SCIENTIFIC REASONING: RESEARCH, DEVELOPMENT, AND ASSESSMENT

DISSERTATION

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

By

Jing Han, B.S., M.S.

*****

The Ohio State University 2013

Dissertation Committee:

Professor Lei Bao, Adviser
Professor Fengyuan Yang
Professor Andrew F. Heckler
Professor Evan R. Sugarbaker

Approved by

_______________________
Adviser
Physics Graduate Program
ABSTRACT

Education in Science, Technology, Engineering, and Math (STEM) is emphasized worldwide. Reports from large-scale international studies such as TIMSS and PISA continually rank U.S. students behind many other nations. As a result, the U.S. has increased its emphasis on the implementation of a more extensive science and mathematics curriculum in K-12 education.

In STEM education, widely accepted teaching goals include not only the development of solid content knowledge but also the development of general scientific abilities that will enable students to successfully handle open-ended real-world tasks in future careers. One such ability, scientific reasoning, is closely related to a wide range of general cognitive abilities such as critical thinking and reasoning. Existing research has suggested that scientific reasoning skills can be trained and transferred. Training in scientific reasoning may also have a long-term impact on student academic achievement. In the STEM education community, it has been widely agreed that student development of transferable general abilities is at least as important as certain learned STEM knowledge. Therefore, it is important to investigate how to implement a STEM education program that can help students develop both STEM content knowledge and scientific reasoning.
In order to develop such a knowledge base and to assess and evaluate the impact and effectiveness of education methods and resources, we need good assessment tools that can be easily applied in large scale and produce valid results comparable across a wide range of populations.

In the area of scientific reasoning, there exists a practical tool, the Lawson’s Classroom Test of Scientific Reasoning, and since its initial development in the late 1970s and early 1980s, the test has undergone several revisions with the current version released in 2000. Although the Lawson’s test has provided much useful information for research and assessment purposes, the test itself hasn’t been systematically validated and several issues have been observed in large scale applications concerning item designs and scalability of the results (details will be provided in later sections). Therefore, there is an urgent need for systematic research on validating the Lawson’s test and further development of validated standardized assessment tools on scientific reasoning for K-18 students.

This dissertation project establishes a first step to systematically improve the assessment instrumentation of scientific reasoning. A series of studies have been conducted:

(1) A detailed validation study of the Lawson’s test, which has identified a number of validity issues including item/choice design issues, item context issues, item structure and wording issues (e.g. two-tier design), the limited scale of measurement range, and the ceiling effect for advanced students.
(2) A study to determine the basic measurement features of the Lawson’s test with large scale data.

(3) A data-mining study of Lawson’s test data, which helps identify learning progression behaviors of selected scientific reasoning skills. The results also provide evidence for researchers to evaluate and model the scoring methods of two-tiered questions used in the Lawson’s test.

(4) A study with randomized testing to investigate the learning progression of the skill of control of variables (COV), which showed a series of fine grained intermediate levels of COV skills.

This project produces rich resources for sustained research and development on scientific reasoning. It establishes a valuable baseline for teachers and researchers to apply the Lawson’s test in research and teaching and a solid foundation for researchers to further develop the next generation assessment instruments on scientific reasoning.
ACKNOWLEDGMENTS

I am extremely grateful to my adviser, Prof. Lei Bao, for all his support since my initial meeting to find out about working with the physics education research group.

I would like to thank Prof. Andrew Heckler, Prof. Fengyuan Yang, and Prof. Evan Sugarbaker for serving on my committee.

I also would like to thank fellow graduate students in the physics education research group for all the helpful discussions and support.

I thank my family and friends for their unconditional support.

Most importantly, I would like to thank the students for their participation and cooperation with this project.
VITA

2008-Present…………………………… Research Assistant, Department of Physics, 
The Ohio State University

2010…………………………………… M.S. Physics, The Ohio State University

2006…………………………………… B.S. Physics, Capital Normal University, Beijing, China

PUBLICATIONS

Jing Han, Li Chen, Yibing Zhang, Shaona Zhou, Lei Bao, “Seeing what students see in doing the Force Concept Inventory”, Am. J. Phys., In Press.

Lei Bao, Amy Raplinger, Jing Han, Yeounsoo Kim, “Assessment of Students’ Cognitive Conflicts an Anxiety”, Journal of Research in Science Teaching, Submitted.

Li Chen, Jing Han, Lei Bao, Jing Wang, Yan Tu, “Comparisons of Item Response Theory Algorithms with Force Concept Inventory Data”, Research in Education Assessment and Learning, 2 (02), 26-34, (2011).


Jing Han, Dan Li, “Research in Moral and Law Education in XuanWu District”, Excellent Thesis of Special Undergraduate Grant for Research CNU, 2006. ( in Chinese)

FIELDS OF STUDY

Major Field: Physics
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td></td>
<td>ii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td></td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td></td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td></td>
<td>xii</td>
</tr>
<tr>
<td>List of Tables</td>
<td></td>
<td>xiv</td>
</tr>
</tbody>
</table>

Chapters:

1. Introduction to Research on Scientific Reasoning

2. Research on Assessment of Scientific Reasoning

1.1 What is Scientific Reasoning? | 1
1.2 Why is Scientific Reasoning Important? | 9
1.3 How is Scientific Reasoning Learned? | 12
1.4 How is Scientific Reasoning Assessed? | 13
1.5 Scientific Reasoning, an Important Component of the 21st Century Skills
   1.5.1 Skills Gap between Schools and Workplaces | 17
   1.5.2 What are 21st Century Skills? | 19
1.6 Outline of the Thesis | 22

2.1 Theoretical Background of Scientific Reasoning | 24
2.2 Existing Research and Tools on Assessment of Scientific Reasoning
   2.2.1 Group Assessment of Logical Thinking (GALT) | 29
   2.2.2 The Test of Logical Thinking (TOLT) | 31
   2.2.3 Lawson’s Classroom Test of Scientific Reasoning (LCTSR) | 32
2.3 Expanding the Dimensions of Skills for Assessment of Scientific Reasoning
   2.3.1 Control of Variables | 37
   2.3.2 Proportions and Ratios | 42
Chapter 3. Validity of the Lawson’s Classroom Test of Scientific Reasoning........ 60

3.1 A Historical Review on the Development of the Lawson’s Test.............. 60
3.2 Content Evaluation of Lawson’s Test – Item Context Issues................. 66
3.3 A Data-Driven Study on the Validity of the Lawson’s Test.................. 67
3.4 Quantitative Results – Item Score Analysis...................................... 69
3.5 Quantitative Results – Analysis of Two-Tier Score Patterns.................. 72
3.6 Qualitative Results – Analysis of Student Interviews........................... 75
3.7 Consideration on Two-Tier Question Structure.................................. 84
3.8 The Ceiling Effect and Measurement Saturation of the Lawson’s Test...... 85
3.9 Conclusions......................................................................................... 88

Chapter 4. The Developmental Metric of Scientific Reasoning....................... 89

4.1 Context of the Study............................................................................ 89
4.2 Data Collection.................................................................................... 90
4.3 The Developmental Scales of the Lawson’s Test Scores....................... 93
  4.3.1 The Learning Evolution Index Curve of the Lawson’s Test............. 93
  4.3.2 The Developmental Scales of the Six Skill Dimensions of the Lawson’s Test.......................................................... 98
  4.3.3 The Developmental Curve of the Conservation of Mass and Volume... 99
  4.3.4 The Developmental Curve of the Proportional Reasoning............... 101
  4.3.5 The Developmental Curve of the Control of Variables.................. 103
  4.3.6 The Developmental Curve of the Probabilistic Reasoning............. 105
  4.3.7 The Developmental Curve of the Correlation Thinking.................. 108
  4.3.8 The Developmental Curve of the Hypothetical-deductive Reasoning... 110
4.4 Summary............................................................................................ 112

Chapter 5. Study Learning Progression of Scientific Reasoning Skills through Pattern Analysis of Responses to the Lawson’s Test............................................. 114

5.1 Context of the Study............................................................................ 114
5.2 Review of Learning Progression in the Context of Scientific Reasoning... 114
5.3 Research Design to Study Learning Progress in Scientific Reasoning....... 123
5.4 Data Collection.................................................................................... 130
5.5 Data Analysis and Results................................................................. 131
  5.5.1 Result 1: Defining a new level in the scoring of the Lawson Test.... 136
5.5.2 Result 2: Combined patterns of responses as indicators for performance levels ................................................................. 139
5.5.3 Result 3: Proposing a three-level scoring system for the Lawson’s Test.. 145
5.6 Conclusions ................................................................................................................. 149

Chapter 6. A Case Study on Fine Grained Learning Progression of Control of Variables ........................................................................... 152

6.1 Context of the Study ................................................................................................. 152
6.2 Review of Research on Control of Variables ....................................................... 153
6.3 Research Design ..................................................................................................... 159
6.3.1 Research Questions and Goals ......................................................................... 159
6.3.2 The Design of the Assessment Instrument ...................................................... 160
6.3.3 Data Collection .................................................................................................. 165
6.4 Data Analysis and Results ...................................................................................... 166
6.4.1 Impact of giving experimental data in a question on student performance ........ 166
6.4.2 Impact of question context on student performance ....................................... 170
6.4.3 Impact of embedded relationships between variables on student performance ................................................................. 174
6.6 Conclusions and Discussions ............................................................................... 179

Chapter 7. Summary ..................................................................................................... 183

References ..................................................................................................................... 187

Appendix A: Group Assessment of Logical Thinking (GALT) ......................... 206
Appendix B: The Test of Logical Thinking (TOLT) ................................................... 212
Appendix C: Lawson’s Class Room Test of Scientific Reasoning ......................... 218
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 4.1.</td>
<td>The developmental trend of Chinese and U.S. students’ total LCTSR scores</td>
<td>96</td>
</tr>
<tr>
<td>Figure 4.2.</td>
<td>The developmental trends on conservation of matter and volumes</td>
<td>103</td>
</tr>
<tr>
<td>Figure 4.3.</td>
<td>The developmental trends on proportional reasoning.</td>
<td>104</td>
</tr>
<tr>
<td>Figure 4.4.</td>
<td>The developmental trends on control of variables</td>
<td>107</td>
</tr>
<tr>
<td>Figure 4.5.</td>
<td>The developmental trends on probabilistic reasoning.</td>
<td>109</td>
</tr>
<tr>
<td>Figure 4.6.</td>
<td>The developmental trends on correlation thinking.</td>
<td>112</td>
</tr>
<tr>
<td>Figure 4.7.</td>
<td>The developmental trends on hypothetical-deductive reasoning.</td>
<td>114</td>
</tr>
<tr>
<td>Figure 5.1.</td>
<td>Items from the Lawson’s Test used in this study.</td>
<td>131</td>
</tr>
<tr>
<td>Figure 5.2.</td>
<td>Percentage of grades 3-12 at the six levels of Lawson Test performance.</td>
<td>146</td>
</tr>
<tr>
<td>Figure 5.3.</td>
<td>Percentage of grade groupings at each of the six levels.</td>
<td>148</td>
</tr>
<tr>
<td>Figure 5.4.</td>
<td>Traditional, individual, and three-level Lawson Test scoring.</td>
<td>152</td>
</tr>
</tbody>
</table>
Figure 6.1. Test questions on COV with experimental data. Question 1 poses a COV situation using a real-life context. Questions 2 and 3 are in physics contexts and are based on the tasks used in Boudreaux et al. (2008).

Figure 6.2. Mean scores on Test A (data not given) and Test B (data given). The error bars (approximately ±0.04) represent the standard error.

Figure 6.3. The mean scores on Test A (data not given) and Test B (data given) for each context. The real-life context shows a greater difference between the means of Tests A and B than the physics context. The error bars (±0.04) indicate the standard errors.
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 3.1. Lawson Test two-tiered response patterns of U.S. college freshmen (N=1699).</td>
<td>75</td>
</tr>
<tr>
<td>Table 3.2. Comparison of Lawson test total percentage scores of U.S. college freshmen (N=1699) calculated with paired-score vs. single-question-score methods.</td>
<td>75</td>
</tr>
<tr>
<td>Table 4.1. Summary of Lawson’s Test Data from USA and China. COV (Control of Variables), HD (Hypothetical Deductive)</td>
<td>94</td>
</tr>
<tr>
<td>Table 4.2. The model fit parameters and the root-mean-square deviations (RMSD) of the fit for the mean scores and population.</td>
<td>98</td>
</tr>
<tr>
<td>Table 4.3. The six skill dimensions of the Lawson’s test.</td>
<td>101</td>
</tr>
<tr>
<td>Table 5.1. Traditional scoring on a two-tier item from the Lawson Test.</td>
<td>128</td>
</tr>
<tr>
<td>Table 5.2. Student performance on two easy and two difficult questions.</td>
<td>130</td>
</tr>
<tr>
<td>Table 5.3. Distribution of collected student data across different grade levels.</td>
<td>134</td>
</tr>
<tr>
<td>Table 5.4. Responses to Lawson Test items P1, P2, F1, and F2 from grades 6-7, 9-10, and college.</td>
<td>138</td>
</tr>
</tbody>
</table>
Chapter 1. Introduction to Research on Scientific Reasoning

1.1 What is Scientific Reasoning?

Scientific reasoning, also referred to as “formal reasoning” (Piaget, 1965) or “critical thinking” (Hawkins and PEA, 1987) in early studies, represents the ability to systematically explore a problem, formulate and test hypotheses, control and manipulate variables, and evaluate experimental outcomes (Zimmerman, 2007; Bao et al., 2009). It represents a set of domain general skills involved in science inquiry supporting the experimentation, evidence evaluation, inference and argumentation that lead to formation and modification of concepts and theories about the natural and social world.

