th>
<th>Con. 1.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fact. 1.1</td>
<td>0.1978</td>
<td>0.2543</td>
<td>-0.0274</td>
<td>0.0560</td>
<td>0.1941</td>
<td>-0.0040</td>
<td>0.1449</td>
<td>0.0974</td>
<td>0.0859</td>
<td>0.0215</td>
<td></td>
</tr>
<tr>
<td>Fact. 1.2</td>
<td>0.1978</td>
<td>0.0636</td>
<td>0.1325</td>
<td>0.1141</td>
<td>0.0554</td>
<td>-0.0217</td>
<td>0.0349</td>
<td>0.1985</td>
<td>0.1568</td>
<td>0.1531</td>
<td></td>
</tr>
<tr>
<td>Fact. 1.3</td>
<td>0.2543</td>
<td>0.0636</td>
<td>0.0917</td>
<td>0.1871</td>
<td>0.2343</td>
<td>0.1291</td>
<td>0.1550</td>
<td>0.1557</td>
<td>0.1673</td>
<td>0.1130</td>
<td></td>
</tr>
<tr>
<td>Fact. 1.4</td>
<td>-0.0274</td>
<td>0.1325</td>
<td>0.0917</td>
<td>0.1110</td>
<td>0.0792</td>
<td>0.0454</td>
<td>0.0822</td>
<td>0.1469</td>
<td>0.0309</td>
<td>0.1154</td>
<td></td>
</tr>
<tr>
<td>Fact. 1.6</td>
<td>0.0560</td>
<td>0.1141</td>
<td>0.1871</td>
<td>0.1110</td>
<td>0.0433</td>
<td>0.2371</td>
<td>0.2264</td>
<td>0.1269</td>
<td>0.2309</td>
<td>0.2229</td>
<td></td>
</tr>
<tr>
<td>Con. 1.1</td>
<td>0.1941</td>
<td>0.0554</td>
<td>0.2343</td>
<td>0.0792</td>
<td>0.0433</td>
<td>0.1609</td>
<td>0.1116</td>
<td>0.2137</td>
<td>0.0031</td>
<td>0.2010</td>
<td></td>
</tr>
<tr>
<td>Con. 1.2</td>
<td>-0.0040</td>
<td>-0.0217</td>
<td>0.1291</td>
<td>0.0454</td>
<td>0.2371</td>
<td>0.1609</td>
<td>0.0685</td>
<td>0.1517</td>
<td>0.1174</td>
<td>0.1358</td>
<td></td>
</tr>
<tr>
<td>Con. 1.3</td>
<td>0.1449</td>
<td>0.0349</td>
<td>0.1550</td>
<td>0.0822</td>
<td>0.2264</td>
<td>0.1116</td>
<td>0.0685</td>
<td>0.0451</td>
<td>0.1737</td>
<td>0.3785</td>
<td></td>
</tr>
<tr>
<td>Con. 1.4</td>
<td>0.0974</td>
<td>0.1985</td>
<td>0.1557</td>
<td>0.1469</td>
<td>0.1269</td>
<td>0.2137</td>
<td>0.1517</td>
<td>0.0451</td>
<td>0.1396</td>
<td>0.2562</td>
<td></td>
</tr>
<tr>
<td>Con. 1.5</td>
<td>0.0859</td>
<td>0.1568</td>
<td>0.1673</td>
<td>0.0309</td>
<td>0.2309</td>
<td>0.0031</td>
<td>0.1174</td>
<td>0.1737</td>
<td>0.1396</td>
<td>0.2252</td>
<td></td>
</tr>
<tr>
<td>Con. 1.6</td>
<td>0.0215</td>
<td>0.1531</td>
<td>0.1130</td>
<td>0.1154</td>
<td>0.2229</td>
<td>0.2010</td>
<td>0.1358</td>
<td>0.3785</td>
<td>0.2562</td>
<td>0.2252</td>
<td></td>
</tr>
</tbody>
</table>
### Table 8.6.4: Treatment #2 correlations ($P \leq .05$)

<table>
<thead>
<tr>
<th>Fact. 2.1</th>
<th>Fact. 2.3</th>
<th>Fact. 2.4</th>
<th>Fact. 2.5</th>
<th>Fact. 2.6</th>
<th>Con. 2.1</th>
<th>Con. 2.2</th>
<th>Con. 2.3</th>
<th>Con. 2.4</th>
<th>Con. 2.5</th>
<th>Con. 2.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1875</td>
<td>0.3064</td>
<td>0.3266</td>
<td>0.2133</td>
<td>0.2869</td>
<td>0.3982</td>
<td>0.3581</td>
<td>0.0711</td>
<td>0.1924</td>
<td>0.2869</td>
<td></td>
</tr>
<tr>
<td>0.1875</td>
<td>0.3561</td>
<td>0.3174</td>
<td>0.2610</td>
<td>0.2443</td>
<td>0.2531</td>
<td>0.0556</td>
<td>0.2361</td>
<td>0.2275</td>
<td>0.3268</td>
<td></td>
</tr>
<tr>
<td>0.3064</td>
<td>0.3561</td>
<td>0.2884</td>
<td>0.2269</td>
<td>0.1680</td>
<td>0.4392</td>
<td>0.1656</td>
<td>0.1530</td>
<td>0.2793</td>
<td>0.2117</td>
<td></td>
</tr>
<tr>
<td>0.3266</td>
<td>0.3174</td>
<td>0.2884</td>
<td>0.2964</td>
<td>0.2269</td>
<td>0.2022</td>
<td>0.1594</td>
<td>0.1878</td>
<td>0.2481</td>
<td>0.3041</td>
<td></td>
</tr>
<tr>
<td>0.2133</td>
<td>0.2610</td>
<td>0.2964</td>
<td>0.2269</td>
<td>0.2022</td>
<td>0.1594</td>
<td>0.1878</td>
<td>0.2481</td>
<td>0.3041</td>
<td>0.3352</td>
<td></td>
</tr>
<tr>
<td>0.2869</td>
<td>0.2443</td>
<td>0.2371</td>
<td>0.1680</td>
<td>0.2022</td>
<td>0.1521</td>
<td>0.3272</td>
<td>0.1804</td>
<td>0.3434</td>
<td>0.2739</td>
<td></td>
</tr>
<tr>
<td>0.3982</td>
<td>0.2531</td>
<td>0.2490</td>
<td>0.4392</td>
<td>0.1594</td>
<td>0.2320</td>
<td>0.2920</td>
<td>0.2829</td>
<td>0.3680</td>
<td>0.2320</td>
<td></td>
</tr>
<tr>
<td>0.3581</td>
<td>0.0556</td>
<td>0.1407</td>
<td>0.1656</td>
<td>0.1878</td>
<td>0.1521</td>
<td>0.2920</td>
<td>0.1308</td>
<td>0.2661</td>
<td>0.2739</td>
<td></td>
</tr>
<tr>
<td>0.0711</td>
<td>0.2361</td>
<td>0.2256</td>
<td>0.1530</td>
<td>0.2481</td>
<td>0.3272</td>
<td>0.2829</td>
<td>0.1308</td>
<td>0.3628</td>
<td>0.2259</td>
<td></td>
</tr>
<tr>
<td>0.1924</td>
<td>0.2275</td>
<td>0.2175</td>
<td>0.2793</td>
<td>0.3041</td>
<td>0.1804</td>
<td>0.3680</td>
<td>0.2061</td>
<td>0.3628</td>
<td>0.2809</td>
<td></td>
</tr>
<tr>
<td>0.2869</td>
<td>0.3268</td>
<td>0.2371</td>
<td>0.2117</td>
<td>0.3352</td>
<td>0.3434</td>
<td>0.2320</td>
<td>0.2739</td>
<td>0.2259</td>
<td>0.2809</td>
<td></td>
</tr>
</tbody>
</table>

### Table 8.6.5: Treatment #3 correlations ($P \leq .05$)

<table>
<thead>
<tr>
<th>Fact. 3.1</th>
<th>Fact. 3.2</th>
<th>Fact. 3.3</th>
<th>Fact. 3.4</th>
<th>Fact. 3.5</th>
<th>Fact. 3.6</th>
<th>Con. 3.1</th>
<th>Con. 3.2</th>
<th>Con. 3.4</th>
<th>Con. 3.5</th>
<th>Con. 3.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3259</td>
<td>0.0943</td>
<td>0.0808</td>
<td>0.3311</td>
<td>0.1566</td>
<td>0.1393</td>
<td>0.1256</td>
<td>0.2553</td>
<td>0.1137</td>
<td>0.2672</td>
<td></td>
</tr>
<tr>
<td>0.2359</td>
<td>0.2128</td>
<td>0.1173</td>
<td>0.0623</td>
<td>0.1558</td>
<td>0.0545</td>
<td>0.2041</td>
<td>0.2205</td>
<td>0.1534</td>
<td>0.1503</td>
<td></td>
</tr>
<tr>
<td>0.0943</td>
<td>0.2128</td>
<td>0.0757</td>
<td>0.1722</td>
<td>0.1118</td>
<td>0.3170</td>
<td>0.1722</td>
<td>0.2598</td>
<td>0.2680</td>
<td>0.2530</td>
<td></td>
</tr>
<tr>
<td>0.0808</td>
<td>0.1173</td>
<td>0.0757</td>
<td>-0.0524</td>
<td>0.2943</td>
<td>0.2451</td>
<td>0.1483</td>
<td>0.1519</td>
<td>0.1882</td>
<td>0.1868</td>
<td></td>
</tr>
<tr>
<td>0.3311</td>
<td>0.0623</td>
<td>0.1722</td>
<td>-0.0524</td>
<td>-0.0076</td>
<td>0.1755</td>
<td>0.0512</td>
<td>0.2997</td>
<td>0.1179</td>
<td>0.1669</td>
<td></td>
</tr>
<tr>
<td>0.1566</td>
<td>0.1558</td>
<td>0.1118</td>
<td>0.2943</td>
<td>-0.0076</td>
<td>0.0860</td>
<td>0.1409</td>
<td>0.1890</td>
<td>0.1534</td>
<td>0.2260</td>
<td></td>
</tr>
<tr>
<td>0.1393</td>
<td>0.0545</td>
<td>0.3170</td>
<td>0.2451</td>
<td>0.1755</td>
<td>0.0860</td>
<td>0.3359</td>
<td>0.2840</td>
<td>0.4518</td>
<td>0.2329</td>
<td></td>
</tr>
<tr>
<td>0.1256</td>
<td>0.2041</td>
<td>0.1722</td>
<td>0.1483</td>
<td>0.0512</td>
<td>0.1409</td>
<td>0.3359</td>
<td>0.1356</td>
<td>0.2525</td>
<td>0.1576</td>
<td></td>
</tr>
<tr>
<td>0.2355</td>
<td>0.2205</td>
<td>0.2598</td>
<td>0.1519</td>
<td>0.2997</td>
<td>0.1890</td>
<td>0.2840</td>
<td>0.1356</td>
<td>0.3680</td>
<td>0.1786</td>
<td></td>
</tr>
<tr>
<td>0.1137</td>
<td>0.1534</td>
<td>0.2680</td>
<td>0.1882</td>
<td>0.1179</td>
<td>0.1534</td>
<td>0.4518</td>
<td>0.2525</td>
<td>0.3680</td>
<td>0.1902</td>
<td></td>
</tr>
<tr>
<td>0.2672</td>
<td>0.1503</td>
<td>0.2530</td>
<td>0.1868</td>
<td>0.1669</td>
<td>0.2260</td>
<td>0.2329</td>
<td>0.1576</td>
<td>0.1786</td>
<td>0.1902</td>
<td></td>
</tr>
</tbody>
</table>
### Table 8.6.6: Indirect interest correlations (all significant at $P \leq 0.05$)