There exists a large body of research on the multifaceted aspects of scientific reasoning. Zimmerman (2007) made a comprehensive review on the related work using the Klahr’s (2000, 2005) Scientific Discovery as Dual Search (SDDS) model as the general framework that organizes the main empirical findings in three areas including experiential skills, evidence evaluation skills, and integrated approaches in self-directed experimentation (Klahr, 2000, 2005; Zimmerman, 2007). Kuhn (2002) has argued that the defining feature of scientific thinking is the set of skills involved in differentiating and coordinating theory and evidence (Kuhn, 1989, 2002). The specific set of skills in scientific reasoning included the isolation and control of variables, producing the full set of factorial combinations in multivariable tasks, selecting an appropriate design or a
conclusive test, generating experimental designs or conclusive tests, record keeping, the
inductive skills implicated in generating a theory to account for a pattern of evidence, and
general inference skills involved in reconciling existing beliefs with new evidence that
either confirms or disconfirms those beliefs (Zimmerman, 2007). Elements concerning
casual mechanisms (Koslowski, 1996) and epistemological understandings (Chinn and
Malhotra, 2002) have also been carefully examined and debated.

From a more operational perspective, scientific reasoning is assessed (and
operationally defined) in terms of a set of basic reasoning skills that are commonly
needed for students to successfully conduct scientific inquiry, which includes exploring a
problem, formulating and testing hypotheses, manipulating and isolating variables, and
observing and evaluating the consequences. The Lawson’s Test of Scientific Reasoning
(LTSR) provides a solid starting point for assessing scientific reasoning skills (Lawson,
1978, 2000). The test is designed to examine a small set of dimensions including (1)
conservation of matter and volume, (2) proportional reasoning, (3) control of variables,
(4) probability reasoning, (5) correlation reasoning, and (6) hypothetical-deductive
reasoning. These skills are important concrete components of the broadly defined
scientific reasoning ability.

Although there exists a wide range of understandings on what constitutes scientific
reasoning, the literature seems to generally agree that scientific reasoning represents an
important component of science inquiry. Therefore, a good understanding of the nature
of scientific reasoning requires extended knowledge of science inquiry.
Scientific inquiry has its roots in the early research on constructivism and reasoning. Vygotsky (1978) stated that children learn constructively when new tasks fall into the zone of proximal development. That is, if a task is one that a child can do with an adult’s help, then the child can eventually learn to do this task on their own by following the adult’s example. The idea that children build on existing knowledge is also reflected in Inhelder and Piaget’s (1958) work with formal reasoning development. Their model describes the levels through which children progress from birth (sensorimotor stage) to adulthood (formal operational stage).

This pioneering work is the foundation for two schools of thought on student learning: cognitive conflict and scaffolding. Cognitive conflict occurs because students often come into the classroom with established beliefs based on their life experiences. Many of these beliefs are non-scientific with some being strongly held and difficult to change. Therefore, helping students to “change” their non-scientific preconceptions to the expert beliefs has been the main goal of many of the studies on conceptual change. Through research, it has been found that by explicitly recognizing the discrepancy between their current beliefs and the scientific ones (often referred to as the experience of a cognitive conflict), students can be motivated to change their current beliefs, which starts the processes of conceptual change. Posner et al. (1982) identified four requirements for successful conceptual change. Students must have (1) dissatisfaction with their current conceptions, and they must see the new conception as (2) intelligible, (3) plausible, and (4) fruitful. Simply put, students need to first recognize that there is a conflict between their current views and the new information to be learned, and if they
are going to reject their old views, the new idea needs to make sense to them. This type of constructivist learning is the basis for courses such as Physics by Inquiry (McDermott et al., 1996) where conflicts are elicited by the coursework and confronted and resolved by the student with help from the text, peers, and instructors.

Scaffolding is a fundamentally different process from cognitive conflict. While cognitive conflict can lead to rapid changes in student conceptions, scaffolding avoids conflict and uses small steps to build on previous understanding. Coltman, Petyaeva, and Anghileri (2002) describe scaffolding as the way adults supportively interact with children during the learning process. The child can solve a problem on their own (perhaps unintentionally), but the meaning of the achievement or the process that led to it can be lost on the child unless an adult brings it to the child’s attention. In this way, children build on their previous knowledge without being forced to confront conflict.

Both of the conceptual change and scaffolding frameworks play important roles in the current education system regarding the implementation and evaluation of scientific inquiry, which is generally understood as the process used in developing scientific knowledge (Schwarz, Lederman, & Crawford, 2004). The stages of this process include identifying variables, forming a hypothesis, designing an experiment, making observations, collecting and analyzing data, and drawing conclusions. The process is cyclic in nature. Once a conclusion has been reached, the original hypothesis can be revised, which leads to further experimentation. Scientific inquiry is seen in the real-world work of scientists, but it is also used in student-centered open-ended classroom
activities to teach students how to gain scientific knowledge (Roth & Roychoudhury, 1993).

Scientific reasoning skills support the stages of inquiry. Identifying variables and forming hypotheses are primary skills. The skills underlying experimental design are identification and control of variables. Observation and data collection require data-taking and data-organization skills as well as identification of hidden variables. Analyzing data and drawing conclusions are arguably the most complex stages as they require an understanding of correlational and causal relations (with single and multiple variables) as well as the ability to interpret graphical information. At all stages, students need written and oral communication skills to present their ideas coherently.

The operational definition of scientific reasoning includes the necessary skills that support scientific inquiry such as control of variables, hypothetical deductive reasoning, causal and correlational reasoning, proportions and ratios, deductive and inductive reasoning, and probabilistic reasoning. This is not a complete list as one can argue that other dimensions could be included, but the literature suggests that these are commonly agreed-upon scientific reasoning skills. Some dimensions have been studied in great detail, and the results of a selection of these studies are summarized below.

In the realm of experimental design, one reasoning skill that needs to be used is control of variables (COV). In a recent study, Boudreaux et al. (2008) found that college students and in-service teachers had difficulties with basic methods in COV which included failure to control variables, assuming that only one variable can influence a system’s behavior, and rejection of entire sets of data due to a few uncontrolled
experiments. Boudreaux et al. concluded that students and teachers typically understand that it is important to control variables but often encounter difficulties in implementing the appropriate COV strategies to interpret experimental results. Other research has analyzed the capabilities of younger students. Chen and Klahr (1999) asked students to design experiments involving a ball rolling down a ramp to test a given variable and then state what they could conclude from the outcomes. With increasing complexity by involving more variables in contexts of ramps, springs, and sinking objects, Penner and Klahr (1996) and Toth, Klahr, and Chen (2000) had students design and conduct experiments, justify their choices, and consolidate and summarize their findings.

Kuhn (2007) tied together COV and causal reasoning. This study had fourth-graders use computer software to run experiments relating to earthquakes (and ocean voyages). This study used more variables than previous studies mentioned and asked students to determine whether each variable was causal, non-causal, or indeterminate. Kuhn found that students made progress in learning COV skills but struggled when faced with handling multivariable causality.

Student understanding of probability has been studied as well. Fox and Levav (2004) focused on conditional probability and found that irrelevant information and problem wording influenced responses. Denison et al. (2006) found that children as young as four years old can understand random sampling and perform simple probability tasks. Khazanov (2005) reported that college students hold misconceptions about probability and that this can interfere with their learning of inferential statistics. Interestingly,
Khazanov found that confronting these misconceptions lead to better results than traditional instruction.

Correlational reasoning has been widely studied. Many groups have found that prior beliefs about the relationship between two variables have a high influence on student judgment of the correlation between those variables (e.g. Kuhn, Amsel, & O’Loughlin, 1988). Other research shows that subjects have difficulty in handling negative correlations (e.g. Erlick, 1966, and Batanero, Estepa, & Godino, 1997). Furthermore, subjects have a tendency to make causal claims based on a correlational relationship (e.g. Shaklee & Tucker, 1980). Carlson et al. (2002) found that students have troubles with creating and interpreting graphical displays relating to correlation. Batanero, Estepa, & Godino (1997) found that use of technology seems to improve the strategies that subjects use to analyze correlation.

Hypothetical deductive reasoning is a skill that Lawson (2000) describes as a pattern of “If…And…Then…And/But…Therefore…” seen in experimentation. For example, if a ball is denser than water and the ball is placed in a bucket of water, then it will sink to the bottom; but it is observed that the ball does not sink; therefore the ball is not denser than water. This system of reasoning can be applied to any experiment. Lawson et al. (2000) studied biology students’ use of hypothetical deductive reasoning and labeled three stages of development of this skill (not able to test hypotheses, able to test hypotheses for observable causal agents, and able to test hypotheses for unobservable causal agents). They found that instruction that targeted hypothetical deductive strategies improved student performance on hypothetical deductive reasoning assessment items.
There was also a positive correlation between hypothetical deductive reasoning ability and content performance, but Lawson et al. note that content knowledge is not enough to ensure scientific reasoning ability.

Regarding the development of scientific reasoning skills in formal and informal education settings, it is clearly lay out in the National Research Council’s (1996) National Science Education Standards that the scientific methods and skills students are expected to learn at different grade levels. For example, these standards declare that in fifth through eighth grades, students should learn how to analyze evidence and data, design and conduct experiments, and think critically and logically in making connections between data and explanations.

As science has continued to become fundamental to modern society, there is a growing need to pass on the essential aspects of scientific inquiry and with it the need to better impart such knowledge. Previous studies have indicated that scientific reasoning is critical in enabling the successful management of real-world situations in professions beyond the classroom. For example, in K-12 education, the development of scientific reasoning skills has been shown to have a long-term impact on student academic achievement (Adey & Shayer, 1994). Positive correlations between student scientific reasoning abilities and measures of students’ gains in learning science content have been reported (Coletta & Phillips, 2005), and reasoning ability has been shown to be a better predictor of success in biology courses than prior biology knowledge (Johnson & Lawson, 1998). The above findings support the consensus of the science education community on the need for K-12 students to develop an adequate level of scientific
reasoning skill along with a solid foundation of content knowledge. Zimmerman (2007) claims that investigation skills and content knowledge bootstrap one another, creating a relationship that underlies the development of scientific thinking. Research has been conducted to determine how these scientific thinking skills can best be fostered and which teaching strategies contribute most to learning, retention, and transfer of these skills. Zimmerman found that children are more capable in scientific thinking than was originally thought, and that adults are less so. She also states that scientific thinking requires a complex set of cognitive skills, the development of which requires much practice and patience. It is important, then, for educators to understand how scientific reasoning abilities develop.

A great deal of work has been done analyzing student use of scientific reasoning skills, and an understanding of these dimensions is important in defining scientific reasoning in a broad context. Our current work involves compiling scientific reasoning assessment questions, data, and resources that can be made available to teachers and researchers. We are developing a new assessment instrument, “Inquiry for Scientific Thinking and Reasoning” (iSTAR). An website (www.istarassessment.org) has also been developed, which is focused on research on scientific reasoning and science learning. The site contains compilations of existing research, examples of assessment items, and thorough descriptions of the scientific reasoning dimensions (iSTARAssessment.org).

1.2 Why is Scientific Reasoning Important?
Science inquiry has been widely accepted as the core component of STEM education. Since scientific reasoning represents a set of skills and abilities that are necessary for successfully conducting science inquiry tasks, it has also been widely emphasized in science education standards and curriculum. Much research has also been conducted to understanding how scientific reasoning interacts with other areas of learning. For example, research has shown that scientific reasoning skills have a long-term impact on student academic achievement (Adey & Shayer, 1994). Researchers have found positive correlations between student scientific reasoning abilities and measures of learning gains in science content (Coletta & Phillips, 2005; Lawson et al., 2000). Another study found that students who learned probability in an inquiry-based environment outperformed students who learned in a traditional environment (Vahey, Enyedy, & Gifford, 2000). Shayer and Adey (1993) performed a study comparing students who received scientific reasoning-based teaching with those who did not. Three years after the lessons occurred, the reasoning-based group outperformed the control group on tests in not only science but also English and mathematics. Shayer and Adey argue that instruction in scientific reasoning has a permanent impact on general learning ability.

Scientific reasoning skills are also important because they enter every domain of society. Their place is evident in the educational domain. A desire for students to acquire scientific thinking skills is driving some curriculum development. National standards in education outline various skills that students should have at each grade level (NRC, 1996). While scientific reasoning skills typically fall under the science education standards, teachers in any classroom can promote creative thinking and inquiry learning.
In any subject, teachers have the option of teaching to the test or using an inquiry-based environment to help students develop a full set of skills that can be used beyond the classroom.

Scientists are not the only people who use scientific reasoning skills on the job. In the workplace domain, employers look for individuals who can learn new tasks and utilize problem solving skills. Scientific reasoning skills are the tools that allow one to obtain new knowledge and think critically. Furthermore, inquiry learning can generate an appreciation for exploration that makes students eager and able to try new things, learn from mistakes, and be their own teachers, which is what employers want. Bauerlein (2010) reports results from a study that found 89% of employers said written and oral communication skills are the most important skills for employees to have; 81% of employers listed critical thinking and analytical reasoning. Clearly, scientific reasoning skills are necessary to be competitive in the working world.

Finally, in the social domain, those with scientific thinking skills are capable of handling the wealth of information presented to them on a daily basis. Advertisements, political campaigns, and scientific reports made to the general public all use data to convince the consumer, voter, or citizen of a message. It is important to take a step back and analyze the information, and scientific reasoning skills make this possible.

Certain reasoning skills help in everyday decision-making and problem-solving. Ratio skills are used in determining gas mileage or finding the cheapest brand at the grocery store. Inductive reasoning is used whenever a conclusion is made from limited observations and information. Causal reasoning and probability are used in predicting
weather and assessing insurance rates among many other things. Hypothetical deductive reasoning skills are used in everyday problem-solving. For example, if you are trying to figure out why your television remote is not working, you may test the hypothesis that the batteries are dead by inserting new batteries. If this solves the problem, the experiment is done; if it does not, a new hypothesis is developed. While it may not consciously cross one’s mind explicitly, hypothetical deductive reasoning is being used. This is true of many of these skills – they become part of one’s set of abilities and are used automatically.

1.3 How is Scientific Reasoning Learned?

Scientific reasoning ties in very closely with science inquiry. In developing scientific reasoning, research has shown that inquiry-based science instruction can promote scientific reasoning abilities (Adey and Shayer, 1990; Lawson, 1995; Marek and Cavallo, 1997; Benford and Lawson, 2001; Gerber, Cavallo and Marek, 2001). Additionally, studies have shown that students had higher gains on scientific reasoning abilities in inquiry classrooms over non-inquiry classrooms (Bao et al., 2009). Examples of such learning settings include Physics by Inquiry (McDermott et al., 1996), RealTime Physics (Sokoloff, Thornton, & Laws, 2004), the CUPLE (Comprehensive Unified Physics Learning Environment) Physics Studio (Wilson, 1994), and The SCALE-UP (Student-Centered Activities for Large Enrollment Undergraduate Programs) Project (Beichner, 2008). The goal of these classrooms is to engage students in a way that fosters the development of scientific reasoning skills. Such skills are not inherently learned by the
student, and rigorous scientific education is not enough. It is not what is taught, but rather how it is taught, that makes the difference (Bao et al., 2009). Scientific reasoning skills need to be directly addressed during the course (Schwartz, Lederman, & Crawford, 2004). In a study of pre-service teachers, Schwartz et al. found that providing explicit opportunities to reflect on scientific reasoning (using journals) strengthened views on reasoning. Teachers serve as guides, but having students take time to discuss and reflect on scientific reasoning has a positive impact. The role of the teacher is still important, though, as instructors with higher levels of scientific reasoning skills are found to be more effective in using inquiry methods in teaching science courses (Benford & Lawson, 2001).

1.4 How is Scientific Reasoning Assessed?

In order to understand if students are learning scientific reasoning skills, it is important to have an accurate assessment instrument. Such tools need to be easy to use, practical, and applicable to a variety of educational settings.

Traditionally, the Piagetian clinical interview is used to assess students' formal reasoning abilities, but such a method requires experienced interviewers, special materials and equipment, and is usually time consuming (Inhelder & Piaget, 1958; Lawson, 1978). A number of researchers have used the Piagetian method as a basis for developing their own measurement tools in assessing students' scientific reasoning abilities. Outcomes of this work include the Group Assessment of Logical Thinking Test (GALT) (Roadrangka, Yeany, & Padilla, 1982), the Test of Logical Thinking (TOLT)
Among the various assessment instruments, the Lawson Test has gained wide popularity in science education communities. In the development of his test, Lawson (1978) aimed for a balance between the convenience of paper and pencil tests and the positive factors of interview tasks. Test items were based on several categories of scientific reasoning: isolation and control of variables, combinatorial reasoning, correlational reasoning, probabilistic reasoning, and proportional reasoning. The original format of the test had an instructor perform a demonstration in front of a class, after which the instructor would pose a question to the entire class and the students would mark their answers in a test booklet. The booklet contained the questions followed by several answer choices. For each of the 15 test items, students had to choose the correct answer and provide a reasonable explanation in order to receive credit for that item. In its current form, the Lawson test is a 24-item, two-tier, multiple choice test. Treagust (1995) describes a two-tier item as a question with some possible answers followed by a second question giving possible reasons for the response to the first question. The reasoning options are based on student misconceptions that are discovered via free response tests, interviews, and the literature.

In physics, many education researchers have been using the Lawson test to study the relations between students’ scientific reasoning abilities and physics learning. In a recent study, Coletta and Phillips (2005 & 2007) reported significant correlations ($r \approx 0.5$) between pre-post normalized gain on the Force Concept Inventory and students’
reasoning abilities measured with Lawson’s test. However, research to critically inspect the existing assessment instruments and to develop new instruments for scientific reasoning is largely missing in the literature.

Regarding the Lawson test, for which there are some studies (Stefanich et al., 1983; Pratt & Hacker, 1984) investigating the validity of the Lawson’s 1978 version of the formal reasoning test, there is little work on its 2000 version. Even though the 2000 edition has become a standard assessment tool in physics education research, the test itself has not been systematically validated. Through research, we have also observed several issues concerning the question designs and data interpretations. In Chapter 2, I will give a more in-depth review of the different assessment instruments on scientific reasoning and discuss in detail of the validity and reliability of the Lawson’s test.

1.5 Scientific Reasoning, an Important Component of the 21st Century Skills

We live in an ever-changing world – demographic change, rise of automation and workforce structural change, globalization, and corporate change are some major driving forces that demand fundamental transformations in education and skills on an individual level. Across the globe, work is becoming increasingly bi-polar with jobs sorting out into two clusters - a low-wage, lower-skilled, routine work cluster, going to the lowest global bidder qualified to do the work, and increasingly to automation; and a fast growing, high-paying, creative work cluster requiring a combination of complex technical skills like problem-solving and critical thinking, and strong people skills like collaboration and clear communication. In the U.S., the demand for non-routine skills (expert thinking and
complex communication) is rising fast, as the need for routine and manual skills falls (1960-2002).

Advances in digital technology and telecommunications now enable companies to send works and tasks to be done wherever they can be completed best and cheapest. Meanwhile, political and economic changes in developing countries such as India, China and Mexico have freed up many more workers who can adequately perform such jobs. As a result, not only do Americans have to compete for jobs with foreigners in a rising global labor market, but increasing competition will also center on highly skilled workers for more intellectually demanding and higher paying jobs.

Due to technology development and globalization, companies have gone through radical restructure with less hierarchy and lighter supervision where workers experience greater autonomy and personal responsibility. Work has also become much more collaborative and employees must adapt to new challenges and demands when tackling projects and solving problems.

Consequently, a growing number of educators, business leaders and politicians have called for “21\textsuperscript{st} century skills” being taught as part of everyone’s education. Global competition, increased access to technology, digital information and tools are increasing the importance of 21\textsuperscript{st} century knowledge-and-skills, which are critical for a country’s economic success. Advocates base their arguments on a widening gap between the knowledge and skills acquired in school and the knowledge and skills required in 21\textsuperscript{st} century workplaces. That is, today’s curricula do not adequately prepare students to live and work in a technology-based economy and globalized society. Thus, in order to
successfully face career challenges and a globally competitive workforce, schools must be aligned with real world environments by infusing 21st century skills in education practices.

1.5.1 Skills Gap between Schools and Workplaces

Previous studies have demonstrated a huge skill gap between schools and workplace requirements. In 2005 Skills Gap Report, when asking manufacturing employers which types of skills their employees will need more of over the next three years, basic employability skills (attendance, timeliness, work ethics, etc.) and technical skills were the areas most commonly selected (53%). Following that are reading/writing/communication skills, where 51% of the respondents said they will need more of these types of skills over the next three years. Beyond these, there are a number of related skills that will be needed over the next several years that are characteristic of high-performance workforces, such as the ability to work in teams (47%), strong computer skills (40%), the ability to read and translate diagrams and flow charts (39%), strong supervisory and managerial skills (37%), and innovative/creative abilities (31%). Moreover, manufacturing employers see training as a business necessity and their spending on training is increasing – not just for executives, but across all employee groups. The types of training that most employees receive are technical and basic skills training. The next tier of trainings are for problem solving, teamwork, leadership, computer skills, basic or advanced mathematics, basic reading and writing, and interpersonal skills – all standard skills for high-performance workforces.
Another landmark 2006 research study among more than 400 employers, *Are They Really Ready to Work?*, (conducted by Corporate Voices for Working Families, the Conference Board, the Partnership for 21st Century Skills, and the Society for Human Resource Management), clearly spotlighted employers’ concerns about the lack of preparedness of new entrants into the workforce regardless of the level of educational attainment. More specifically, the deficiencies are greatest at the high school level, with 42.4% of employers reporting the overall preparation of high school graduates as deficient; 80.9% reporting deficiencies in written communications; 70.3% citing deficiencies in professionalism; and 69.6% reporting deficiencies in critical thinking. Although preparedness increases with educational level, employers noted significant deficiencies remaining at the four-year college level in written communication (27.8%), leadership (23.8%) and professionalism (18.6%). In addition, employers reported that top five most important skills are critical thinking and problem solving, information technology, teamwork/collaboration, creativity/innovation, and diversity.

A more recent study, “Across the Great Divide”, released March 2011, surveyed 450 businesses and 751 post-secondary educational institutions and found concerning disparities between the goals of higher education and what businesses sought in workers. The skill gap exists along the entire learning-career continuum – colleges, businesses and the students all had different expectations of what was needed to prepare a workforce for today’s and tomorrow’s jobs. According to the report, employers indicated they believed the most important goal of a four-year degree was to prepare individuals for "success in the workplace" (56%). On the other hand, educational leaders saw higher education as a
way of providing individuals with "core academic knowledge and skills" (64%). The study also found that only 15% of the businesses believed hiring those with an associate degree was a good return on investment for their companies.

Both workers and employers believe that the education sector has the primary responsibility to close the workforce readiness gap. Yet, as surveys indicated, majority of companies do not believe schools are doing a good job preparing students for the workplace. Therefore, continuing contact between schools and businesses is critical to developing a prepared workforce. And it is essential for business leaders, policy makers and educators to work together to address the workforce readiness gap.

1.5.2 What are 21st Century Skills?

So what exactly are 21st century skills? The P21 (Partnership for 21st Century Skills - a group of corporations who partnered with the U.S. Department of Education in 2002) has created a framework that identifies the key skills for success. Based on their categorization and definition, ten skills have been identified as the 21st Century skills, in four groups:

*Ways of Thinking*

1. Creativity and innovation
2. Critical thinking, problem solving, decision making
3. Learning to learn, Metacognition

*Ways of Working*

4. Communication
5. Collaboration (teamwork)

*Tools for Working*

6. Information literacy

7. ICT literacy

*Living in the World*

8. Citizenship – local and global

9. Life and career

10. Personal & social responsibility – including cultural awareness and competence

The essence of these skills includes collaboration, communication, creativity and innovation and critical thinking coined the 4Cs by P21. Many other researchers and authors created lists similar to the 4Cs. For example, Tony Wagner from the Harvard Graduate School of Education interviewed more than 600 chief executive officers, and asked them the same essential question: “Which qualities will our graduates need in the 21st-century for success in college, careers and citizenship?” Wagner's subsequent Seven Survival Skills correspond to the 4Cs but also include agility and adaptability, accessing and analyzing information, as well as curiosity and imagination.

There is agreement among all researchers that these skills of collaboration, communication, creativity and critical thinking are necessary and must be integrated into the classrooms. Indeed, states are adopting new standards to ensure these skills are met. For example, Common Core State Standards have been adopted by most states and several territories in the United States. Common Core State Standards are designed to
provide a national, standardized set of academic standards (organized around 21st century skills) as an alternative to those previously developed by the states on an individual basis. The Common Core Standards are sought to be more rigorous; demand higher-order thinking; introduce some concepts at an earlier age; and allow for interstate comparisons.

On the other hand, the modern workplace and lifestyle demand that students balance cognitive, personal, and interpersonal abilities, but current education policy discussions have not defined those abilities well, according to a special report released by the National Research Council of the National Academies of Science in Washington. Based on the report, 21st century skills generally fall into three categories:

- **Cognitive skills**, such as critical thinking and analytic reasoning, which in the context of STEM learning are established as scientific reasoning skills;
- **Interpersonal skills**, such as teamwork and complex communication; and
- **Intrapersonal skills**, such as resiliency and conscientiousness (the latter of which has also been strongly associated with good career earnings and healthy lifestyles).

A relevant concept that we often hear is the “21st century learning skills.” So what is it? Ted Lai, Director of Information Technology for the Fullerton Elementary School District puts it this way:

"In a nutshell, these are the skills that will help people be globally competitive in the 21st Century. Especially with our students, these are skills that include not only the curricular standards but also a host of other essential skills like communication, collaboration, and creativity. Literacy doesn’t merely refer to the ability to read and
write but also the ability to evaluate and synthesize information, media, and other technology. At the heart of 21st Century Learning, in my opinion, is the piece on creating authentic projects and constructing knowledge... essentially making connections between learning and the real world!"

Clearly, “21st century skills” has become the lasted buzz in education, which has also re-kindled a long-standing debate about content vs. skills. Among the three major categories of 21st century skills, the scientific reasoning is a core component of the “Cognitive Skills”. The existing research on scientific reasoning fully supports the current movement towards training skills rather than content as the goal of 21st century education, and provides practical pedagogy, instruments, and curriculum for developing the 21st century skills.