<table>
<thead>
<tr>
<th>Int. 1.1</th>
<th>Int. 1.2</th>
<th>Int. 1.3</th>
<th>Int. 1.4</th>
<th>Int. 2.1</th>
<th>Int. 2.2</th>
<th>Int. 2.3</th>
<th>Int. 2.4</th>
<th>Int. 3.1</th>
<th>Int. 3.2</th>
<th>Int. 3.3</th>
<th>Int. 3.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5889</td>
<td>0.6003</td>
<td>0.5025</td>
<td>0.6436</td>
<td>0.6448</td>
<td>0.5743</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5025</td>
<td>0.6003</td>
<td>0.5881</td>
<td>0.6307</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5277</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 8.6.7: Direct interest correlations (all significant at $P \leq 0.05$)

<table>
<thead>
<tr>
<th>DI:n 1.2</th>
<th>DI:n 1.3</th>
<th>DI:n 1.4</th>
<th>DI:n 2.1</th>
<th>DI:n 2.2</th>
<th>DI:n 2.3</th>
<th>DI:n 2.4</th>
<th>DI:n 3.1</th>
<th>DI:n 3.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3111</td>
<td>0.3654</td>
<td>0.4648</td>
<td>0.3857</td>
<td>0.3563</td>
<td>0.3575</td>
<td>0.2290</td>
<td>0.3142</td>
<td></td>
</tr>
<tr>
<td>0.5059</td>
<td>0.3654</td>
<td>0.2077</td>
<td>0.4767</td>
<td>0.3886</td>
<td>0.5471</td>
<td>0.3589</td>
<td>0.3227</td>
<td>0.3631</td>
</tr>
<tr>
<td>0.2898</td>
<td>0.2077</td>
<td>0.5640</td>
<td>0.2792</td>
<td>0.4027</td>
<td>0.4844</td>
<td>0.3718</td>
<td>0.3105</td>
<td>0.5983</td>
</tr>
<tr>
<td>0.2582</td>
<td>0.2582</td>
<td>0.2792</td>
<td>0.4027</td>
<td>0.4844</td>
<td>0.3718</td>
<td>0.3105</td>
<td>0.5983</td>
<td>0.6573</td>
</tr>
<tr>
<td>0.2290</td>
<td>0.3427</td>
<td>0.5983</td>
<td>0.6573</td>
<td>0.5322</td>
<td>0.4819</td>
<td>0.4214</td>
<td>0.6262</td>
<td>0.4317</td>
</tr>
<tr>
<td>0.3142</td>
<td>0.3631</td>
<td>0.5020</td>
<td>0.2379</td>
<td>0.3105</td>
<td>0.5983</td>
<td>0.6573</td>
<td>0.5322</td>
<td>0.4819</td>
</tr>
</tbody>
</table>
Table 8.6.8: Pilot test correlations ($P \leq 0.05$)

<table>
<thead>
<tr>
<th></th>
<th>Pilot #1</th>
<th>Pilot #2</th>
<th>Pilot #3</th>
<th>Pilot #4</th>
<th>Pilot #5</th>
<th>Pilot #6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot #1</td>
<td>0.4646</td>
<td>0.3931</td>
<td>0.4008</td>
<td>0.0605</td>
<td>–0.0182</td>
<td></td>
</tr>
<tr>
<td>Pilot #2</td>
<td>0.4646</td>
<td>0.2720</td>
<td>0.2681</td>
<td>0.2401</td>
<td>–0.0762</td>
<td></td>
</tr>
<tr>
<td>Pilot #3</td>
<td>0.3931</td>
<td>0.2720</td>
<td>0.2681</td>
<td>0.2401</td>
<td>0.0723</td>
<td></td>
</tr>
<tr>
<td>Pilot #4</td>
<td>0.4008</td>
<td>0.2681</td>
<td>0.2681</td>
<td>0.2137</td>
<td>–0.0282</td>
<td>0.2593</td>
</tr>
<tr>
<td>Pilot #5</td>
<td>0.0605</td>
<td>0.2401</td>
<td>0.2401</td>
<td>0.2137</td>
<td>0.2593</td>
<td></td>
</tr>
<tr>
<td>Pilot #6</td>
<td>–0.0182</td>
<td>–0.0762</td>
<td>0.0723</td>
<td>–0.0282</td>
<td>0.2593</td>
<td></td>
</tr>
</tbody>
</table>

8.6.2 Basic Statistics

Table 8.6.9: Science self-efficacy statistics

<table>
<thead>
<tr>
<th>Science Self-Efficacy ($N=148$)</th>
<th>Mean</th>
<th>S.D.</th>
<th>Levels</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sc. Self-Efficacy Item #1</td>
<td>4.5541</td>
<td>0.6093</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Sc. Self-Efficacy Item #2</td>
<td>4.3311</td>
<td>0.8114</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Sc. Self-Efficacy Item #3</td>
<td>3.1757</td>
<td>1.1530</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Sc. Self-Efficacy Item #4</td>
<td>3.7770</td>
<td>0.8869</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Sc. Self-Efficacy Item #5</td>
<td>3.8108</td>
<td>1.1087</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Sc. Self-Efficacy Item #6</td>
<td>3.5878</td>
<td>1.1305</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Sc. Self-Efficacy Item #7</td>
<td>3.6622</td>
<td>0.9516</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Sc. Self-Efficacy Item #8</td>
<td>3.7568</td>
<td>0.8299</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

| Full Inventory                  | 30.6554  | 5.3245   | 33     | 12   | 40   |
Table 8.6.10: Indirect interest level statistics

<table>
<thead>
<tr>
<th>Indirect Interest Level</th>
<th>Subjects</th>
<th>Mean</th>
<th>S.D.</th>
<th>Levels</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest Item #1.1</td>
<td>146</td>
<td>2.3425</td>
<td>1.1107</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Interest Item #1.2</td>
<td>146</td>
<td>2.4726</td>
<td>1.0583</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Interest Item #1.3</td>
<td>146</td>
<td>2.3630</td>
<td>0.9605</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Interest Item #1.4</td>
<td>146</td>
<td>2.4932</td>
<td>1.0051</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Interest #1 Inventory</td>
<td>146</td>
<td>9.6712</td>
<td>3.4424</td>
<td>17</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>Interest Item #2.1</td>
<td>143</td>
<td>2.6503</td>
<td>0.9587</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Interest Item #2.2</td>
<td>143</td>
<td>2.8462</td>
<td>1.0022</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Interest Item #2.3</td>
<td>143</td>
<td>2.6434</td>
<td>0.9816</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Interest Item #2.4</td>
<td>143</td>
<td>2.7692</td>
<td>0.9907</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Interest #2 Inventory</td>
<td>143</td>
<td>10.9091</td>
<td>3.3143</td>
<td>17</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>Interest Item #3.1</td>
<td>148</td>
<td>2.6216</td>
<td>1.1029</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Interest Item #3.2</td>
<td>148</td>
<td>2.7635</td>
<td>1.1331</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Interest Item #3.3</td>
<td>148</td>
<td>2.6351</td>
<td>1.0636</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Interest Item #3.4</td>
<td>148</td>
<td>2.6351</td>
<td>1.0826</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Interest #3 Inventory</td>
<td>148</td>
<td>10.6554</td>
<td>3.8505</td>
<td>17</td>
<td>4</td>
<td>20</td>
</tr>
</tbody>
</table>
### Table 8.6.11: Direct interest level statistics

<table>
<thead>
<tr>
<th>Direct Interest (N=137)</th>
<th>Mean</th>
<th>S.D.</th>
<th>Range</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adj. D.I. Pos. #1.2</td>
<td>0.5036</td>
<td>0.7389</td>
<td>-1…1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Adj. D.I. Pos. #2.1</td>
<td>0.3066</td>
<td>0.7628</td>
<td>-1…1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Adj. D.I. Pos. #2.2</td>
<td>0.6788</td>
<td>0.5933</td>
<td>-1…1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Adj. D.I. Pos. #3.1</td>
<td>0.3066</td>
<td>0.8875</td>
<td>-1…1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Adj. D.I. Pos. #3.2</td>
<td>0.6569</td>
<td>0.6352</td>
<td>-1…1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Adj. D.I. Neg. #1.2</td>
<td>0.3431</td>
<td>0.8084</td>
<td>-1…1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Adj. D.I. Neg. #2.1</td>
<td>0.2920</td>
<td>0.8059</td>
<td>-1…1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Adj. D.I. Neg. #2.2</td>
<td>0.5182</td>
<td>0.7583</td>
<td>-1…1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Adj. D.I. Neg. #3.1</td>
<td>-0.1387</td>
<td>0.5027</td>
<td>-1…1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Adj. D.I. Neg. #3.2</td>
<td>0.5839</td>
<td>0.7826</td>
<td>-1…1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Adj. Dir. Int. Inventory</td>
<td>4.0511</td>
<td>4.9428</td>
<td>-10…10</td>
<td>-9</td>
<td>9</td>
</tr>
</tbody>
</table>

### Table 8.6.12: Difficulty assessment statistics

<table>
<thead>
<tr>
<th>Difficulty Assessment</th>
<th>Subjects</th>
<th>Mean</th>
<th>S.D.</th>
<th>Levels</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty Item #1</td>
<td>146</td>
<td>3.5548</td>
<td>1.0105</td>
<td>6</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Difficulty Item #2</td>
<td>143</td>
<td>2.7832</td>
<td>1.1015</td>
<td>6</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Difficulty Item #3</td>
<td>148</td>
<td>3.1351</td>
<td>1.0982</td>
<td>6</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>
Table 8.6.13: Session duration statistics