Although reading, writing, mathematics and science are cornerstones of today’s education, curricula must go further to include skills such as scientific reasoning, critical thinking, collaboration and digital literacy that will prepare students for 21st-century employment and ensure students’ success in the real world. Establishing new forms of assessment can begin a fundamental change in how we approach education worldwide.

1.6 Outline of the Thesis

The research discussed in this thesis focuses on the assessment aspect of scientific reasoning, which is organized in five main parts. Chapter 2 gives a detailed review of the related literature on prior research and existing assessment instruments on scientific
reasoning. In current literature, the Lawson’s test of scientific reasoning is the most widely used quantitative tool in assessing scientific reasoning. However, the test’s validity has not been thoroughly studied. Chapter 3 introduces a study to evaluate the validity of the Lawson’s test. The research has shown a number of test design issues with the current version of the Lawson’s test and also suggested ways to improve the instruments. In Chapter 4, I discuss the study on mapping out a longitudinal developmental scale from 3rd grad to graduate level of scientific reasoning measured with the Lawson’s test. The developmental trends of students from both USA and China are also compared. Chapter 5 introduces a data-mining study of Lawson’s test data, which helps identify learning progression behaviors of selected scientific reasoning skills. The results also provide evidence for researchers to evaluate and model the scoring methods of two-tiered questions used in the Lawson’s test. Chapter 6 gives another case study that investigates the learning progression of the skill of control of variables (COV), which showed a series of fine grained intermediate levels of COV skills. The thesis ends with Chapter 7, which summarizes the entire scope of the work and makes suggestions for future work and development.
Chapter 2. Research on Assessment of Scientific Reasoning

2.1 Theoretical Background of Scientific Reasoning

Research on scientific reasoning is rooted in the early studies on cognitive development of “formal reasoning” (Piaget, 1965) and “critical thinking” (Hawkins & PEA, 1987). Traditionally, the Piagetian clinical interview is used to assess students’ formal reasoning abilities. In Piaget’s cognitive developmental theory, an individual moves to the next cognitive level when presented with challenges in the environment that cause him or her to change, to alter his or her mental structures in order to meet those challenges (Fowler, 1981). Piaget used the word schema to refer to anything that is generalizable and repeatable in an action (Piaget & Inhelder, 1969). As children grow and mature, these mental structures are described as organized abstract mental operations actively constructed by the children.

As their cognitive structures change, so do their adaptation techniques, and these periods of time in a child’s life are referred to as stages. The first is the sensorimotor stage of the children 2 years of age and younger (Piaget & Inhelder, 1969), an important period of time when the child is constructing all of the necessary cognitive substructures for later periods of development. These constructions, without representation or thought,
are developed through movement and perceptions. The movements and reflexes of the child in this period form habits that later form intelligence. This happens through 6 successive sub-stages: modification of reflexes, primary circular reactions, secondary circular reactions, coordination of secondary schemas, tertiary circular reactions, and invention of new means through mental combinations (Miller, 2002). During this stage, three important concepts are believed to be acquired (a) object permanence, when the child understands the object did not cease to exist just because it is hidden from view; (b) space and time, important to solving “detour” problems; (c) causality, which is when the child begins to realize cause and effect by his or her own actions and in various other objects (Piaget & Inhelder, 1969).

The second is the preoperational stage of 2- to 7-year-old children, transitions from the sensorimotor period with the development of mental representations through semiotic function, where one object stands for another (Miller, 2002). Signs and symbols are learned as similar objects and events that signify real ones. Though mental representation has advanced from its previous stage, children in this period cannot think in reversible terms (Piaget & Inhelder, 1969). Miller helps to describe other characteristics of this level, including rigidity of thought, semilogical reasoning, and limited social cognition. Rigidity of thought is best described with the example of two identical containers that have equal amounts of liquid. When the contents of a container are poured into a thinner and taller container or shorter and wider container, children at this level freeze their thought on the height and assume the volume is more or less, depending on the height of the container. The height becomes their only focus, rather than the transition of volume.
If the liquid is poured from one container into another, children focus on the states of the containers rather than the process of pouring the same amount of liquid.

Cognitively, children are unable to reverse direction of the poured liquid and imagine it being poured back into the original container and containing the same amount. They can, however, understand the identity of the liquid, that it may be poured from one container to another and still be the same kind of liquid. In this level, causal relationships are better understood outside of self, as pulling the cord more makes the curtain open more, though they may not be able to explain how it happened. Rather than thinking logically, children in this level reason semi-logically, often explaining natural events by human behavior or as tied to human activities (Miller, 2002).

Most children in ages 8 to 11 is often categorized as in the concrete operational stage in Piaget’s theory of cognitive development. According to Miller (2002, p. 52) the mental representations of children in this concrete operational period come alive with the ability to use operations, “an internalized mental action that is part of an organized structure.” In the example of the liquid in containers, children now understand the process and can reason the liquid is the same amount though in different sized containers. This ability to use operations may come at different times during this period. Concrete children begin to better understand reversibility and conservation. Classifications based on the understanding of sizes of an included class to the entire class are achieved (Piaget & Inhelder, 1969). Relations and temporal-spatial representations are additional operations evident in concrete operational children (e.g., children can understand differences in
height and length and include the earth’s surface in drawing their perception of things). All of these operations strengthen gradually over time.

The formal operational period is the fourth and final of the periods of cognitive development in Piaget’s theory (Piaget & Inhelder, 1969). This stage, which follows the Concrete Operational stage, commences at around 11 years of age and continues into adulthood. In this stage, individuals move beyond concrete experiences and begin to think abstractly, reason logically and draw conclusions from the information available, as well as apply all these processes to hypothetical situations. Rather than simply acknowledging the results of concrete operations, individuals in this final period can provide hypotheses about their relations based on logic and abstract thought. This abstract thought looks more like the scientific method than did thought in previous periods. In the concrete operational period, children could observe operations and lack the ability to explain the process. In the formal operational period, they are able to problem-solve and imagine multiple outcomes. One of Piaget’s common tasks in determining if a child has reached formal operational thought is the pendulum problem. The formal operational thinker demonstrates hypothetico-deductive thought by imagining all of the possible rates that the pendulum may oscillate, observing and keeping track of possible results, and ultimately arriving at possible conclusions (Piaget & Inhelder, 1969).

As adolescents grow into adulthood and throughout adulthood, formal operations are still developing and abstract thought is applied to more situations. Miller contends Piaget
ended his periods of developmental logical thought with formal operations. Beyond this point, individuals’ thought only change in content and stability of rather than in structure.

2.2 Existing Research and Tools on Assessment of Scientific Reasoning

In the early works on measurement of cognitive development, Piaget used multiple problems to test a child's operations of thought (Piaget & Inhelder, 1969). Miller (2002) defined Piaget's methodology as the “clinical method,” which involves a chainlike verbal interaction between the experimenter and the child. In this interaction, the experimenter asks a question or poses a problem, and the subsequent questions are then asked based on the response the child gave to the previous question. Piaget developed this interaction in order to understand the reasoning behind the children's answers.

Cook and Cook (2005) noted that through Piagetian tasks, Piaget could better understand preoperational children's thinking. He found these children showed centration, focusing on only one thing at a time rather than thinking of several aspects. This means they were centered on the static endpoints, the before and after, rather than the process. The next aspect of logical thinking noticed in Piaget's finding was preoperational children's lack of a sense of reversibility. The task of liquid conservation is simple to the logical thinking child. Water from a short and wide container is poured into a tall and skinny container. A preoperational thinker would focus only on the height of the liquid and the fact that the water was first low, then it was at a higher level in the second container; therefore, there must be more water in the second container. With a lack of a grasp for reversibility, the preoperational child does not have true operational thought to
allow him or her to imagine the pour reversed and realize the same amount of water is in both containers. The other two conservation tasks are similar to the liquid task. They each show a beginning state, a transformation, and an ending state where something has changed. The importance of children's operational and newer logical thought “is not so much that children are no longer deceived by the problem, but rather that they have now learned some basic logical rules that become evident in much of their thinking” (Lefrancois, 2001, p. 383).

Guided by Piagetian tasks, a number of researchers (Lawson, 1978a; Shayer & Adey, 1981; Tisher & Dale, 1975) have developed their own measurements in assessing students' scientific reasoning abilities, such as the Group Assessment of Logical Thinking Test (GALT) (Roadrangka, Yeany, and Padilla 1982), the Test of Logical Thinking (TOLT) (Tobin and Capie, 1981), and the Lawson's Classroom Test of Scientific Reasoning (LCTSR) (Lawson, 1978). Below, I will briefly review the three instruments and their measures. A more detailed discussion on validity of the Lawson’s test will be given in Chapter 3.

2.2.1 Group Assessment of Logical Thinking (GALT)

Roadrangka, Yeany, and Padilla (1983) compiled reliable and valid test items for the Group Assessment of Logical Thinking (GALT). In the pilot testing, Piagetian interview tasks were administered to a sub-sample of students for purposes of validation. The 21-item GALT test is given in appendix A. The first 18 items present multiple-choice problems to be answered by the individual as well as a selection of reasoning choices to
support his or her answer. The final three items are scored upon the child's inclusion of all possible answers and patterns to classify these answers.

GALT measures 6 logical operations, including conservation, correlational reasoning, proportional reasoning, controlling variables, probabilistic reasoning, and combinatorial reasoning. They also used a multiple-choice style to present answers and possible reasoning behind those answers. The GALT is sufficiently reliable and valid in its ability to distinguish between students at Piagetian stages of development. Reliability was tested by administering the GALT to students and administering Piagetian Interview Tasks to a sub-sample of those students. They found a strong correlation, \( r = .80 \) (Roadrangka et al., 1983). The question selection derived from other reliable and valid instruments helped make this a reliable and valid assessment. The Cronbach’s reliability coefficient for internal consistency of the GALT was reported as \( \alpha = .62-.70 \) (Bunce & Hutchinson, 1993).

One of the six modes measures concrete operations and the other five measure formal operations (Bunce et al., 1993). The answers to the GALT items 1 to 18 were considered correct only if the best answer and reason were both correct. For item 19, children must (1) show a pattern and (2) have no more than one error or omission, and for item 20, children must also show a pattern in answers given, having no more than two errors or omissions. To be labeled as concrete operational thinkers, the children had to score 0 to 4. Transitional thinkers was indicative of the score 5 to 7, and abstract operational thinkers would have been those children who scored 8 to 12 (Roadrangka et al., 1983).
Researchers, predominantly in the field of science education have utilized the GALT to determine a developmental level to gauge student performance, phases in the learning cycle, and cognitive/motivational characteristics. In addition, researchers have administered the GALT to determine the best method of teaching a particular subject based on the students’ logical thinking ability (Niaz & Robinson, 1992; Allard & Barman, 1994; Kang, Scharmann, Noh, & Koh, 2005). Through use of the GALT test, Allard and Barman assessed the reasoning of 48 college biology students and found 54% of these students would benefit from concrete methods of instruction. Sampling 101 more science students in a basic science course showed these researchers that 72% of these students would benefit from concrete methods rather than a traditional lecture approach in the classroom.

2.2.2 The Test of Logical Thinking (TOLT)

The Test of Logical Thinking (TOLT) is a 10-item test developed by Tobin and Capie (1981). It measures five skill dimensions of reasoning including proportional reasoning, controlling variables, probabilistic reasoning, correlational reasoning, and combinational reasoning. A high internal consistency reliability ($\alpha = 0.85$) and a reasonably strong one-factor solution obtained from factor analysis of performance on the 10 items suggested that the items were measuring a common underlying dimension. The test is included in appendix B. We can see that the items bare a lot of similarities to the ones used in GALT and the Lawson’s test, and therefore, will not be discussed in details.
2.2.3 Lawson’s Classroom Test of Scientific Reasoning (LCTSR)

Lawson (1978) originally designed his test of formal reasoning to address the need for a reliable, convenient assessment tool that would allow for diagnosis of a student’s developmental level. A valid form of measurement prior to the Lawson Test was the administration of Piagetian tasks. This method, however, is time-consuming and requires experienced interviewers, special materials, and equipment. A paper and pencil test would be more practical for classroom use, but there are also problems with this method. Paper and pencil tests require reading and writing ability, test takers have no motivation from materials or equipment to use, and it is not as relaxed as a clinical interview setting.

In the development of his test, Lawson (1978) aimed for a balance between the convenience of paper and pencil tests and the positive factors of interview tasks. He studied eighth- through tenth-grade students to determine their scientific reasoning skill level. Lawson breaks scientific reasoning into several categories: isolation and control of variables, combinatorial reasoning, correlational reasoning, probabilistic reasoning, and proportional reasoning. Test items were based on these dimensions. The original format of the test had an instructor perform a demonstration in front of a class, after which the instructor would pose a question to the entire class and the students would mark their answers in a test booklet. The booklet contained the questions followed by several answer choices. For each of the 15 test items, students had to choose the correct answer and provide a reasonable explanation in order to receive credit for that item.

To establish the validity of his test, Lawson (1978) compared test scores to responses to interview tasks, which were known to reflect the three established levels of reasoning.
He found that the majority of students were classified at the same level by both the test and interview tasks but that the classroom test may slightly underestimate student abilities. Validity was further established by referencing previous research on what the test items were supposed to measure as well as performing item analysis and principal-components analysis. The reliability of the Lawson’s test (Ver. 2000) has been evaluated by researchers who used this test. Typical internal consistency in terms of Cronbach's $\alpha$ range from 0.61 to 0.78 (Lee & She, 2010)

The popularly used version of Lawson's Classroom Test of Scientific Reasoning was released in the year 2000. It is a 24 item two-tier, multiple choice test. Treagust (1995) describes a two-tier item as a question with some possible answers followed by a second question giving possible reasons for the response to the first question. The reasoning options are based on student misconceptions that are discovered via free response tests, interviews, and the literature.

In the 2000 version, the combinational reasoning is replaced with correlation reasoning and hypothetic-deductive reasoning. The test is also converted into pure multiple choice format containing 24 items in 12 pairs. The changes of the target skill dimensions and the items are summarized in Table 2.1. With a typical two-tier structure, the first 10 pairs (items 1-20) each begins with a question for a reasoning outcome followed by a question soliciting students’ judgment on several statements of reasoning explanations. Items 21-24 are also structured in two pairs, designed to assess students’ hypothetical-deductive reasoning skills concerning unobservable entities (Lawson, 2000). Partially due to the pathways of hypothesis testing processes, these two pairs follow
different response patterns. In the item pair of 21-22, the lead question asks for selection of an experimental design suitable for testing a set of given hypothesis. The follow up question asks students to identify the data pattern that would help draw conclusion about the hypotheses. In the item pair of 23-24, both questions ask students to identify the data pattern that would support the conclusions about the given hypotheses.

The Lawson’s test is widely used in the science education community. Based on literature and through our own research we have also observed a number of issues in the Lawson’s test regarding its question design and validity. In Chapter 3, I will give a more detailed discussion on the observed issues with the Lawson’s test and possible solutions to improve the assessment. Since all the existing instruments were developed over four decades ago and many have limitations and design issues, it is then important to further develop a more up-to-date version of assessment on scientific reasoning for the education and research community. To do so, we first determine an extended set of scientific reasoning skills that need to be assessed with the new instrument. The next section gives a list of skill dimensions and brief reviews of the related research.
<table>
<thead>
<tr>
<th>Scheme Tested</th>
<th>Item Number (1978)</th>
<th>Item Number (2000)</th>
<th>Nature of Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conservation of weight</td>
<td>1</td>
<td>1, 2</td>
<td>Varying the shapes of two identical balls of clay placed on opposite ends of a balance.</td>
</tr>
<tr>
<td>Conservation of volume</td>
<td>2</td>
<td>3,4</td>
<td>Examining the displacement volumes of two cylinders of different densities.</td>
</tr>
<tr>
<td>Proportional reasoning</td>
<td>3, 4</td>
<td>5, 6, 7, 8</td>
<td>Pouring water between wide and narrow cylinders and predicting levels.</td>
</tr>
<tr>
<td>Proportional reasoning</td>
<td>5, 6</td>
<td></td>
<td>Moving weights on a beam balance and predicting equilibrium positions.</td>
</tr>
<tr>
<td>Control of variables</td>
<td>7</td>
<td>9,10</td>
<td>Designing experiments to test the influence of length of string on the period of a pendulum</td>
</tr>
<tr>
<td>Control of variables</td>
<td>8</td>
<td></td>
<td>Designing experiments to test the influence of weight of bob on the period of a pendulum</td>
</tr>
<tr>
<td>Control of variables</td>
<td>9, 10</td>
<td></td>
<td>Using a ramp and three metal spheres to examine the influences of sphere weight and release position on collisions.</td>
</tr>
<tr>
<td>Control of variables</td>
<td></td>
<td>11, 12, 13, 14</td>
<td>Using fruit flies and tubes to examine the influences of red/blue light and gravity on flies’ responses.</td>
</tr>
<tr>
<td>Combinational reasoning</td>
<td>11</td>
<td></td>
<td>Computing combinations of four switches that will turn on light.</td>
</tr>
<tr>
<td>Combinational reasoning</td>
<td>12</td>
<td></td>
<td>Listing all possible linear arrangements of four objects representing stores in a shopping center</td>
</tr>
<tr>
<td>Probability</td>
<td>13, 14, 15</td>
<td>15, 16, 17, 18</td>
<td>Predicting chances of withdrawing colored wooden blocks from a sack</td>
</tr>
<tr>
<td>Correlation reasoning</td>
<td>19, 20</td>
<td></td>
<td>Predicting whether correlation exits between the size of the mice and the color of their tails through presented data</td>
</tr>
<tr>
<td>Hypothetic-deductive reasoning</td>
<td>21, 22</td>
<td></td>
<td>Designing experiments to find out why the water rush up into the glass after the candle goes out</td>
</tr>
<tr>
<td>Hypothetic-deductive reasoning</td>
<td>23, 24</td>
<td></td>
<td>Designing experiments to find out why the red blood cells become smaller after adding a few drops of salt water</td>
</tr>
</tbody>
</table>

Table 2.1. The Comparison of Lawson’s Classroom Test of Formal Reasoning between the 1978 version and the 2000 version.
2.3 Expanding the Dimensions of Skills for Assessment of Scientific Reasoning

In our current research on assessment of scientific reasoning, we focus on a set of basic reasoning skills that are commonly needed for students to systematically conduct scientific inquiry, which includes exploring a problem, formulating and testing hypotheses, manipulating and isolating variables, and observing and evaluating the consequences. The Lawson’s Test of Scientific Reasoning (LTSR) provides a solid starting point for assessing scientific reasoning skills (Lawson, 1978, 2000). The test is designed to examine a small set of dimensions including (1) conservation of matter and volume, (2) proportional reasoning, (3) control of variables, (4) probability reasoning, (5) correlation reasoning, and (6) hypothetical-deductive reasoning. These skills are important concrete components of the broadly defined scientific reasoning ability.

To fully assess students’ ability and provide fine-tuned guidance for teachers, we have been working to expand the measurement capability of standardized assessment on scientific reasoning by incorporating sub-categories within the existing skill dimensions and new dimensions that are not included in the Lawson’s test. For example, we have developed questions on conditional probability and Bayesian statistics within the general category of probability reasoning as well as questions on an extended list of additional skill dimensions such as categorization, combinations, logical reasoning, causal reasoning, and advance hypothesis forming and testing.

In addition, for each skill dimension, multiple questions are designed using a wide variety of scientific and social contexts and with different levels of complexity, so that
we can measure students with different background and strengths from school age through college levels. These new dimensions and designs will improve the measurement capability to target students at a wider range of grade levels and backgrounds, and also provide more detailed information for researchers and teachers to address the development of scientific reasoning skills and the interactions of these skills with other aspects of learning is STEM education. Based on the literature and our own research, the following dimensions have been identified for assessment of scientific reasoning.

- Control of Variables
- Proportions and Ratios
- Probability
- Correlational Reasoning
- Deductive Reasoning
- Inductive reasoning
- Causal Reasoning
- Hypothetical-Deductive Reasoning

2.3.1 Control of Variables

In a scientific inquiry process involving many variables, the relationship between the variables needs to be determined. To do so, we form a hypothesis and test it experimentally. When designing these experiments, it is important to design controlled experiments rather than confounded experiments. This means we have to control all other variables in order to analyze the relationship between key variables without
interference. For example, when considering the relationship between age and frequency of delinquent activity, the variable of gender has to be treated as a variable to be controlled.

Control of variables is a necessary strategy in designing unconfounded experiments and in determining whether a given experiment is controlled or confounded. Control of variables strategy is used in creating and conducting experiments. Because variables interact with each other, the experimenter needs to make inferences in order to appropriately control variables and interpret the results (Chinn, C. A., & Hmelo-Silver, C. E., 2002). Usually, experimenters focus on the effect of a single variable of interest (Kuhn, D., & Dean, D. 2005).

Control of variables strategy is used in a logical sense to distinguish controlled and confounded experiments, which is necessary in determining whether an experiment can lead to a conclusive result. The logical aspects of control of variables include the ability to make appropriate inferences from the outcomes of unconfounded experiments and to understand the inherent indeterminacy of confounded experiments. In short, control of variables is the fundamental idea underlying the design of unconfounded experiments from which valid, causal, inferences can be made (Chen & Klahr, 1999).

Control of variables (COV) is a core construct supporting a wide range of higher-order scientific thinking skills and it is also an important skill fundamental to understanding physics concepts and experiments. In a recent study, Boudreaux et al. (2008) found that college students and in-service teachers had difficulties with basic methods in COV which included failure to control variables, assuming that only one
variable can influence a system’s behavior, and rejection of entire sets of data due to a few uncontrolled experiments. Boudreaux et al. concluded that students and teachers typically understand that it is important to control variables but often encounter difficulties in implementing the appropriate COV strategies to interpret experimental results.

In learning, control of variables is one of the National Research Council’s (1996) aspects of “designing and conducting a scientific investigation” (p. 145). Several sub-skills are identified in this category including “systematic observation, making accurate measurements, and identifying and controlling variables” (p. 145). Control of variables is a component of scientific inquiry, which is broadly understood to mean skill in discovering or constructing knowledge for oneself (Dean, D., & Kuhn, D. 2007). It is also one of several types of procedural knowledge – or “process skills” – that are deemed central to early science instruction (Klahr, D., & Nigam, M. 2004).

In everyday life, a real world situation is often complicated involving many different kinds of variables, therefore, when solving real problems, people need to determine which variables influence the outcome. To do so, the variable of interest is changed while other variables must be controlled. For example, if a city planner wants to find out whether temperature affects the comfort levels of a city, all variables other than temperature, such as humidity and cloud cover, must be held constant while temperature is varied.
Example of Control of Variable Question

1. This is a modified version of a question in the Lawson’s Test:

   Shown are drawings of three strings hanging from a bar. The three strings have metal weights attached to their ends. String 1 and String 3 are the same length. String 2 is shorter. A 10 unit weight is attached to the end of String 1. A 10 unit weight is also attached to the end of String 2. A 5 unit weight is attached to the end of String 3. The strings (and attached weights) can be swung back and forth, and the time it takes to make a swing can be timed.

   Suppose you want to find out whether the length of the string has an effect on the time it takes to swing back and forth. Which strings would you use to find out?

   a. only one string
   b. all three strings
   c. 2 and 3
   d. 1 and 3
   e. 1 and 2

   Answer: e

   Note: In this problem, there are two variables that may influence the time it takes to swing back and forth: the length of the string and the mass of the attached weight. Students are asked to determine the relationship between the length of the string and the time it takes to swing back and forth, so the size of the weight needs to be controlled (held constant). The weights attached to the end of string 1 and 2 are the same, but the
lengths of these two strings are different. They can be chosen to test the relationship between length and swing time.

2. This is an revised version of a question from the Lawson’s test:

Twenty fruit flies are placed in each of four glass tubes. The tubes are sealed. Tubes I and II are partially covered with black paper; Tubes III and IV are not covered. The tubes are placed as shown. Then they are exposed to red light for five minutes. The number of flies in the uncovered part of each tube is shown in the drawing. This experiment shows that flies respond to (move to or away from):

a. red light but not gravity
b. gravity but not red light
c. both red light and gravity
d. neither red light nor gravity

Answer: b

In this problem, there are two variables that would affect the distribution of the fruit flies -- gravity and red light. To investigate the relationship between red light and the distribution of fruit flies, gravity is treated as a controlled variable. Tubes II and IV are compared since gravity is not having an effect on those tubes. The two tubes have a very similar distribution of fruit flies, so we know red light did not have an impact. To test the
relationship between gravity and the distribution of fruit flies, red light is the controlled variable. By comparing Tubes I and II (or III and IV), we see that gravity does have an impact on the distribution.

2.3.2 Proportions and Ratios

In mathematics and physics, proportionality is a mathematical relation between two quantities. There are two different views of this “mathematical relation.” One is based on ratios, and the other is based on functions.

1. A Ratio Viewpoint

In many books, proportionality is expressed as an equality of two ratios:

\[
\frac{a}{b} = \frac{c}{d}
\]

Given the values of any three of the terms, it is possible to solve for the fourth term.

2. A Functional Viewpoint

Consider the following equation for gravitational force:

\[ F = G \frac{m_1 m_2}{r^2} \]

A scientist would say that the force of gravity between two masses is directly proportional to the product of the two masses and inversely proportional to the square of the distance between the two masses. From this perspective, proportionality is a functional relationship between variables in a mathematical equation.

Proportional reasoning is associated with the formal operational stage of thought, according to Piaget’s theory of intellectual development. In Research proportional
reasoning can be conceptualized in the following ways: identification of two extensive variables that are applicable to a problem and application of the given data and relationships to find (i) an additional value for one extensive variable (missing value problems) or (ii) comparison of two values of the intensive variable computed from the data (comparison problem). (Karplus et al., 1983)

In learning, proportional reasoning is recognized as a fundamental reasoning construct necessary for mathematics and science achievement (McLaughlin, 2003). In scientific inquiry, we can define useful quantities through proportional reasoning. For example, we define density, speed, and resistance with ratios. Krajcik and Haney (1987) analyzed the American Chemical Society Exam and found that over 50% of the test involved tasks requiring proportional reasoning. This implies that proportional reasoning is the primary reasoning construct required for success in chemistry, and complete development of this skill is crucial for achieving understanding of the many formal concepts associated with the content. Akatugba and Wallace (1999) contend that almost every concept in physics requires a proficient understanding of proportional reasoning, and students who are not capable of this type of reasoning will have difficulty mastering the concepts.