<table>
<thead>
<tr>
<th>Session Duration</th>
<th>Subjects</th>
<th>Mean</th>
<th>S.D.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session #1 Duration</td>
<td>146</td>
<td>14:47.75</td>
<td>5:14.43</td>
<td>5:30</td>
<td>36:18</td>
</tr>
<tr>
<td>Session #2 Duration</td>
<td>143</td>
<td>13:27.76</td>
<td>4:52.51</td>
<td>5:19</td>
<td>32:57</td>
</tr>
<tr>
<td>Session #3 Duration</td>
<td>148</td>
<td>13:40.66</td>
<td>5:45.86</td>
<td>3:35</td>
<td>47:14</td>
</tr>
</tbody>
</table>

8.6.1  Errata

Table 8.6.14: Aggregate score correlations

<table>
<thead>
<tr>
<th>Test Item Set</th>
<th>by</th>
<th>Test Item Set</th>
<th>r</th>
<th>Subjects</th>
<th>Lo. 95%</th>
<th>Up. 95%</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual #1 Set by Rev Set</td>
<td>0.3290</td>
<td>139</td>
<td>0.1719</td>
<td>0.4698</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factual #2 Set by Rev Set</td>
<td>0.4310</td>
<td>133</td>
<td>0.2814</td>
<td>0.5601</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factual #3 Set by Rev Set</td>
<td>0.4192</td>
<td>138</td>
<td>0.2710</td>
<td>0.5479</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #1 Set by Rev Set</td>
<td>0.3802</td>
<td>139</td>
<td>0.2282</td>
<td>0.5142</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #2 Set by Rev Set</td>
<td>0.4481</td>
<td>133</td>
<td>0.3008</td>
<td>0.5745</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #3 Set by Rev Set</td>
<td>0.5394</td>
<td>138</td>
<td>0.4092</td>
<td>0.6481</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factual #1 Set by Factual #1 Set</td>
<td>0.4582</td>
<td>133</td>
<td>0.3124</td>
<td>0.5830</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factual #3 Set by Factual #1 Set</td>
<td>0.4995</td>
<td>136</td>
<td>0.3616</td>
<td>0.6161</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factual #3 Set by Factual #2 Set</td>
<td>0.5095</td>
<td>135</td>
<td>0.3726</td>
<td>0.6246</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #1 Set by Factual #1 Set</td>
<td>0.4065</td>
<td>146</td>
<td>0.2613</td>
<td>0.5337</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #1 Set by Factual #2 Set</td>
<td>0.3732</td>
<td>133</td>
<td>0.2168</td>
<td>0.5110</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #1 Set by Conceptual #3 Set</td>
<td>0.3975</td>
<td>136</td>
<td>0.2456</td>
<td>0.5304</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #2 Set by Factual #1 Set</td>
<td>0.4642</td>
<td>133</td>
<td>0.3192</td>
<td>0.5880</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #2 Set by Conceptual #1 Set</td>
<td>0.5941</td>
<td>143</td>
<td>0.4764</td>
<td>0.6909</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #2 Set by Conceptual #2 Set</td>
<td>0.5595</td>
<td>135</td>
<td>0.4313</td>
<td>0.6655</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #2 Set by Conceptual #3 Set</td>
<td>0.4304</td>
<td>133</td>
<td>0.2807</td>
<td>0.5596</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #3 Set by Conceptual #1 Set</td>
<td>0.4720</td>
<td>136</td>
<td>0.3299</td>
<td>0.5932</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #3 Set by Conceptual #2 Set</td>
<td>0.4880</td>
<td>135</td>
<td>0.3477</td>
<td>0.6069</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #3 Set by Conceptual #3 Set</td>
<td>0.5381</td>
<td>148</td>
<td>0.4126</td>
<td>0.6436</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #3 Set by Conceptual #1 Set</td>
<td>0.4724</td>
<td>136</td>
<td>0.3303</td>
<td>0.5935</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conceptual #3 Set by Conceptual #2 Set</td>
<td>0.5273</td>
<td>135</td>
<td>0.3933</td>
<td>0.6393</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8.6.15 summarizes the reliability estimates for the batteries and inventories. In the case of tests, values below $\alpha=0.60$ are considered dubious, values at or above $\alpha=0.60$ are considered acceptable, and values at or above $\alpha=0.70$ are considered good. In the case of psychological inventories, the standards are raised by an increment of 0.1, so that $\alpha=0.80$ becomes the threshold for a “good” rating. Values that exceed the minimum for a “good” rating by 0.1 are considered excellent. The review test covered 7th grade content originally addressed by participating students across three semesters of instruction. Thus, its reliability estimate should be reconsidered, despite a “dubious” rating—the test measured more than one distinct construct.
### Table 8.6.15: Summary of reliability estimates

<table>
<thead>
<tr>
<th>Summary</th>
<th>$\alpha$</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review Battery</td>
<td>0.5284</td>
<td>Dubious</td>
</tr>
<tr>
<td>Treatment #1 Full Set</td>
<td>0.6234</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Treatment #2 Full Set</td>
<td>0.7821</td>
<td>Good</td>
</tr>
<tr>
<td>Treatment #3 Full Set</td>
<td>0.7156</td>
<td>Good</td>
</tr>
<tr>
<td>Factual #2 Set</td>
<td>0.6456</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Conceptual #2 Set</td>
<td>0.6757</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Conceptual #3 Set</td>
<td>0.6386</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Other Fact. &amp; Con. Sets</td>
<td>&lt;.6000</td>
<td>Dubious</td>
</tr>
<tr>
<td>Science Self-Efficacy</td>
<td>0.8507</td>
<td>Good</td>
</tr>
<tr>
<td>Interest #1 Inventory</td>
<td>0.8510</td>
<td>Good</td>
</tr>
<tr>
<td>Interest #2 Inventory</td>
<td>0.8638</td>
<td>Good</td>
</tr>
<tr>
<td>Interest #3 Inventory</td>
<td>0.9013</td>
<td>Excellent</td>
</tr>
<tr>
<td>Task Difficulty Scales</td>
<td>0.7339</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Session Duration as a Set</td>
<td>0.8908</td>
<td>Good</td>
</tr>
<tr>
<td>Adj. Direct Interest</td>
<td>0.8649</td>
<td>Good</td>
</tr>
</tbody>
</table>
8.7 Analysis of Variance (ANOVA) Tables

8.7.1 The Sex Variable (Where Significant)

Table 8.7.1: ANOVA table, review test by sex

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>1</td>
<td>34.1558</td>
<td>34.1558</td>
<td>11.1994</td>
<td>0.0010</td>
</tr>
<tr>
<td>Error</td>
<td>146</td>
<td>445.2699</td>
<td>3.0498</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>147</td>
<td>479.4257</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8.7.2: ANOVA table, combined treatment tests by sex

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>1</td>
<td>258.0284</td>
<td>258.028</td>
<td>6.3487</td>
<td>0.0130</td>
</tr>
<tr>
<td>Error</td>
<td>123</td>
<td>4999.0596</td>
<td>40.643</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>124</td>
<td>5257.0880</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8.7.3: ANOVA table, science self-efficacy by sex

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>1</td>
<td>198.5676</td>
<td>198.568</td>
<td>7.3046</td>
<td>0.0077</td>
</tr>
<tr>
<td>Error</td>
<td>146</td>
<td>3968.8581</td>
<td>27.184</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>147</td>
<td>4167.4257</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8.7.4: ANOVA table, combined duration by sex

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>1</td>
<td>3580685</td>
<td>3580685</td>
<td>4.7288</td>
<td>0.0316</td>
</tr>
<tr>
<td>Error</td>
<td>123</td>
<td>93135752</td>
<td>757201</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>124</td>
<td>96716438</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8.7.2 The Class Variable (Where Significant)

Table 8.7.5: ANOVA table, combined treatment tests by class

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>7</td>
<td>628.5290</td>
<td>89.7899</td>
<td>2.2697</td>
<td>0.0334</td>
</tr>
<tr>
<td>Error</td>
<td>117</td>
<td>4628.5590</td>
<td>39.5603</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>124</td>
<td>5257.0880</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8.7.6: ANOVA table, combined duration by class

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>7</td>
<td>13987723</td>
<td>1998246</td>
<td>2.8260</td>
<td>0.0093</td>
</tr>
<tr>
<td>Error</td>
<td>117</td>
<td>82728715</td>
<td>707083</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>124</td>
<td>96716438</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8.7.3 Group Equivalence

Table 8.7.7: ANOVA table, review test by group

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq. Group</td>
<td>5</td>
<td>8.8030</td>
<td>1.7606</td>
<td>0.5312</td>
<td>0.7523</td>
</tr>
<tr>
<td>Error</td>
<td>142</td>
<td>470.6226</td>
<td>3.3142</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>147</td>
<td>479.4257</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8.7.8: ANOVA table, combined treatment tests by group

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq. Group</td>
<td>5</td>
<td>221.6285</td>
<td>44.3257</td>
<td>1.0475</td>
<td>0.3932</td>
</tr>
<tr>
<td>Error</td>
<td>119</td>
<td>5035.4595</td>
<td>42.3148</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>124</td>
<td>5257.0880</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8.7.9: ANOVA table, science self-efficacy by group

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq. Group</td>
<td>5</td>
<td>197.8232</td>
<td>39.5646</td>
<td>1.4153</td>
<td>0.2223</td>
</tr>
<tr>
<td>Error</td>
<td>142</td>
<td>3969.6025</td>
<td>27.9549</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>147</td>
<td>4167.4257</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8.7.1 Independent Variable Inquiry

Table 8.7.10: ANOVA table, treatment #1 test by form

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forms #1</td>
<td>2</td>
<td>0.8364</td>
<td>0.4182</td>
<td>0.0692</td>
<td>0.9332</td>
</tr>
<tr>
<td>Error</td>
<td>143</td>
<td>864.1773</td>
<td>6.0432</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>145</td>
<td>865.0137</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8.7.11: ANOVA table, treatment #2 test by form

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forms #2</td>
<td>2</td>
<td>56.6214</td>
<td>28.3107</td>
<td>4.0152</td>
<td>0.0202</td>
</tr>
<tr>
<td>Error</td>
<td>140</td>
<td>987.1129</td>
<td>7.0508</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>142</td>
<td>1043.7343</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

253
Table 8.7.12: ANOVA table, factual #2 set by form

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forms #2(f)</td>
<td>2</td>
<td>11.7759</td>
<td>5.8879</td>
<td>3.6368</td>
<td>0.0289</td>
</tr>
<tr>
<td>Error</td>
<td>140</td>
<td>226.6577</td>
<td>1.6190</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>142</td>
<td>238.4336</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8.7.13: ANOVA table, conceptual #2 set by form