Proportional reasoning is considered a milestone in students’ cognitive development and is at the heart of middle grade mathematics. Proportional reasoning is associated with Piaget’s formal operational stage of thought. Many Piagetian and neo-Piagetian researchers identify the formal operational stage in subjects by having them perform tasks that require the use of ratios and proportions (Roth & Milkent, 1991).
Proportional reasoning is widely applied in everyday life. For example, gas mileage and unit price are ratios that may be grouped under the general notion of “rates” (Karplus et al. 1983).

Example of Proportion and Ratio Question

1. A fifth grade class has 18 students. At lunch time, the teacher brings in 12 bottles of orange juice, which fully fill all students’ cups (no juice is left). How many cups can be filled with 16 bottles of orange juice?
   a. 20  b. 24  c. 28
   d. 30  e. 32  f. other
   Answer: b

   The question gives that 12 bottles of juice fill 18 cups, so each cup holds \( \frac{12}{18} \) of a bottle of juice. If there are 16 bottles of orange juice, a proportional relationship can be set up: \( \frac{12}{18} = \frac{16}{x} \). Solving for \( x \) yields 24.

2. This question is selected from the Lawson’s Test. Below are drawings of a wide and a narrow cylinder. The cylinders have equally spaced marks on them. Water is poured into the wide cylinder up to the 4\(^{th}\) mark (see A). This water rises to the 6\(^{th}\) mark when poured into the narrow cylinder (see B).
Both cylinders are emptied (not shown) and water is poured into the wide cylinder up to the 6th mark. How high would this water rise if it were poured into the empty narrow cylinder?

a. to about 8  
b. to about 9  
c. to about 10  
d. to about 12  
e. none of these answers is correct

Answer: b

To solve the problem, let the height of the water in the wide cylinder be \( h_w \), the height of the water in the narrow cylinder be \( h_n \), the size of the wide cylinder be \( S_w \), and the size of the narrow cylinder be \( S_n \). The volume of water is conserved, so \( h_w S_w = h_n S_n \). This can be rearranged into a proportional relation \( \frac{h_w}{h_n} = \frac{S_n}{S_w} \). Since \( S_n \) and \( S_w \) are constants, the ratio between \( h_w \) and \( h_n \) is constant. In part A, \( h_w = 4 \) and \( h_n = 6 \). In part B, \( h_w = 6 \) and \( h_n = x \). This provides us with a new proportion: \( \frac{6}{x} = \frac{4}{6} \). Solving for \( x \) yields 9.

2.3.3 Probability

There are two main interpretations of probability, one that could be termed “objective” and the other “subjective.” A probabilistic situation is a situation in which we are interested in the fraction of the number of repetitions of a particular process that produces a particular result when repeated under identical circumstances a large number
of times. The process itself, together with noting the results, is often called an experiment. An outcome is a result of an experiment. An event is an outcome or a set of all outcomes of a designated type. An event’s probability is the fraction of the times an event will occur as the outcome of some repeatable process when that process is repeated a large number of times.

The classical interpretation of probability is a theoretical probability based on the physics of the experiment, but does not require the experiment to be performed. For example, we know that the probability of a balanced coin turning up heads is equal to 0.5 without ever performing trials of the experiment. Under the classical interpretation, the probability of an event is defined as the ratio of the number of outcomes favorable to the event divided by the total number of possible outcomes.

Sometimes a situation may be too complex to understand the physical nature of it well enough to calculate probabilities. However, by running a large number of trials and observing the outcomes, we can estimate the probability. This is the empirical probability based on long-run relative frequencies and is defined as the ratio of the number of observed outcomes favorable to the event divided by the total number of observed outcomes. The larger is the number of trials, the more is accurate the estimate of probability. If the system can be modeled by computer, then simulations can be performed in place of physical trials.

A manager frequently faces situations in which neither classical nor empirical probabilities are useful. For example, in a one-shot situation such as the launch of a unique product, the probability of success can neither be calculated nor estimated from
repeated trials. However, the manager may make an educated guess of the probability. This subjective probability can be thought of as a person’s degree of confidence that the event will occur. In absence of better information upon which to rely, subjective probability may be used to make logically consistent decisions, but the quality of those decisions depends on the accuracy of the subjective estimate.

**Example of Probability Question**

Three red square pieces of wood, four yellow square pieces, and five blue square pieces are put into a cloth bag. Four red round pieces, two yellow round pieces, and three blue round pieces are also put into the bag. All the pieces are then mixed about. Suppose someone reaches into the bag (without looking and without feeling for a particular shape piece) and pulls out one piece. What are the chances that the piece is a red round or blue round piece?

a. cannot be determined
b. 1 chance out of 3
c. 1 chance out of 21
d. 15 chances out of 21
e. 1 chance out of 2

Answer: B

2.3.4 Correlational Reasoning

In the scientific inquiry process of multi-variable contexts, some variables are independent to each other, but some are dependent. In social level, people pay much
attention to series correlation relationship. Such as the correlation between smoking and the chance to get lung cancer; drinking tea and losing weight; weather and market; the physical statures of parents and their offspring; and the correlation between the demand for a product and its price. For any two variables, they may associate with each other closely, weakly or be no dependence at all. Correlation is used to describe the degree of dependence between two variables. (There exists correlation between more than two variables, but our discussion focus on the link between two variables.)

Lawson’s Definition about Correlational Reasoning:

Correlational reasoning is defined as the thought patterns individuals use to determine the strength of mutual or reciprocal relationships between variables. Correlational reasoning is fundamental to the establishment of relationships between variables; such relationships are, in turn, basic prediction and to scientific exploration. (Anton E. Lawson, Helen Adi and Robert Karplus 1979)

Though there are multiple versions about the definition of correlation, there are two typical features when we define it:

1. When we see two variables, there are two different way to look at their relationship. One is to see if there is a link between them. The other is to see how these two variables are related, which is, in another sense, the mechanism of their relationship. Researchers studying correlational reasoning mainly focus on if people think the present data show that two variables are related or not and if people can get prediction from the data. Correlational reasoning does not require
people to see there exists certain mechanisms or causal relationship between two variables.

2. Correlational reasoning is highly related with conditional probability. That means when correlation exists between incidence A and B, the probability of A can influence the probability of B and vice versa.

Example of Correlational Reasoning Question

Brown was observing the mice that live in his field. He discovered that all of them were either fat or thin. Also, all of them had either black tails or white tails. This made him wonder if there might be a link between the size of the mice and the color of their tails. So he captured all of the mice in one part of his field and observed them. The picture shows the mice that he captured.

Based on the captured mice, do you think there is a link between the size of the mice and the color of their tails?

A. appears to be a link
B. appears not to be a link

C. cannot make a reasonable guess

Answer: A

This question is asking people to judge whether or not there exists a correlation between the size of the mice and the color of their tails. We should compare the 4 groups of mice based on their properties (fat or thin, black or white tail), which gives us the following table:

<table>
<thead>
<tr>
<th></th>
<th>Fat mouse</th>
<th>Thin mouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse with Black tail</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Mouse with white tail</td>
<td>3</td>
<td>8</td>
</tr>
</tbody>
</table>

We can see that most of the fat mice have black tails while most of the thin mice have white tails. Therefore, there exists a correlation between the size of the mouse and the color of its tail.

2.3.5 Deductive Reasoning

Deductive arguments are attempts to show that a conclusion necessarily follows from a set of premises. A deductive argument is valid if the conclusion does follow necessarily from the premises, i.e., if the conclusion must be true provided that the premises are true. A deductive argument is sound if it is valid and its premises are true. Deductive arguments are valid or invalid, sound or unsound, but are never false or true. Deductive Reasoning is a method of gaining knowledge.
Park & Han (2002) claim that deductive reasoning can be a factor which can help students recognize cognitive conflict and resolve it. They use deductive reasoning to help students to learn the direction of force acting on a moving object. For example, they show these two premises to students:

Premise 1: If an object is moving more and more slowly, then the net force acts on that object in the opposite direction to that of its motion.

Premise 2: A ball which is thrown vertically upward is moving upward more and more slowly.

They then ask what conclusion can be drawn from these premises. Their research shows that using deductive reasoning can help students change their preconceptions.

Deductive reasoning is a basic logic skill and is very useful in our daily life. We make many deductions from what we already know. For example, say you receive a flower as Christmas gift. You need to put it somewhere. You know all plants need sunshine. Your flower is plant. The flower needs sunshine, so you put it beside the window.

**Example of Deductive Reasoning Question**

You and your friends play a new card game. One side of each card shows an integer number while the other side is either white or gray. After playing for a while, one of your friends discovers that if a card shows an even number on one side, it will always be gray on the other side. Your friend lays out four cards in front of you as shown. If you want to test whether the rule your friend discovered is true or not, which cards should you turn over (choose as few cards as possible)?
a. 3 only
b. 8 only
c. 3 and white
d. 3 and gray
e. 8 and white
f. 8 and gray
g. all four cards

Answer: e

2.3.5 Inductive Reasoning

1) The basic definition of inductive reasoning

"Induction is a major kind of reasoning process in which a conclusion is drawn from particular cases. It is usually contrasted with deduction, the reasoning process in which the conclusion logically follows from the premises, and in which the conclusion has to be true if the premises are true. In inductive reasoning, on the contrary, there is no logical movement from premises to conclusion. The premises constitute good reasons for accepting the conclusion. The premises in inductive reasoning are usually based on facts or observations. There is always a possibility, though, that the premises may be true while the conclusion is false, since there is not necessarily a logical relationship between premises and conclusion." (Grolier's 1994 Multimedia Encyclopedia)
Inductive reasoning is used when generating hypotheses, formulating theories and discovering relationships, and is essential for scientific discovery.

2) The definitions of inductive reasoning in research:

Induction can be defined as the process whereby regularities or order are detected and, inversely, whereby apparent regularities, seeming generalizations, are disproved or falsified. This is achieved by finding out, for instance, that all swans observed so far are white or, on the contrary, that at least one single swan has another color. To put it more generally, one can state that the process of induction takes place by detecting commonalities through a process of comparing. However, with inductive reasoning it is not enough to compare whole objects globally to each other. Instead, they have to be compared with respect to their attributes or to the relations held in common. That is the reason why all inductive reasoning processes are processes of abstract reasoning.

Example of Inductive Reasoning Question

Question 2: What should be in ( )

A  B  C  D  E
Answer: D

From the front 5 pictures we can induce that the square spot moves clockwise and the semicircle and semi square exchange their position in turn. So the [?] should be D.

2.3.6 Causal Reasoning

Causal reasoning concerns with establishing the presence of causal relationships among events. When causal relationships exists, we have good reason to believe that events of one sort (the causes) are systematically related to events of some other sort (the effects), it may become possible for us to alter our environment by producing (or by preventing) the occurrence of certain kinds of events.

Most studies of students’ ability to coordinate theory and evidence focus on what is best described as inductive causal inference (i.e., given a pattern of evidence, what inferences can be drawn?).

If there are causal relationship between variable x and y, there are several kinds of causes:

1) Necessary causes: If x is a necessary cause of y, then the presence of y necessarily implies the presence of x with the probability of 100%. The presence of x, however, does not imply that y will occur.

2) Sufficient causes: If x is a sufficient cause of y, then the presence of x necessarily implies the presence of y with the probability of 100%. However, another cause z may alternatively cause y. Thus the presence of y does not imply the presence of
x. For instance, in the case of ‘losing the breath means the death of a person’, losing the breath is a sufficient cause of death of a person, but death of a person is a necessary cause of losing the breath.

3) Contributory causes: If x is a contributory cause of y, it means the presence of x makes possible the presence of y, but not with the probability of 100%. In other words, a contributory cause may be neither necessary nor sufficient but it must be contributory.

For instance, in the case of ‘having the cancer causes the death of a person’, ‘having the cancer’ is a contributory causes. It is neither a necessary nor sufficient cause of the death of a person. Because, firstly, having cancer is not sufficiently cause a person to die (some cancer can be treated); secondly, the death of the person necessarily caused by cancer (some other factor such as car accident, or suicide also may cause a person to die. But everyone can’t deny that if a person has cancer, it will probably lead to his death.

**Example of Casual Reasoning Question**

A zoologist travels to Africa to study the natural breeding environment of giraffes. While there, he notices a type of tall tree that produces a special fruit that only grows at the top of the tree. He also notices that giraffes that frequently eat this fruit appear to be stronger and taller than those who cannot reach the fruit. He concludes that the fruit contains rich nutrients which make the giraffes that eat the fruit grow stronger and taller.

Which one of the following statements do you agree with?
a. When a giraffe frequently eats this special fruit, it grows stronger and taller.

b. The nutrients in the fruit can help the giraffe grow stronger and taller.

c. Both A and B are correct.

d. The result is not sufficient to demonstrate that eating the fruit causes a giraffe to grow stronger and taller.

e. None of the above statements is reasonable.

Answer: D

In this question, the zoologist observes a positive correlation between the frequency of eating fruit from a tall tree and the height of the giraffes. He then concludes that eating the fruit cause the giraffes to be taller. This is a typical “correlation implies causation” fallacy. It is quite possible that tall giraffes eat from the tall trees just because they are tall. We cannot make any conclusion about whether or not the fruit makes the giraffe taller.

2.3.7 Hypothetical-deductive reasoning

Hypothetical-deductive method (HD method) is a very important method for testing theories or hypotheses. The HD method is one of the most basic methods common to all scientific disciplines including biology, physics, and chemistry. Its application can be divided into five stages:

1) Form many hypotheses and evaluate each hypothesis

2) Select a hypothesis to be tested

3) Generate predications from the hypothesis
4) Use experiments to check whether predictions are correct

5) If the predictions are correct, then the hypothesis is confirmed. If not, the hypothesis is disconfirmed.

HD reasoning involves starting with a general theory of all possible factors that might affect an outcome and forming a hypothesis; then deductions are made from that hypothesis to predict what might happen in an experiment.

In scientific inquiry, HD reasoning is very important because, in order to solve science problems, you need to make hypotheses. Many hypotheses can't be tested directly; you have to deduce from a hypothesis and make predictions which can be tested through experiments.

According to Piaget’s theory of intellectual development, HD reasoning appears in the formal operational stage (Inhelder & Piaget, 1958). Lawson et al. (2000) claim that there are two general developmentally-based levels of hypothesis-testing skill. The first level involves skills associated with testing hypotheses about observable causal agents; the second involves testing hypotheses about unobservable entities. The ability to test alternative explanations involving unseen theoretical entities is a fifth stage of intellectual development that goes beyond Piaget’s four stages.

Example of Hypothetical-deductive Reasoning Question

A student put a drop of blood on a microscope slide and then looked at the blood under a

- Magnified Red Blood Cells
- After Adding Salt Water
microscope. As you can see in the diagram below, the magnified red blood cells look like little round balls. After adding a few drops of salt water to the drop of blood, the student noticed that the cells appeared to become smaller.

This observation raises an interesting question: Why do the red blood cells appear smaller? Here are two possible explanations:

I. *Salt ions (Na\(^+\) and Cl\(^-\)) push on the cell membranes and make the cells appear smaller.*

II. *Water molecules are attracted to the salt ions so the water molecules move out of the cells and leave the cells smaller.*

To test these explanations, the student used some salt water, a very accurate weighing device, and some water-filled plastic bags, and assumed the plastic behaves just like red-blood-cell membranes. The experiment involved carefully weighing a water-filled bag in a salt solution for ten minutes and then reweighing the bag.

What result of the experiment would best show that explanation I is probably wrong?

*A. the bag loses weight*

*B. the bag weighs the same*

*C. the bag appears smaller*

Answer: *A*

What result of the experiment would best show that explanation II is probably wrong?

*A. the bag loses weight*

*B. the bag weighs the same*

*C. the bag appears smaller*
Answer: B

This question gives students two alternative hypotheses and the experiment to test these two hypotheses. Students need to make a prediction about the results of the experiment according to each hypothesis and consider what result could confirm or disconfirm the hypothesis.

If hypothesis I is right, then the weight of the bag won't change because there are no molecules or ions coming into or going out of the bag. If hypothesis II is right, the bag will lose weight because water molecules move out of the bag. From HD reasoning, we know the answers are A and B.
Chapter 3. Validity Evaluation of the Lawson’s Classroom Test of Scientific Reasoning

3.1 A Historical Review on the Development of the Lawson’s Test

Currently, there are increasing interests and activities among many STEM education communities in research on students’ ability in scientific reasoning, in which the Lawson’s test is often used as the tool for assessing scientific reasoning. However, research to critically inspect the existing assessment instruments and to develop new instruments for scientific reasoning is largely missing in the literature.

Regarding the Lawson test, for which there are some studies (Stefanich et al., 1983; Pratt & Hacker, 1984) investigating the validity of the Lawson's 1978 version of the formal reasoning test, there is little work on its 2000 version. Even though the 2000 edition has become a standard assessment tool in physics education research, the test itself has not been systematically validated. Through research, we have also observed several issues concerning the question designs and data interpretations. In this Chapter, we first consider the inception and early examination of the validity of the Lawson test as was originally created in 1978. We then report our research on the validity of the 2000 version of the Lawson test based on large-scale assessment data and follow-up interviews amongst a large swath of education levels. Our findings show both the validity and predicaments with the current examination design for measuring formal reasoning skills.
During the 1960s and 70s, various researchers wanted to create an examining tool to uncover the level of understanding and formal reasoning students possessed. The impetuous came from the fact that the most accurate and informative tool was through clinical-style interviews, but this tool was extremely demanding. It required both the aptitude in implementation as well as significant amounts of time; it may also require equipment for demonstrations to help probe the minds of their subjects. A more effective method was needed so that teachers could accurately determine the standing of the students in their classes. That method should be something that could be administered by any trained instructor and scored in an objective manner.

Most tests developed were paper-and-pencil methods which could more easily be graded (Longeot, 1965; Raven, 1973; Burney, 1974; Tisher & Dale, 1975; Tomlishen-Keasey, 1975; Tobin & Capie, 1980). Other tests may require equipment for the students to work with and pamphlets for them to fill out (Rowell & Hoffman, 1975), but these sorts of tests were time-consuming and needed all the said instruments; such issues made the implementation of this type of examination more restricted and tended to have smaller sample sizes. Another version of this sort of exam had an instructor conduct a demonstration for the class, which minimized the time and equipment requirements (Shayer & Wharry, 1975). This helped strike a decent balance between the power of interviews with a format that could be more readily implemented.

The test of interest that came out of this person was Lawson (1978a) which had been developed out of Piagetian methods and questions of formal reasoning based on Piaget’s developmental theory (Piaget, 1965). Lawson argued that most pencil-and-paper tests
could tend to examine reading and writing skills more than formal reasoning abilities, and he found the questions of Shayer and Wharry (1975) to be insufficient in variety. The test that he developed utilized questions that others had created in different contexts to test formal reasoning for a variety of operations, including control of variables, combinatorial reasoning, correlational reasoning, probabilistic reasoning, and proportional reasoning.

In his test design, Lawson had fifteen items all of which had a demonstration conducted by the instructor to help pose a question to the students. Each pupil had their own test pamphlet which had two parts to each testing time: a multiple-choice question for the correct prediction to the posed question, and a written section for the student to explain their answer. An item was scored as correct only if the prediction was correct and their reasoning satisfactory. Another version of the test was created as well that had only 10 items instead (Lawson, 1978b).

To test his investigative tool, Lawson administered the exam to 513 students from eighth to tenth grade of which 72 were randomly picked for a battery of Piagetian tasks in a clinical interview. The comparison between interview data and the test results would indicate how well the two are correlated. Lawson also had several judges assess the quality of his test questions by those with expertise in Piagetian research; all six judges agreed that the questions tested both concrete and/or formal reasoning. For comparing the interview data and test data, two statistical tools were employed: parametric statistics and principle components analysis. The former tool found an overall correlation of 0.76 that was statistically significant (p< 0.001). The later tool indicated that 66% of the variance in scores would be accounted for if there were three principle factors that were measured
by this examination; with the small number of students, this result was considered tentative. Overall, the assessments indicated that the Lawson test was able to measure formal reasoning and to correlate reasonably well with clinical methods.

Lawson (1978a) also created a ranking system to help instructors understand what a score on this test would mean. From his results of over 500 students, Lawson created three classes of reasoning subjects: concrete, transitional, and formal. The first would have a score from 0-5, the second 6-11, and the last 12-15. About one third (35.3%) of those examined were classified at the concrete level, about half (49.5%) transitional, and the remaining (15.2%) at formal. When compared to interview assessments, more than half of the 72 interviewees fit into the reasoning levels developed by Lawson. Of those that did not fit into the schema, the data indicated that the Lawson test may have underestimated those students.

After publication, the Lawson test began to be used in classrooms, and not long after other researchers began to assess how well the test was performing. For example, Stephanich et al. (1983) tested the correlation between clinical interviews and students taking various assessment tests including that of Lawson. A weaker correlation than reported by Lawson was found and the test was said to overestimate reasoning abilities rather than underestimate. However, this examination had a small sample size (N=27) and no estimate of statistical significance was provided, so it cannot overturn what was found in the previous study. Nonetheless, an appreciable correlation was found (0.50) between interviews and the pencil-and-paper examination, so we can say that the test from 1978 seems valid.
Another study comes from Pratt and Hacker (1984), who implemented the exam on 150 students to uncover if the Lawson test measured one or several factors. Taking issue with the factor analysis of Lawson (1978a), the researchers used another model to indicate that indeed the Lawson test measurement was multi-factorial rather than singular (this was even more strongly indicated in the test from Lawson [1978b]), which they took to be as a weakness of the test. This sort of examination was repeated by Hacker (1989) who found the same results: Lawson’s test is multi-factorial. Other researchers (Harrison, 1986; Reckase, Ackerman & Carlson, 1988) do not find a multi-factorial examination to be problematic, especially if formal reasoning is multifaceted, so we take Pratt and Hacker (1984) to prove a point that Lawson (1978a) was uncertain of in his original article.

Later an Israeli study conducted the Lawson test to find how strongly these test scores correlated to success in the sciences and mathematics (Hofstein & Mandler, 1985). In their results, it was found that formal reasoning students outperformed transitional and concrete reasoning students, though the latter two levels of ability were not distinguishable. Concerning performance in STEM, only one of the items (probability reasoning) was found to be predictive in all analyzed sciences, but overall the test was a good indicator for the success of biology students alone. This limitation thus indicates that this formal reasoning test is not sufficient in determining success in STEM overall, and perhaps no formal reasoning test can achieve this result.

After more than twenty years since its original edition, Lawson in 2000 produced a new version of his examination, but this time it was completely multiple-choice and
without demonstrations. Also unlike before where there were fifteen items, the 2000 test had twelve items in question pairs, making twenty-four total questions. Another change was that a score was a simple count of the number of right answers rather than the number of items where both questions had to be correctly answered. Combinational reasoning items were also replaced by more correlational reasoning and hypothetical-deductive reasoning items. A complete list of changes is given in Table 2.1 of Chapter 2.

The first ten items of the 2000 version in its two-question format had, as originally developed, one question for correctly predicting the outcome to some particular situation, and the second question was to find the correct reasoning behind that selection in the first question. The last two items then introduced hypothetical-deductive questions. The first question of item 11 concerned the experimental design, and the second question asked for what outcome would support a stated hypothesis. Item 12 was similar to the prior item, but both questions concerned the data pattern that would support a stated hypothesis.

The utility of the new version of the Lawson test is most obvious in that a completely multiple-choice exam is more quickly and objectively scored, and thus it becomes better accommodated to use by instructors to gauge the reasoning abilities of their pupils. However, the 2000 Lawson test was not presented in a formal study proving its efficacy, instead resting on the laurels of its earlier incarnation. In one study since its distribution, it was shown that there was a correlation between increases in scores on the Force and Concept Inventory (FCI) and Conceptual Survey of Electricity and Magnetism (SCEM) and Lawson test scores among the community college students (with and without calculus) to whom the various tests were administered (Diff & Tache, 2007). The
correlations were small but significant nonetheless, and similar findings have also been reported (Dubson & Pollock, 2006; Coletta & Phillips, 2005 & 2007; Coletta et al., 2008), so the new Lawson test still appears to be a useful measure of formal reasoning.

Nonetheless, a proper analysis of the Lawson 2000 exam had not been done. Moreover, no investigation has analyzed the test across a large swath of education levels. This renders the current study necessary in determining what weaknesses it may have and where problems may develop.

3.2 Content Evaluation of Lawon’s Test – Item Context Issues

To lead the validity evaluation of the Lawon’s test, a content analysis of some the questions provides useful insights on how experts or examinees may disagree over information included in the questions. The Lawon’s test items were designed with simple scientific context scenarios that may show up as examples in K-12 science courses, such as pendulums, graduated cylinders, candle lighting experiment, etc. Although simple, these contexts do stand out with a scientific lab flavor that may intimidate certain students who are weak in science.

In addition, since these contexts may be used in courses which are hard to control, there is a potential complication of content interference. For example, Q21 to Q24 on the Lawon’s test assess students’ hypothetical deductive reasoning ability (i.e., the ability on hypothesis forming and testing). The contexts used are candle lighting (oxygen and carbon dioxide experiment) and cells in salty water (osmosis experiment). Through interviews, we have observed that a significant number of high school and college
students reporting that they responded to the questions by recalling the exact experiments that they have done or observed before and didn’t try to reason through the problem. Therefore, to these students, the questions become a content test rather than a test on reasoning.

Practically speaking, the content interference cannot be totally removed, however, in order to correctly interpret assessment results, systematic research with detailed interviews is needed to provide valid information on the extent to which the content interference may have on students of different age groups and backgrounds.

3.3 A Data-Driven Study on the Validity of the Lawson’s Test

The sections below take the data-driven approach to evaluate the validity of the Lawson’s tests based on students’ quantitative and qualitative responses. For this study with the 2000 version of the Lawson test, data was collected in three forms: (1) large scale quantitative data with students from 3rd grade to graduate level; (2) an added short answer free-response explanation to each item of the Lawson test given to college students in their freshmen year; (3) think-aloud interviews with freshman college students. The first form provides our quantitative data which can indicate particular issues with questions that need to be assessed by other means. The second form provides both qualitative and quantitative data which can more clearly indicate reasoning difficulties or testing worries. The third form extracts detailed information about students’ thought processes in answering questions. There was additionally, though on a smaller scale, the collection of eye-tracking data for students that took the exam with
computers rather than pencil and paper. This information could help indicate what sorts of questions may make students more hesitant in answering.