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forms #2(c)</td>
<td>2</td>
<td>16.9364</td>
<td>8.4682</td>
<td>2.8951</td>
<td>0.0586</td>
</tr>
<tr>
<td>Error</td>
<td>140</td>
<td>409.4972</td>
<td>2.9250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>142</td>
<td>426.4336</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8.7.14: ANOVA table, treatment #3 test by form

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forms #3</td>
<td>2</td>
<td>9.1250</td>
<td>4.5625</td>
<td>0.7094</td>
<td>0.4937</td>
</tr>
<tr>
<td>Error</td>
<td>145</td>
<td>932.6250</td>
<td>6.4319</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>147</td>
<td>941.7500</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 8.7.15: ANOVA table, difficulty #1 rating by form

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty #1</td>
<td>2</td>
<td>0.8859</td>
<td>0.4430</td>
<td>0.4304</td>
<td>0.6511</td>
</tr>
<tr>
<td>Error</td>
<td>143</td>
<td>147.1757</td>
<td>1.0292</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>145</td>
<td>148.0616</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 8.7.16: ANOVA table, difficulty #2 rating by form

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty #2</td>
<td>2</td>
<td>3.2052</td>
<td>1.6026</td>
<td>1.3270</td>
<td>0.2686</td>
</tr>
<tr>
<td>Error</td>
<td>140</td>
<td>169.0746</td>
<td>1.2077</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>142</td>
<td>172.2797</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 8.7.17: ANOVA table, difficulty #3 rating by form

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty #3</td>
<td>2</td>
<td>3.4176</td>
<td>1.7088</td>
<td>1.4250</td>
<td>0.2439</td>
</tr>
<tr>
<td>Error</td>
<td>145</td>
<td>173.8797</td>
<td>1.1992</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>147</td>
<td>177.2973</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8.7.18: ANOVA table, indirect interest #1 rating by form

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest #1</td>
<td>2</td>
<td>46.4288</td>
<td>23.2144</td>
<td>1.9857</td>
<td>0.1411</td>
</tr>
<tr>
<td>Error</td>
<td>143</td>
<td>1671.7904</td>
<td>11.6908</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>145</td>
<td>1718.2192</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8.7.19: ANOVA table, indirect interest #2 rating by form

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest #2</td>
<td>2</td>
<td>3.3387</td>
<td>1.6694</td>
<td>0.1502</td>
<td>0.8607</td>
</tr>
<tr>
<td>Error</td>
<td>140</td>
<td>1556.4794</td>
<td>11.1177</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>142</td>
<td>1559.8182</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8.7.20: ANOVA table, indirect interest #3 rating by form

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest #3</td>
<td>2</td>
<td>91.5185</td>
<td>45.7593</td>
<td>3.1779</td>
<td>0.0446</td>
</tr>
<tr>
<td>Error</td>
<td>145</td>
<td>2087.9071</td>
<td>14.3994</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>147</td>
<td>2179.4257</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### 8.8.1 Group Equivalence

#### Table 8.8.1: T-test, combined treatment tests by class

<table>
<thead>
<tr>
<th>Level</th>
<th>Level</th>
<th>Difference</th>
<th>St. Err. Dif.</th>
<th>Lower CL</th>
<th>Upper CL</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher A, Class 3 – Teacher B, Class 3</td>
<td>7.1032</td>
<td>2.2413</td>
<td>2.6643</td>
<td>11.5420</td>
<td>0.0020</td>
<td></td>
</tr>
<tr>
<td>Teacher B, Class 1 – Teacher B, Class 3</td>
<td>5.9018</td>
<td>2.3018</td>
<td>1.3432</td>
<td>9.4604</td>
<td>0.0116</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1 – Teacher B, Class 3</td>
<td>5.4018</td>
<td>2.3018</td>
<td>0.8432</td>
<td>9.9604</td>
<td>0.0206</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 3 – Teacher A, Class 2</td>
<td>5.1111</td>
<td>2.0966</td>
<td>0.9590</td>
<td>9.2633</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 3 – Teacher B, Class 2</td>
<td>4.9444</td>
<td>2.0966</td>
<td>0.9590</td>
<td>9.2633</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1 – Teacher A, Class 2</td>
<td>5.4018</td>
<td>2.3018</td>
<td>0.8432</td>
<td>9.9604</td>
<td>0.0206</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 3 – Teacher B, Class 3</td>
<td>5.1111</td>
<td>2.0966</td>
<td>0.9590</td>
<td>9.2633</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1 – Teacher A, Class 2</td>
<td>5.4018</td>
<td>2.3018</td>
<td>0.8432</td>
<td>9.9604</td>
<td>0.0206</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 3 – Teacher B, Class 3</td>
<td>5.1111</td>
<td>2.0966</td>
<td>0.9590</td>
<td>9.2633</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1 – Teacher A, Class 2</td>
<td>5.4018</td>
<td>2.3018</td>
<td>0.8432</td>
<td>9.9604</td>
<td>0.0206</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 3 – Teacher B, Class 3</td>
<td>5.1111</td>
<td>2.0966</td>
<td>0.9590</td>
<td>9.2633</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1 – Teacher A, Class 2</td>
<td>5.4018</td>
<td>2.3018</td>
<td>0.8432</td>
<td>9.9604</td>
<td>0.0206</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 3 – Teacher B, Class 3</td>
<td>5.1111</td>
<td>2.0966</td>
<td>0.9590</td>
<td>9.2633</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1 – Teacher A, Class 2</td>
<td>5.4018</td>
<td>2.3018</td>
<td>0.8432</td>
<td>9.9604</td>
<td>0.0206</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 3 – Teacher B, Class 3</td>
<td>5.1111</td>
<td>2.0966</td>
<td>0.9590</td>
<td>9.2633</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1 – Teacher A, Class 2</td>
<td>5.4018</td>
<td>2.3018</td>
<td>0.8432</td>
<td>9.9604</td>
<td>0.0206</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 3 – Teacher B, Class 3</td>
<td>5.1111</td>
<td>2.0966</td>
<td>0.9590</td>
<td>9.2633</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1 – Teacher A, Class 2</td>
<td>5.4018</td>
<td>2.3018</td>
<td>0.8432</td>
<td>9.9604</td>
<td>0.0206</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 3 – Teacher B, Class 3</td>
<td>5.1111</td>
<td>2.0966</td>
<td>0.9590</td>
<td>9.2633</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1 – Teacher A, Class 2</td>
<td>5.4018</td>
<td>2.3018</td>
<td>0.8432</td>
<td>9.9604</td>
<td>0.0206</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 3 – Teacher B, Class 3</td>
<td>5.1111</td>
<td>2.0966</td>
<td>0.9590</td>
<td>9.2633</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1 – Teacher A, Class 2</td>
<td>5.4018</td>
<td>2.3018</td>
<td>0.8432</td>
<td>9.9604</td>
<td>0.0206</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 3 – Teacher B, Class 3</td>
<td>5.1111</td>
<td>2.0966</td>
<td>0.9590</td>
<td>9.2633</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1 – Teacher A, Class 2</td>
<td>5.4018</td>
<td>2.3018</td>
<td>0.8432</td>
<td>9.9604</td>
<td>0.0206</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 3 – Teacher B, Class 3</td>
<td>5.1111</td>
<td>2.0966</td>
<td>0.9590</td>
<td>9.2633</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1 – Teacher A, Class 2</td>
<td>5.4018</td>
<td>2.3018</td>
<td>0.8432</td>
<td>9.9604</td>
<td>0.0206</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 3 – Teacher B, Class 3</td>
<td>5.1111</td>
<td>2.0966</td>
<td>0.9590</td>
<td>9.2633</td>
<td>0.0163</td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1 – Teacher A, Class 2</td>
<td>5.4018</td>
<td>2.3018</td>
<td>0.8432</td>
<td>9.9604</td>
<td>0.0206</td>
<td></td>
</tr>
</tbody>
</table>
### Table 8.8.2: T-test, session duration by class

<table>
<thead>
<tr>
<th>Level</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher A, Class 4</td>
<td>Teacher A, Class 3</td>
<td>1114.91</td>
<td>299.65</td>
<td>521.48</td>
<td>1708.35</td>
<td>0.0003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 4</td>
<td>Teacher A, Class 1</td>
<td>1061.67</td>
<td>307.73</td>
<td>452.22</td>
<td>1671.12</td>
<td>0.0008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 4</td>
<td>Teacher B, Class 2</td>
<td>1032.08</td>
<td>299.65</td>
<td>438.64</td>
<td>1625.12</td>
<td>0.0008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 4</td>
<td>Teacher B, Class 3</td>
<td>707.93</td>
<td>317.82</td>
<td>78.50</td>
<td>1337.36</td>
<td>0.0278</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 4</td>
<td>Teacher B, Class 4</td>
<td>616.68</td>
<td>338.80</td>
<td>–54.30</td>
<td>1287.65</td>
<td>0.0713</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 4</td>
<td>Teacher A, Class 1</td>
<td>842.86</td>
<td>299.65</td>
<td>249.42</td>
<td>1436.29</td>
<td>0.0058</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 4</td>
<td>Teacher B, Class 1</td>
<td>585.73</td>
<td>307.73</td>
<td>–23.71</td>
<td>1195.18</td>
<td>0.0594</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher B, Class 2</td>
<td>Teacher A, Class 3</td>
<td>498.24</td>
<td>321.81</td>
<td>–125.84</td>
<td>1018.54</td>
<td>0.1251</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher B, Class 4</td>
<td>Teacher B, Class 1</td>
<td>444.99</td>
<td>329.35</td>
<td>–207.27</td>
<td>1097.26</td>
<td>0.1793</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher B, Class 3</td>
<td>Teacher A, Class 1</td>
<td>446.35</td>
<td>288.92</td>
<td>–157.43</td>
<td>1018.54</td>
<td>0.1251</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1</td>
<td>Teacher A, Class 3</td>
<td>272.06</td>
<td>280.29</td>
<td>–283.05</td>
<td>827.16</td>
<td>0.3337</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher B, Class 1</td>
<td>Teacher A, Class 2</td>
<td>257.12</td>
<td>288.92</td>
<td>–315.07</td>
<td>829.32</td>
<td>0.3753</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher B, Class 2</td>
<td>Teacher A, Class 3</td>
<td>218.81</td>
<td>288.92</td>
<td>–353.38</td>
<td>791.00</td>
<td>0.4504</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher B, Class 3</td>
<td>Teacher B, Class 1</td>
<td>189.22</td>
<td>280.29</td>
<td>–365.89</td>
<td>744.33</td>
<td>0.5010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1</td>
<td>Teacher A, Class 2</td>
<td>134.93</td>
<td>299.65</td>
<td>–458.51</td>
<td>728.36</td>
<td>0.6533</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher B, Class 1</td>
<td>Teacher A, Class 3</td>
<td>122.20</td>
<td>307.73</td>
<td>–487.27</td>
<td>731.64</td>
<td>0.6920</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher B, Class 4</td>
<td>Teacher B, Class 2</td>
<td>82.83</td>
<td>280.29</td>
<td>–365.89</td>
<td>637.94</td>
<td>0.7681</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher A, Class 1</td>
<td>Teacher A, Class 3</td>
<td>53.24</td>
<td>288.92</td>
<td>–315.07</td>
<td>625.44</td>
<td>0.8541</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher B, Class 2</td>
<td>Teacher A, Class 1</td>
<td>29.59</td>
<td>288.92</td>
<td>–542.60</td>
<td>601.78</td>
<td>0.9186</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

258
8.8.2 Independent Variable Inquiry

Table 8.8.3: T-test, treatment #2 test by form

<table>
<thead>
<tr>
<th>Level</th>
<th>–</th>
<th>Level</th>
<th>Difference</th>
<th>St. Err. Dif.</th>
<th>Lower CL</th>
<th>Upper CL</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form #2: FI Version</td>
<td>–</td>
<td>Form #2: PP Version</td>
<td>1.4509</td>
<td>0.5328</td>
<td>0.3976</td>
<td>2.5042</td>
<td>0.0073</td>
</tr>
<tr>
<td>Form #2: PS Version</td>
<td>–</td>
<td>Form #2: PP Version</td>
<td>1.0469</td>
<td>0.5427</td>
<td>–0.0260</td>
<td>2.1199</td>
<td>0.0557</td>
</tr>
<tr>
<td>Form #2: FI Version</td>
<td>–</td>
<td>Form #2: PS Version</td>
<td>0.4039</td>
<td>0.5633</td>
<td>–0.7096</td>
<td>1.5175</td>
<td>0.4745</td>
</tr>
</tbody>
</table>

Table 8.8.4: T-test, factual #2 set by form

<table>
<thead>
<tr>
<th>Level</th>
<th>–</th>
<th>Level</th>
<th>Difference</th>
<th>St. Err. Dif.</th>
<th>Lower CL</th>
<th>Upper CL</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form #2: FI Version</td>
<td>–</td>
<td>Form #2: PP Version</td>
<td>0.6433</td>
<td>0.2553</td>
<td>0.1386</td>
<td>1.1481</td>
<td>0.0129</td>
</tr>
<tr>
<td>Form #2: PS Version</td>
<td>–</td>
<td>Form #2: PP Version</td>
<td>0.5189</td>
<td>0.2601</td>
<td>0.0048</td>
<td>1.0331</td>
<td>0.0479</td>
</tr>
<tr>
<td>Form #2: FI Version</td>
<td>–</td>
<td>Form #2: PS Version</td>
<td>0.1244</td>
<td>0.2699</td>
<td>–0.4092</td>
<td>0.6580</td>
<td>0.6457</td>
</tr>
</tbody>
</table>

Table 8.8.5: T-test, conceptual #2 set by form

<table>
<thead>
<tr>
<th>Level</th>
<th>–</th>
<th>Level</th>
<th>Difference</th>
<th>St. Err. Dif.</th>
<th>Lower CL</th>
<th>Upper CL</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form #2: FI Version</td>
<td>–</td>
<td>Form #2: PP Version</td>
<td>0.8076</td>
<td>0.3432</td>
<td>0.1291</td>
<td>1.4860</td>
<td>0.0200</td>
</tr>
<tr>
<td>Form #2: PS Version</td>
<td>–</td>
<td>Form #2: PP Version</td>
<td>0.5280</td>
<td>0.3496</td>
<td>–0.1631</td>
<td>1.2191</td>
<td>0.1332</td>
</tr>
<tr>
<td>Form #2: FI Version</td>
<td>–</td>
<td>Form #2: PS Version</td>
<td>0.2796</td>
<td>0.3628</td>
<td>–0.4377</td>
<td>0.9968</td>
<td>0.4422</td>
</tr>
</tbody>
</table>

259
Table 8.8.6: T-test, indirect interest #3 by form

<table>
<thead>
<tr>
<th>Level</th>
<th>Level</th>
<th>Difference</th>
<th>St. Err. Dif.</th>
<th>Lower CL</th>
<th>Upper CL</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form #3: PS Version</td>
<td>Form #3: PP Version</td>
<td>1.9151</td>
<td>0.7739</td>
<td>0.3855</td>
<td>3.4447</td>
<td>0.0145</td>
</tr>
<tr>
<td>Form #3: FI Version</td>
<td>Form #3: PP Version</td>
<td>1.3627</td>
<td>0.7808</td>
<td>-0.1804</td>
<td>2.9059</td>
<td>0.0830</td>
</tr>
<tr>
<td>Form #3: PS Version</td>
<td>Form #3: FI Version</td>
<td>0.5523</td>
<td>0.7443</td>
<td>-0.9188</td>
<td>2.0235</td>
<td>0.4592</td>
</tr>
</tbody>
</table>

8.9  

Linear Regression

8.9.1  

Assessing Task Difficulty According to Performance

Figure 8.9.1: Difficulty performance-basis
### Table 8.9.1: Difficulty performance-basis summary of fit

<table>
<thead>
<tr>
<th>Summary of Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
</tr>
<tr>
<td>RSquare Adj</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>Mean of Response</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
</tr>
</tbody>
</table>

### Table 8.9.2: Difficulty performance-basis ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>91.7327</td>
<td>91.7327</td>
<td>17.6446</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>123</td>
<td>639.4673</td>
<td>5.1989</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>124</td>
<td>731.2000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 8.9.3: Difficulty performance-basis parameter estimate

| Term        | Estimate     | Std Error | t Ratio | Prob > |t| |
|-------------|--------------|-----------|---------|---------|---|
| Intercept   | 12.164818    | 0.707290  | 17.20   | <.0001  |
| Test Scores | -0.132096    | 0.031447  | -4.20   | <.0001  |
Task Difficulty Explanations

Treatment #1 session, “very easy”:

1) It never got hard all you had to do is read then pick question
2) It allowed you to look back at it and figure out the answer like taking a test while using your notes
3) Because almost all of the answers were in the reading.
4) It was very straightforward with the information it was telling me and the answers were very simple ones where you got most of them from the text. The pictures also helped to add another way to get the information.

Treatment #2 session, “very easy”:

1) Because if you read it first then you would have all the information. Everything you needed to answer the questions was easily found and read in the text. It gave you enough information so that you wouldn't be confused about it.
2) No reading they need to put more reading in it
3) It was easy because it told us that lakes are bigger than a pond
4) I didn't find this hard at all
5) It was easy to understand and I was more positive I got the question right.
6) BECAUSE THERE WERE NOT AS MUCH READING AND MORE EX.
7) Once I read the handout thoroughly all the question were very easy. I was like "dub" way too easy.
8) because the answers was right there when or if you read good.
9) because the handout was clear
10) I think that it was very easy because most of the answers were common sense things or could be found in the text.
11) The handout easily broke down the concepts of lakes and ponds so when answering the questions it was easy to remember the information that was on the handout.
12) Because the answers were almost all in the handout
13) The questions were easy and the information was organized well. The pictures also helped to get a better feel about the information.
14) This was very easy because luckily, this NCSU Student gave me very good information, in the handout, to help me answer these questions. This informative handout, really helped me understand much better than I probably would have from a . . . textbook. GREAT Job!.
15) I found this easy because of the fact that you had all the answers in front of you and they are easy to find. Also, you had just read what you were answering questions about, so everything was fresh in your mind.
16) I deal with these entities on a normal bases as a fisher, scout and regular person.

Treatment #3 session, “very easy”:

1) it was easy to read
2) Because the questions were lame and didn't want me to try.
3) because the information was clear and it was spread out. The handout had bolded vocabulary words and pictures to keep you interested.
4) it was very easy
5) Yes it was very easy with the handout.
6) it was very easy to read and the questions were very easy aswell
7) I knew so stuff about it so it helped me with it plus the knew information helped me.
8) By reading the text you learn all of the answers.
9) Well the answers are simple and really you don't need the articles to help you because most people know about it. Meen it's just an everyday problem society has that everybody knows about.
10) I was very straightforward with its questions and answers and was organized well. The pictures also helped too.
11) Because I already knew most of the material.

Treatment #1 session, “easy”:

1) I found it easy, because the text was clear and so were the questions.
2) I founded it easy, because of I had help from the handout.
3) I found this easy because reading a 2 pages then answering about 15 questions shouldn't be hard. I also find this to be easy because I have heard of the same thing in a different science class when I went to my cousin's class to volunteer.
4) Because all of the answers are given to you in the partcet.
5) I found working with the handout easy because it was something we had already covered.
6) Working with the handouts were easy to me because all you had to do was read really good and then you'll get the answers right there in the text.
7) Because I read carefully through the handout.
8) Because the answers was in the book.
9) It gave me all the information I needed to find the answers to the questions.

Treatment #2 session, “easy”:

1) I was able to look back in the handout and find the answer to my question. That's what made it easy.
2) It was easy because the reading was easy to read and showed lots of examples.
3) Because the questions were simple to answer.
4) I felt that I could always look back and check my work to see if it was correct. If I didn't understand the question, I could always go back to the reading for a review.
5) I thought it was a bit interesting and fun and all the answers were in the book if you actually read.
6) I got the questions and the topic was not boring like the last one was.
7) Because if I didn't get a question I could read the article again to see which answer best suit it.
8) I think that this handout was easier than the last because the information is clear and I can clearly see the diagrams and pictures to match.
9) Because I was taking my time to read the hand out. But the questions was kind id easy for me.
10) Because they were lots of Pictures and captions.
11) When I was working on this exercise I found it much easier because I could look back, I could keep the handout on the computer and read or what ever beside me and on a computer I would have to keep on switching back and forth with the website and the handout.
12) It helps to find the answer more easier than trying to remember.
13) I found working with the handout easy because of how i had something that i could referece to. It was something that i could look at if i forgot the answer to something.
14) This handout was very easy because i had the artical plased in my hand, from thre i can read, answer and learn about the Lakes, and Ponds.
15) it was easy to find the information in the text and it was easy to understand.
16) To me it was easy and mostly I guess because of the pictures.
17) All of the answers to the questions were right in the handout.
18) I found that the answers were all in the handout.
19) The information that was displayed in the pamphlet was very clear and solid. That made it easy to answer the questions and understand the information.
20) It was easy working with the handout because all you have to do is read the handout over to find the answer.
21) It was really easy the pictures helped me understand what was going on it was a little boring at a time since most people don't really care about fish in the summer or the winter but the hand-out explained how things happen in a lake or pond it also helped me determine which is which so everything was every easily displayed it was a little better to use than just pictures and a bunch of arrows pointing towards things or just a bunch of words
22) I think that I can get more questions right and understand easier if I read something textbook style. I struggle with diagrams because they are hard for me to take in.
23) I found this to be easy because I just had to read 2 pages then answer a few questions. One thing about the questions is that one was asking about the turnover and from what I read in the paper none of the answers were right. But other then that it was an easy assignment.
24) The handout was clear and didn't leave anything out. Also, there were lots of illustrations so I didn't have any confusion and didn't have to think really hard and visualize the words myself. It was also fairly easy because the information was broken up into sections instead of one huge blob of words on my paper.
25) Because there were pictures to describe the writing they gave us.
26) I found it easy because all of the information was there and you only had to remember what it was instead of guess at what the correct answer was
27) It was an easy subject that I learned in fifth grade.
28) Because I know almost all of it all ready
29) The handout told me all the answers so that made answering the questions easy. I also found that the many diagrams/pictures helped me understand about lakes and ponds.
30) Because the question made me feel confident about my answers and I did not have any problems on answering them
31) I could read then question, then look right back at the packet and quickly find the answer to the problem which made it really easy.
32) The information on the survey/test is found somewhere in the handout
33) I thought it was easy because the answers to all of the questions were in the handout, and if you read the handout once really well, or twice through, then you would have no problem answering the questions. But some of the questions you did have to think about, as they were not directly in the handout.
34) It presented most of the information in a way that was fairly easy to understand.
35) I could find all of the information that I needed to answer the questions.
36) It is really just talking about lakes and ponds and, all I had to do was re-read over them until I found the answer or just read the whole packet over and over again until I understand or just found the answer. That's why I say it's easy.
37) Because all the answers you need is in the handout
38) I found this easy because I had the packet with all of the information that I would need.
39) Because all the questions they were asking you, the answers were right in the text and you could go back and find them easily
40) I almost understood everything in the handout and there wasn't anything real hard. The handout also explained it really good and made it simple.
41) Because it was easy to learn, after I read it once I had all the information in my head, and I already know some about lakes and ponds.
42) I have heard about lakes and ponds before and compared to the previous surveys we have taken, this was the easiest one.
43) Because this was much easier to understand since you can tell the differences easily and you know what to expect from those types of bodies of water.
44) The answers were easily presented and the questions were worded so that they were easy to understand.
45) There were pictures and diagrams for a better understanding.
46) It was very straightforward, and sorta easy to understand.
47) Because it gave me the most important answers and it gave me the answers to the questions much easier and there wasn't a lot of words

Treatment #3 session, “easy”:

264
1) because fossil fuel is fun  
2) Yes because it was fun and had more major facts than a whole article which would have many unneeded facts. But it also had many pictures to describe the way the things were happening.  
3) I understood the reading  
4) it showed pictures that explained how things work  
5) It was easy because it was a source of reference. It was something that i could look back on if i forgot something. It also helped me get a better understanding of the information.  
6) I found it easy because all of the answers were right in the handout and were easy to access.  
7) This handout was easy because all the information was clearly given to you and it allowed me to easily access the information when I wanted to refer back to it.  
8) well since I read it all the answer were just there why wouldn't it be easy? If you actually read the stuff and read the questions and answers then it's not hard at all.  
9) I found working with this topic to be easy because I have already studied this topic and all the information was in the book of course.  
10) All the information that we needed was right there and I do remember a lot about this from last year's science class.  
11) They us the pepper and we read then we answer the pilom  
12) because of course we were able to use it.  
13) Because it was clear and easy to understand  
14) because i knew most of the stuff that was in there it was not something new to me  
15) I think that some of the questions were not very good and the answers did not make much sense sometimes  
16) it was easy because if you read the article carefully all the answers except, what is peat! where all there  
17) If you read the passage then take the questions then it'll be much more easier than just looking for the information.  
18) Lots of diagrams that helped me understand all about the fossil fuels.  
19) There was clues that i could find  
20) I though that the information was easy to read and that it gave all of the needed information  
21) I had all the information in the packet  
22) I found this to be easy because the pictures gave me more of a better visual of what the story is talking about.  
23) Using this handout, made things, much easier. A lot of this information, I didn't know before reading this handout. So Good Job NC State University Student :) !.  
24) Because it has examples and showed you the process in which coal is made and how they are used and also the affects that it may have.  
25) It helps to understand the disadvantages and advantages of using fossil fuel over solar energy.  
26) It was easy because I already knew a good amount of information on this topic.  

Treatment #1 session, “somewhat easy”:

1) Because, most of the question on the vocab didn't have the vocab in the book. I did enjoy learning about How fresh water flow and freezes on earth! :)  
2) I found it somewhat easy because i was able to go back in look in the passage to answer and the questions.  
3) They got a handout so they can get the answers.  
4) I felt that I could always look back on the handout to check and see if my thinking was right or not.  
5) Because I like visual learning alittle more than reading and most of the time I think I learn more quickly.  
6) Because you can read the paper, and if you confused you can reread it.  
7) It was actually fun, because you had to read the handout and learn about the information, and then go and answer the questions. This really worked  
8) It was somewhat easy because I did not have to look back through the reading to remember the answers.  
9) Because first you read the handout. For me was somewhat easy.  
10) Because The handout had pictures to show you examples of the different things.  
11) Because most of the content I knew pretty well and was not to hard to answer from the handout given while some questions were unclear to the information from the handout.  
12) I think it is somewhat easy because I can see with my eyes what is on the paper and having to go back and forth on the computer can get my micks up with other things.  
13) I found working with the handout to be somewhat easy because it gave me a point of reference while I was doing the survey. It also helped me because I used the picture to make guesses about the way the
water flowed and how the different basins where shaped.

14) I found this to be easy because I had a chance to read it at my article.

15) It tells you what you need to know.

16) There were some parts were the answers didn't make any sense like the gravity and divides relationship questions.

17) Because of the pictures.

18) Part of this handout was common knowledge, but some of the questions included things that actually made me think.

19) Because I did not understand all of the questions.

20) Because all you had to do was read the information and you had to use common sense to.

21) Working with a handout is easier because sometimes the handout has the answers & you can just look for it.

22) Because if previous grades, we usually work with handouts on science related material, and I am a good reader so in absorbing the information and hopefully I will know the answers to the questions.

23) I think it was kind of easy because what I did was look at pictures and read the captions I was able to visualize what was actually happening in the picture but what kind of threw me off were the questions that were kind of hard and you had to get your thoughts organized to answer them.

24) The handout gave me a visual that I could always go back to and I could see what the questions were asking me instead of having to search through words like in an article or textbook.

25) Because the information was in the passage and it was easy to find the information.

26) I thought that it wasn't super easy but the questions weren't hard to understand and to me most of the answers were clear.

27) It was easy to work with the handout because it had a visual, and it described things easily.

28) The reason I found this handout somewhat easy was because there were pictures to explain what was written either above or below it. This helped me alot with understanding key vocabulary throughout the hole handout.

29) There were some questions I did not know the answers too but the article was very easy to learn from.

30) Because, when the teacher told us to read through it before starting on the survey, I had got the understanding of the article and plus I know some what stuff about fresh water and how it forms and where it flows. So when I started on the survey I knew stuff about it and if I didn't really understand I looked back into the article for more help on the questions.

31) Because it about the water and sea and we had the a handout to read with it read more.
I found it easy because we had just read about the things we were answering questions about, so it was easier to remember.

It guided me through it with the pictures.

Because the only thing you have to do is look at the question then find it in the package.

Well cause all of the questions where in the handout and some of the answers where in bold and that was a big help.

Because it didn’t exactly give you too much information to answer some of the questions that you had to answer.

I found working with the handout somewhat easy because it mostly helped me out with answering most of the questions.

Yes.

Because, all of the answers are written there clearly so you don’t have to search for it.

Because in this passage I read, most of my answers were in the passage.

Treatment #2 session, “somewhat easy”:

1) there wasn’t a lot of information to look through so the information was just there
2) because it is like it is giving you all the answer I think
3) I’m able to re-read the passage to look for the answers.
4) I found it to be somewhat easy because it had more words and pictures to go along with it. I got more of the information with the words, rather than just the picture.
5) I learn about kinds of stuff.
6) Because some of the answers were just really in the handout all you had to do was think!
7) cuz
8) because it was easy but not too easy
9) I think all should have been easy to ask and the questions were based on the content
10) Because it was something that I had to learn in one day and only understand a little bit of the information.
11) because some of the questions were easy.
12) I found it somewhat easy because I was able to answer every single question easily except one about the sameness of a lake during the fall and summer seasons
13) Who looked interesting so I put more interest into it. It looked sort of easier then the other we have done.
14) It help me look back in the handout so I can have a chance to get the answers.
15) because of the pics
16) The handout was easy in terms, but it had the information scattered. The problem was trying to find it all.
17) Because some questions I did not understand
18) The questions for me seemed like some of it didn’t match the reading.
19) I think it was somewhat easy because the answers to the questions were right there in the picture and in the little sit words and I just had to look for the right answers and they were all there so I thought it was somewhat easy.
20) I think it was somewhat easy because it had pictures to help but a little bit of a lot of description.
21) Because only some problems were hard to me.
22) sorry, I can’t explain that.
23) I found it somewhat easy because the story gave an example in pictures, and in writing.
24) Most of the answers were very easy to find in the text and the drawing helped a lot
25) I really wouldn’t say it was that easy because, it was some what difficult with some parts at the same time. But most questions I answered it was some what easy! And plus the questions was mostly not the same for the handout.
26) All of the information was in the packet I had just read.
27) IT WAS ALL IN THE BOOK IF YOU READ IT
28) because you will know it easily if it is a lake or pond and I kind of know about it, just little bit
29) Because the question was kind of confusing. They were harder to understand than the question we do in our regular science class. But, the reading wasn’t that confusing. The reading was kind of easy to understand.
30) It was somewhat easy because it had multiple diagrams but the wording of the text was a little difficult to understand.
31) Because I could find the answers easily and I understood everything!
32) BECAUSE SOME OF THESE THINGS WERE SIMPLE AND I REREAD THE HANDOUT
33) The handout was easy because when I answered the questions, almost all of them were either in the reading, or implied.
34) I found it easy because, it has both pictures and words to help understand the concept.
It was like in the middle of being hard or easy. Because all you had to do is read the article and answer the questions. There were some tricky answers and there seemed to be 2 answers that I was sure was right. Well I liked the selection but I don't have much background information on this topic. Also I didn't realize we were going to take another test. Although I didn't know much on the subject the test was easy and understandable. Because I could remember the stuff. Because I knew where the information would be found. Well it was difficult because really there aren't any words, there just little sections of stuff and that confuses me. It was easy because some of the stuff I already knew. Well in the handout all the answers were in the reading. It was organized and if I needed to go back to find an answer it was already there. It was also somewhat different because there was not a title at the top of the paragraphs so you didn't really know which one was which. I did not really get the hand out only some of the handout. I don't really why, but I guess this handout was a bit easy than the other one. Because it has a lot of information and it guides you through it. The only the only thing you have to do is read and answer. All of the answers where in the text except for a few which you had to infer a little bit. The words seemed to be stated clearly in an easy to understand way and the pictures helped give you a visual of what was happening and why it's happening. Because you let us keep the handout. Because it wasn't hard like the last one and I could find clues easier than the last time and it was easier to answer the questions this time.

Treatment #3 session, “somewhat easy”:

1) Because it did not seem to hard.
2) because it told me all the answers and didn't make me think except for one of the questions about peat.
3) There were some questions that were a little difficult but other than that the rest was simple.
4) Because there was a lot of information in the packet.
5) Had enough information to answer questions.
6) It wasn't too hard to answer.
7) I found working with the handout to be somewhat easy because I felt that it was kind of like a guide, so I could look back and review my work instead of just reading a selection and going right to the questions without looking back.
8) Cause most of all of the answers were in the passage.
9) I did not have to read a lot but that made it harder to find the answers.
10) There were some questions I didn't get and they confused me.
11) I think all the questions should have been based on the topic and content in the selection. I also recommend that you make the question easier so you can spot the answer in the text easier.
12) I really liked the illustrations, which made the reading easier to understand.
13) Because the questions were kinda easy.
14) Because pretty much every single question except one I was sure I got right.
15) I can work and I can look down when I need to see something.
16) YOU READ THE HANDOUT AND IT GIVES YOU SOMETIMES, AN IDEA OF WHAT THE ANSWER IS AND SOMETIME GIVES YOU THE ANSWER.
17) It was someway easy , because there we're most of the words that I didn't understand , but I go threw it . It was very interesting to learn about Fossil Fuels. I learned a whole lot.
18) Because I don't have to guess what I think the answer is I can just find it in the handout.
19) the pics and the examples.
20) there were times where they talked about something very small in the articel.
21) Most of the answers are right in the handout, just worded differently.
22) The handout supplied all the information, but some was hard to find. I could use it to answer most of the questions but not all of them.
23) cause of all the pics.
24) This paper did have all of the information needed, but there were spots where I had to make guesses because of the paper not directly giving the information, but implying some of it. This paper did have some information I already knew, which made it a good bit easier too.
25) I liked learning about this because this was more intersting but in some questions I did not understand.
26) I think it was somewhat easy because it didn't have enough diagrams.
43) Overall, it was a little easy, but parts were difficult.
44) I found it somewhat easy because it had both pictures and words so I wasn't overloaded with words and pictures. I had an equal mix of both.
45) The handout helped me with most of the information that I needed, but I couldn't find some of the information that I needed.
46) I LIKED IT BECAUSE IT WAS FUN AND IT HELPED ME A LITTLE MORE THAN NORMALLY AND IT WAS MUCH MORE FUNNER THAN MOST OF THE OTHERS THAT WE HAVE SURVEYED ON. AND I LEARNED MORE ABOUT FOSSIL FUELS AND WHAT KIND THERE ARE TO USE OR THAT ARE USED BY ELECTRICITY. IT HELPED ME MORE AND MORE EVERYTIME I DO A HANDOUT OR SURVEY LIKE THIS!
47) It wasn't easy, it wasn't hard
48) Yes
49) Because if you read the passage right and read the questions carefully it was easy.
50) because most of the information is written out for you in the handout
51) This time I read then answerer the questions
52) It was ok but there was one question I didn't understand. But besides that it was alright.
53) there was some words that i didn't understand in the survey even with reading it with context clues. there was about 3 ords that i didn't get of understant what they ment to say even with reading it with the answer choices or trying to read it with context clues.
54) to help me find the answeres to the survey.
55) Some of the answers were hard to find in the selection but other than that it was easy.
56) I had just read the paper, but some of the diagrams were not completely clear to me.
57) because i already know
58) because it wasnet hard and it wasnet interrly easy either
59) The handout gave you some diagrams and shows a person how things work such as a Fossil Fuel Power Station or the process that it takes to make fossil fuels.
60) I found it somehow easy because the handout basically had the answers right their all you had to do was read, but some question i didnt get ’ so they were difficult to answer. most of the questions were easy maybe two or three that i did not quit get .
61) Yes
62) because it didnt give me the exact words but it did give me an idea about the answers
63) Some of it was hard, because I don't really like this kind of stuff so it was harder to understand, but some of it was easy.
64) It is just the way I am, when reading, i don't focus, so, the answers were a little difficult. I also couldn't find some of the question answers in the text.
Treatment #1 session, “somewhat difficult”:

1) because i couldn't find out the Where can rivers and lakes be found?
2) Because im tired.
3) because the answers werent mworded out on the hand out
4) Because it was hard to tell kind of like where the information was and what the hand out was saying. I didnt understand some of the questions and like how to find the answer.
5) its a bit hard to find the information in the sheet
6) It was somewhat difficult because learning about new items takes awhile for me to process.
7) It did not have enough information.
8) I personally do not like to read so this made it very boring and me not really pay attention to the exercise.
9) I found the handout to be somewhat difficult because i could not find all the exact answers. I did not find it to hard but i would have preferred a handout instead of a picture. With words it is easier to find.
10) I thought that answering these questions while using this handout was somewhat difficult because it was not easy to look at a page in the handout and find the answer quickly. You had to read through the information carefully so that you could pick out the correct answer choice for the question. Some question's correct answer choices were easier to locate in this handout than other ones. This process was a bit challenging but overall I believe that I have learned a lot from this handout.
11) The questions did not match the reading I thought.
12) couldnt find everything
13) it was some what difficult because it tells you someparts are at.
14) i dont really know why
15) Because i had a sort of hard time understanding the format of the questions and most of the answers did not mach the information given to me so that i would understand.
16) It was hard and easy at the same time because like thier was alout of questions i did not get and i tried my best to answer this questions.
17) Because I can really read it!! But it helped me on the question.
18) because there was not all of the information on the paper to answer the questions and because some stuff i didnt understand
19) I found it somewhat difficult because the answers were not as clear as they sometimes are.
20) Usually i like to read the questions before reading the text because it helps me focus on the information that is important. This is a good way for me to be successful. Its also part of a stratagy called RUNNERS Read the title and predict Underline key words in the questions Number the paragraphs Now read Enclose or highlight key information in the text Re-read the questions Solve after eliminating answers down to a 50-50 chance
21) Some of the question did not word it so I could understand it but a lot of the questions were easy to read.
22) I did not find all the answers in my hand out
23) some of the questions had some things to do with it and some didnt
24) Because I dont remeber much else about this topic from when we studied it.
25) I found that the handout was a little confusing. Some of the answers to the questions were hard to find, so hard that I had to read back over them. When normally I can just remember the answers. Also I thought where they put certain informations was bizarre.
26) becuase i didnt really understand the way it goes or in other words the order. i felt it a little diffuclt so i did not know how to answer the questions correctly.
27) Only because you really had to look for the answers in the packet because they were just not right in front of your eyes. Also because things in the packets were worded differently than the questions which made it some what confusing.
28) It was hard to find information in hand out. The quistions were confusing.
29) I think questions that were asked that you had to kinda think of yourself were kind of difficult, and had you thinking.
30) They never said where it rained first.
31) the questions in the packet was a diagram in that packet
32) Becuase all of the readings were seperated making it harder to read what was going on around the words and what those things mean't.
33) I did not really understand the mterial, so i couldnt answer the questions easily
34) I think it was difficult because my hand-out was all picters and a few captions.
35) It was a bit hard, in some of the questions, you didn't cover the material.
36) It was hard to interpret what the handout was trying to say just by looking at the pictures and the captions. Also, there were so many arrows that they lost their meaning. The handout had very little words and too many pictures.
37) I couldn't really find much information in the handout that I needed for the questions. I kept looking in the handout, but I didn't seem to find my information.
38) I couldn't find all the answers I needed.
39) because of the questions it asked me, and the handout did really tell me alot about drainage basin's and divide's.
40) it had tricky answers
41) because the answers were easy to find and the questions were worded well so well infact that i got them all and didnt need any help
42) it was hard to understand some of the information because it was all spread out over the page and it wasent orginized and there wasent an order to read it in
43) It was kinda hard to find all the answers in the artical and some questions I didn't get that easily. So it was somewhat difficult.
44) I didn't get to research the topic as much as I normally do and I didn't get to study. I also didn't know that I was going to take a Quiz on this topic today it was not expected.
45) At first I didn't understand lots of things but when I answerd some questions it made it a little easier. It wasn't very difficult, but some questions were kind of hard.
46) what i found out was that so many things can happen to the rivers how they are formed in so many diffrnt way i never that there would be so many diffrnt things that i could learn about. but with some of the questions they were a little hard and some was very easy.
47) Because my head hurts, i can't concentrate, my stomach hurts, and i'm itchy.
48) Well, I enjoyed reading the paper, but I feel like some of the questsins were not in the reading, so they were hard to figure out, but some were easier than others, thats why I didn't say that this handout was "difficult"
49) this hand out was not that dificult i used some of the own stuff i knew
50) I thought the reading part of this activity was fairly easy but became more difficult when it was time to answer the questions.
51) I don't really understand science that much and I don't do well when i don't like something and don't have enough sleep.
52) it seemed like that some of the sentences were worded weird.
53) The names of what each type of thing was
54) I just thought that the article didn't give very much information, although the questions were easily understood.
55) because the questions werent easy and the information wasnt as imformative as the questions wanted.
56) Well because some of the questions were sort of hard to understand and they were confusing. Like some of the questions were hard to understand and stuff and they were hard to find on the handout so it was sorta challenging.

Treatment #2 session, “somewhat difficult”:

1) Just because.
2) because it is mostly words and i don't like reading
3) There wasn't enough information
4) I found working with this handout somewhat difficult because I don't think of myself as a visual learner, so looking at all of the diagrams and pictures was a bit challenging for me. However, some people might have been more successful with the pictures and diagrams instead of just plain text on a page. I think that it was a good idea to create a textbook with all of these pictures and diagrams, I had to adjust to all of the pictures and diagrams on the pages but it was nothing major and wasn't much of a problem. [first name]
5) I felt like some of the questions were a little challenging, and I needed to go back to the reading to figure them out.
6) This was somewhat difficult because I often use the terms "lake" and "pond" interchangeably so it became confusing to realize that they are two different things.
7) Because some of the information was a little hard to comprehend what the question was asking and what the question referred to on the handout.
8) Because of the questsins and becaue the atitlee they gave me was somewhat dificult to read to many stuff and i had to find out in what order to read the aticle
9) There was a lot of information, and some parts i didn't understand such as comparing lakes and
ponds. I couldn't find the information for some of the questions also.
10) I found this activity somewhat difficult because I did not find all the information, for example, one of the questions asks that I state something common about Winter Lakes and Summer Lakes, all the answers I could not find in the text or pictures.
11) the questions were kind of confusing
12) because it was hard the you look back in the handout
13) It was hard to find answers
14) It was somewhat difficult because not all of the questions were clear.
15) Because I did not understand most of the words that are in the text and some of the questions I did not understand.
16) It is kind of hard to find the info
17) Because it had all of the reading on the prompt and it help me well.
18) Because I did not know that much about the topic and the information that was being shared.
19) Working with a handout was not difficult but it was not all that helpful.
20) Because i can learn from paper instead of just powerpoints or from the teacher i quess.

21) Kind of hard, I had to keep looking back at the handout for minor information.
22) because i did not understand the packet
23) because im not very good with the computer or with finding information out of worksheets but i was able to do it becauseim smart. -[full name]
24) I don't know it just seemed harder to me.
25) the some of the things that i didn't get that was not in the handout for me to understand or for me to answer, but most of all did really good.
26) because I get more stressed out with a handout
27) i liked working on this article because it informed you on how lakes or ponds were made talks about how lakes are formed, & what happen in every season.
28) More straigh tforward and every thing of information together
29) Well, it was hard to keep focus on the reading. So i didnt have as much answers as i was hoping because it was hard for me to just continue reading because i always had something else to think about and stuff so it was hard to stay on topic.
30) it was difficult because I just had some problems finding the answers in the article.

Treatment #3 session, “somewhat difficult”:

1) no comment
2) Because some of the words are big words
3) I thought that this article was somewhat diffucult because some of it was a little confusing. Such as the order and what the picture shows.
4) I thought that working with this handout to answer the survey questions was somewhat difficult because I did not know much at all about fossil fuels when I first started to read the handout. Also, I find this subject (fossil fuels) to be a little confusing even with the handout to help me out. However, I thought that the handout was very well constructed, because it had a good balance between the plain text and the diagrams/pictures. This handout suited my learning style better than the one we did last Monday about water turnover, lakes, and ponds, because that one had too many pictures and not enough plain text. The pictures and diagrams kept this handout interesting and the words provided large amounts of information as well. This handout's format made it more simple for me to the answer the questions on the survey. Thanks. [first name]
5) didn't see peat for a while
6) because it had more pictures than words and i do not understand new things that i learned with out having enough detailed descrition.
7) This handout was somewhat difficult because I did not quite understand everything.
8) it was just difficult
9) it was somewhat difficult because it mostly had picture and it was some times hard to now were it was. i will rather read in a paragraph instead of pictures.
10) Becuase some of the word where not in the story. and so that was kind of hard.
11) Because the handout examle explation thing was kind of complacated to read
12) Some of the information was really hard to find in the selection. Example: what "peat" was. It wasn't in the text which is where I usually look for information first, and I read the text about 5 times just to make sure I didn't skip it. And some other information was in the pictures too, which took awhile to make sure it was right. It wasn't all that difficult though, because the other information that was in the text, I got right away, because I somewhat remembered what it said. The questions were pretty easy to get, and weren't something that a 7th grader couldn't understand. I didn't get some
of the words, and had to read over, and use context clues to figure out what they actually meant.
13) It's easier to work with the handout because the handout had the answers so all you had to do was look back & try to find the answer.
14) I found this handout to be difficult because I found it hard to find all the answers in the worksheet and I do not have a lot of experience with fossil fuels.
15) It was somewhat difficult because I got confused by some of the words but I would figure it out so it was kinda hard but kinda easy.
16) because i couldn't see where every thing was located it was just a bunch of words maybe some charts and some more pictures with detailed footnotes
17) All the diagrams were confusing. I found that because pf all the in depth pictures with different colors. I got distracted. The diagrams were confusing because on the bottom left diagram I did nbot know what way you wer supposed to look at it. So I at first looked at it backwards which thoroughly confused me.
18) because some of the question are so hard because i don't get what are they asking
19) There were words that I do not understand .
20) I found this activity somewhat difficult because some of the information that is asked in the questions couldn't be found in the text.
21) Some of the answers where in the book but the the other answers weren't in the story
22) All the information was scattered and not organised that well. Yes the pictures did help but It was very confusing and crazy!
23) It was somewhat difficult, first i didnt get what it was telling me, then i just stopped looking at the questions and read the little packet over 3 times then, still didn't get it, i guessed on one of the questions, then i found the right answers to the rest of them. it was somewhat difficult
24) Because it was to confusing and I am kind of tired so it's harder for me to think properly.
25) i didnt understand some of the questions that were ask about this exercise
26) Well some of the questions you had to think super hard about, an the topic (fossil fuels) are confusing to me
27) i could not find all the information in the last few questions
28) There was a lot of different stages in the type of coal or fossil fue that is being used. It was a little hard to keep up with the different typed or sources of energy. This exercise was fun to learn about and if i had the chance to learn more about it i would.
29) there to many things to real and it to much. some people don't like working with handouts
30) because in the handout it didn't talk about some of the answere and it was hard for me because i didnt know the stuff.
31) I find it difficult when I'm not really into it and I doesn't seem to put my thought to it.
32) Because the hand out didnt really have good wording, and it doesn't walk the reader through it.
33) I like doing information that is straight foward with wording, and it doesn't very interesting
34) I got distracted by all the pictuers and the diagrams were a little too much.
35) well it was hard to get into the selection so it was harder for me to try to remember what i read. also it was hard to find the info i needed it was sorta scattered everywhere

Treatment #1 session, “difficult”:

1) I dont realy get precipitation and water cycles also I did not realy understand it.
2) iT WAS difficult cause back in florida we weren't learning about bodies of water and stuff we were learn about habitat , the circle of life sexual and reproduction all that tpe of stuff. We were learning development.Cause of a Mass extinction
3) I didnt understand the content and some of it did not make sense I wanted most of it to be based off the selections. Very Difficult!
4) I found this handout to be difficult because some of the information was found by looking at the pictures, and I am used to reading the words and ignoring most of the pictures.
5) Some of the information I had to guess about because it wasnt on the handout.
6) because some of the question I did not understand but the pictures on the side helped a little to explain what the meant.
7) It was difficult because there was not a lot of writing and to me i like when there is a lot of writing.
8) The questions were difficult, and confused me. I couldn't have answerd any of the questions without the handout. The handout would make me get confused about to different things such as the divide and the drainage basin.
9) Because I don't well english
10) I found it difficult to do this part of the study because I just felt like the paper wasn't giving enough information. I'm also not very good with the
landscape subject in science, but other subjects are easy for me.

11) Because I couldn't really understand the information that was on it.

12) Sorry, I do not understand

13) I found it difficult because I have never learned this criteria before and I felt I still knew hardly anything about after I read.

14) I did not understand the text. It have hared words that I did not understand.

15) BEAUSE I COULDN'T FIND THE THINGS I NEEDED REALLY TO FIND THE ANSWER I NEEDED OR SO ME OF THE QUESTIONS WAS HARD TO UNDERSTAND CORRECTLY

16) Because the paper we where given did not tell us all the answers to the questions and if the paper did tell us it made us assume a lot of information. Also the toppic the paper was on was not that familiar!

17) Because, it didn't have as much informtaion as i thought it was going to have it it didn't help me as much as i thought it was going to help me. So the information that i did see i just went ahead and answered the questions that were on this survey.

18) I learned about divides and ranges and etc. I'm still confused from reading that though.

19) I did not understand everything that was in the artical.

20) I thought it was difficult because everything I read didn't make any since to me and then the questions were just to hard for me.

21) because i think that if you really didn't pay attention to the handout and read every word you wuldnt know how to do it.

22) There were alot of words and it didnt give me as much specific informations that is needed and alot of the information is mixed up and there are no subheadings.

23) Because, I thought the handout was confusing, it wasn't even in paragraph form! There were words all over the place, and it was hard to find answers.

Treatment #2 session, “difficult”:

1) i found it to be difficult because some questions i didn't get what it was asking me

2) Because, like I said before, I am not good with reading about boring subjects that just don't interest me. If it were Greek Mythology or Science Fiction then it would be different.

3) Because

Treatment #3 session, “difficult”:

1) because this is like hard for me some times and in pluse i dont like taking my time and i am telling the truth

2) This new info to me, Aand it takes time 4 me 2 understaned it.

3) cause it is not interesting at all

4) I found this to be difficult because most of the answers could not be found directly in the passage.

5) Becuase i don't likeing to read a handout al about only coal .

6) Because english is difficult.

7) because the articl was very had to read, understand, and it was hard to form information in my head

8) It was more difficult because it doesnt directly tell you in the reading so that it could help you answer the questions.

9) it was kind of hard to understand the text therefore making the questions hard

10) It was some what hard becuase the questions were not that clear

11) IT HAD MANY ANSWERS THAT SEEMED RIGHT

12) cause it was

Treatment #1 session, “very difficult”:

1) because i did not understand nothing that they were talking about so i just guessed because i was stuck on some of the quesstions

2) Because they were asking me hard questions that caused me to think hard and caused my brain to start hurting.
3) Because the text wasn't easy to appeal to my senses. I say this because I do not like to stare at stuff I am not interested in for even a short time.

4) i am not that good at science and it was really hard for me because it was not easy and it had big stuff i havent learned

Treatment #2 session, “very difficult”:

1) because when i was reading it i barely understood anything that they were talking about so when the qestions came i was stuck so i just guessed in all of the qestions and i am telling the truth

2) Because I didn't read the handout so I had to use my brain and make a good guess.

3) Because of the lack of food i've eaten today.

4) I find this handout very difficult because this is hard for me.

5) this was difficult because i'm really bad at science and i think it was a lot of work

Treatment #3 session, “very difficult”:

1) I didn't get what the questions were asking me

2) because I am not america people, so I do not understand many word.

3) I didn't know most of it and I'm not use to doing this kind of thing.

4) because i di not understand the packet

5) I didn't get much sleep and i forgot to take my medication this morning. I couldn't concentrate at all.

6) because some of the stuff that was the questions wasnt in the reading and it was hard to find and the stuff that wasnt in the reading was stuff i had to think about and i didnt know anything in it