The student populations of the data collection include Chinese students from the 3rd grade to graduate level (N=7131) and U.S. students from the 5th grade to first year of college (N=2777). The Chinese grade school students (N=6258) are from 141 classes in 20 schools from 8 regions around China. The students in China used a version translated by physicists fluent in both languages. The translated versions were also piloted with a small group of undergraduate and graduate students (N~20) to remove language issues.

The U.S. grade school data was collected in several mid-western states from 30 classes of students across 14 private and public schools (N=1078). The schools were selected from a wide range of regional and demographical backgrounds to obtain a representative pool of populations. The college student data were from four U.S. universities (N=1699) and three Chinese universities (N=873). The students tested were first year science and engineering majors enrolled in calculus-based introductory physics courses (N_{China}=458, N_{US}=1370). The tests were administered before any college level instruction of the relevant content topics.

The four U.S. universities are ranked and their backgrounds are given below (based on 2007 U.S. News and World Report Ranking):

- **U1** is a large research-1 state university, U.S. ranking top 60, acceptance rate 59%.
- **U2** is a large research-1 state university, U.S. ranking top 60, acceptance rate 60%.
• U3 is a large tier-4 state university with an acceptance rate of 84%.
• U4 is a large tier-3 state university with an acceptance rate of 69%.

The three Chinese national universities – schools directly funded and supervised by the nation’s department of education – with their national rankings are also given below (based on 2007 ranking from Sina Education News, http://edu.sina.com.cn):
• C1 is a top 30 national university.
• C2 is a top 60 national university.
• C3 is a top 130 national university.

In the selected universities, we targeted those with medium ranking in order to make a more representative pool of the populations. For college students, the Lawson test was administered in the beginning of the fall semester (or quarter) before any college level instructions. For grade school students, the test was administered at times throughout the fall semester (or quarter).

3.4 Quantitative Results – Item Score Analysis

To search for questions with potential validity issues, the average scores of each test item were computed for students at selected grade levels for baseline comparisons of students at various developmental stages. These results are seen in Figure 3.1 which gives the average item scores of U.S. and Chinese college students. In general, the two populations have systematic differences on items measuring different skill dimensions. The Chinese students are higher on proportional reasoning (items 5-8) and hypothetical deductive reasoning (items 21-24) but are lower on probabilistic (items 15-18) and
correlational reasoning (items 19-20). Due to the large sample sizes, the differences between US and Chinese students are statistically significant and the error bars based on standard errors are small and ignored in the plots.

![Figure 3.1. Average scores of the Lawson test items from Chinese (N=248) and U.S. (N=646) college students.](image)

The items scores show large “dips” (divergences from items in the same skill dimension) on item 8 for the Chinese students, item 12 for both U.S. and Chinese students, and item 14 for U.S. students. As these results come from those with a high
level of education, this pleaded for an investigation of how apparent the dips are at earlier stages of education. To inspect if the dips are unique to the college population, we plotted the item scores of three selected age groups of Chinese students in Figure 3.2.

![Figure 3.2. Average scores of the Lawson test items from Chinese students in 7th, 10th, and 12th grades (N7=529, N10=1195, N12=458).](image)

The results suggest that the dip on item 12 is formed early with students in 7th grade as compared to the question in the same category. This dip remains largely unchanged all the way through college students except to become more pronounced. Among the
Chinese students, the dips on items 8 and 14 start to appear at the 10\textsuperscript{th} grade and become more obvious through the 12\textsuperscript{th} grade. These results indicate that the dips are consistently observed in the developmental trend of students from young to senior ages.

All the students’ performances in the test were recorded in Figure 3.1. Using this graph, it can be observed that there are a few items showing abnormal results. On the item pair of Q11 & Q12 at least 80\% of both U.S. college students and Chinese students scored correctly on Q11, but only around a third of each nation’s students correctly answered Q12. Similar situations occurred in other item pairs such as Q7 & Q8, and Q13 & Q14. Such artifacts would suggest that there is no progression from novice to expert, but this runs contrary both to expectations and to the progress seen in the question pairs where one answer significantly improves while the other does not as seen in Figure 3.2. Moreover, it was the question pairs Q11 & Q12 and Q13 & Q14 that were added to the 1978 version of the Lawson test and not analyzed as his prior questions were. This phenomenon led us to investigate the design of the question and discover if it is the wording used which confuses students. To answer this question and verify the validity of the test, both quantitative and qualitative data were applied in this research.

3.5 Quantitative Results – Analysis of Two-Tier Score Patterns

The Lawson 2000 test version is a two-tier test. Usually educators and researchers (Lee & She, 2009; Coletta & Phillips, 2005, 2007; Hofstein & Mandler, 1985; Lawson, 1978a) grade the test using the pair-scoring schema which means items are scored as being correct (a score of 1) if students choose the right answers for both the question and
the proper corresponding reason. The investigation into this two-tier test method is important. For example, there are a large number of students that chose the right answer for the wrong reason, or vice versa. Under the two-tier scoring system, these students are all categorized as not knowing how to solve the problem. However, it is possible that student may have the knowledge and reasoning skills in the domain at some level but were misguided by question design issues, which lead them to answer incorrectly in the content or in the reasoning part. Therefore, to truly test students’ scientific reasoning abilities and verify the rationality of the item design, a careful study of re-analyzing data using the single-scoring schema was utilized.

<table>
<thead>
<tr>
<th>Response Patterns</th>
<th>1-2</th>
<th>3-4</th>
<th>5-6</th>
<th>7-8</th>
<th>9-10</th>
<th>11-12</th>
<th>13-14</th>
<th>15-16</th>
<th>17-18</th>
<th>19-20</th>
<th>21-22</th>
<th>23-24</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>2.9</td>
<td>20.7</td>
<td>31.0</td>
<td>23.2</td>
<td>18.9</td>
<td>41.5</td>
<td>25.0</td>
<td>7.3</td>
<td>12.9</td>
<td>22.2</td>
<td>37.0</td>
<td>19.3</td>
</tr>
<tr>
<td>11</td>
<td>94.2</td>
<td>77.8</td>
<td>60.1</td>
<td>43.1</td>
<td>76.3</td>
<td>24.0</td>
<td>35.8</td>
<td>81.1</td>
<td>80.4</td>
<td>64.4</td>
<td>30.7</td>
<td>37.0</td>
</tr>
<tr>
<td>01</td>
<td>1.3</td>
<td>0.4</td>
<td>4.5</td>
<td>27.1</td>
<td>2.0</td>
<td>7.1</td>
<td>15.9</td>
<td>10.3</td>
<td>2.9</td>
<td>3.4</td>
<td>20.7</td>
<td>35.3</td>
</tr>
<tr>
<td>10</td>
<td>1.6</td>
<td>1.1</td>
<td>4.4</td>
<td>6.6</td>
<td>2.8</td>
<td>27.4</td>
<td>23.4</td>
<td>1.3</td>
<td>3.8</td>
<td>10.0</td>
<td>11.7</td>
<td>8.4</td>
</tr>
<tr>
<td>Sum of 01 &amp; 10</td>
<td>2.9</td>
<td>1.5</td>
<td>8.9</td>
<td>33.7</td>
<td>4.8</td>
<td>34.5</td>
<td>39.3</td>
<td>11.6</td>
<td>6.7</td>
<td>13.4</td>
<td>32.4</td>
<td>43.7</td>
</tr>
</tbody>
</table>

Table 3.1. Lawson Test two-tiered response patterns of U.S. college freshmen (N=1699).

The results of this research are summarized in Table 3.1. There are four possible patterns. The (0,0) pattern means students incorrectly responded to both questions; the (1,1) pattern means they responded correctly on both questions; the (0,1) pattern means
they responded incorrectly on the content question, but had reasoning that should lead them to the correct answer; and the (1,0) pattern means they responded correctly on the content question, but with incorrect reasoning.

<table>
<thead>
<tr>
<th>Response Patterns</th>
<th>Population Percentage</th>
<th>Paired Score</th>
<th>Single Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>21.82</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>11</td>
<td>58.73</td>
<td>58.73</td>
<td>58.73</td>
</tr>
<tr>
<td>01</td>
<td>10.91</td>
<td>0.00</td>
<td>5.45*</td>
</tr>
<tr>
<td>10</td>
<td>8.54</td>
<td>0.00</td>
<td>4.27*</td>
</tr>
<tr>
<td><strong>Total Score</strong></td>
<td><strong>58.73</strong></td>
<td><strong>68.47</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2. Comparison of Lawson test total percentage scores of U.S. college freshmen (N=1699) calculated with paired-score vs. single-question-score methods. *For 10 and 01 patterns, their contributions to total scores are weighted at 0.5.

The (0,1) and (1,0) pattern was summed and the percentage of students choosing one correct answer in a question pair was found. Attaining the right answer for the wrong reason or the wrong answer for the right reason would imply that there may be a problem in the way the question was asked. The response patterns for all 12 pairs of questions are giving in Table 3.1. Table 3.2 gives the sum of all four patterns for the complete Lawson’s Test.
From Table 3.1, the subtotal of (0,1) and (1,0) pattern of most questions are relevantly low, which means both question and reasoning parts are rather consistent question pairs. However, there are several high concentrations of this pattern in such columns as Q7 & Q8, Q11 & Q12, and Q13 & Q14, which have been highlighted in the yellow color (as mentioned in the introduction to this paper, Q21 through Q24 no longer followed the pattern of question and then explanation, so they have been omitted from this study, even though they contain a high sum of the 01 and 10 pattern.) Remarkably, these inconsistent cases are matched with the problematic items presented earlier. This result further verifies our above hypothesis that there may be some discrepancies in question design.

3.6 Qualitative Results – Analysis of Student Interviews

To further validate hypothesis that the high percentage of (0,1) and (1,0) answers is caused by question design, we give the Lawson’s test to 181 college freshman science and engineering majors and asked students to provide open-ended reports on their reasoning to each of the questions in the Lawson’s test. We also conducted follow-up interviews with a subset of these students (N=66) asking them to go over the test after completing it and explaining their reasoning on how they solved each of the questions. These students are from the same pool of college freshmen, with whom the quantitative data were collected. In this section, the three question pairs of interests are discussed in detail.
Q7 & Q8 Pair on Proportional Reasoning

7. Water is now poured into the narrow cylinder (described in Item 5 above) up to the 11th mark. How high would this water rise if it were poured into the empty wide cylinder?
   a. to about 7 1/2
   b. to about 9
   c. to about 8
   d. to about 7 1/3
   e. none of these answers is correct

8. Because
   a. the ratios must stay the same.
   b. one must actually pour the water and observe to find out.
   c. the answer can not be determined with the information given.
   d. it was 2 less before so it will be 2 less again.
   e. you subtract 2 from the wide for every 3 from the narrow.

From the results in Table 3.1, almost one third of students (27.1%) gave the incorrect answer to the concrete question, but selected the correct reason. Conversely, 6.6% of students answered correctly for the incorrect reason. More specifically, almost 11% of the answers for Q8 were choice “e” even though the correct answer was “a”.

In order to further understand how students respond to this question pair, we asked students to provide open-ended reports on their reasoning and also conducted interviews
with a subgroup of these students. Below are two cases selected from the open-ended reasoning.

Case 1:

7. 1) Answer: d

2) Explanation: \( \frac{x}{11} \times \frac{2}{3} \Rightarrow 3x = 22 \Rightarrow x = \frac{22}{3} \text{ or } 7\frac{1}{3} \)

8. 1) Answer: e

2) Explanation: Proportions

Here we can see that the student selected the correct answer on the question for the wrong reason. Even with this student’s obvious math competence, they picked “e” instead of “a” as the answer for the reasoning part.

In the follow-up interviews, there were a couple of students (5 out of 21) whose answers mirrored the logic of the student in Sample 1. They could correctly solve the problem using knowledge of ratios and mathematics for Q7, but they all picked “e”, an incorrect choice for reasoning defined in the answer key of the Lawson’s Test. According to the interviewed students, this was mainly because they thought “e” was a more detailed description of the ratio, which is specific to the question, whereas “a” is just a
vague generally true statement. Therefore, these students chose the answer that they took
to be more descriptive out of two choices that appeared to both be correct.

In addition, during the administration of the test, quite a few students would raise
their hands to ask the differences between the choices “a” and “e”. They believed these
two choices were fundamentally equivalent. This further reinforces our contention that
the wording leads to perplexity and can upset an accurate measure of their reasoning
abilities.

Case 2:
7. Answer B
Explanation: \[ \frac{y}{6} = \frac{6}{x} \rightarrow yx = 36 \rightarrow x = 9, \text{just use ratios} \]
8. Answer A
Explanation: use ratios

In this case, by contrast, the student failed in obtaining the correct ratio the question
7, but picked the correct reason in question 8. The explanation is rather simplistic by
restating that the question is about the ratio. Apparently, the student was not fluent in
properly using a ratio to solve this problem, but knew the idea to “use ratios”. Therefore,
an inference can be made that the student had a general idea the answer would be related
with the concept of ratios, which led them to pick “a” – the vague true statement about
ratio. This student was able to describe the method to solve the problem, but did not
actually know how to do the operations. This indicates that the design of this question is
problematic to those that are competent in the required skills but more forgiving to those that are less competent, leading to assessment uncertainties in terms of both false positives and false negatives.

Through the above analysis of this question pair, it is obvious that the wording is problematic which has adverse impact on the validity of the assessment.

Q11 & Q12 and Q13 & Q14 Pairs on Correlation Reasoning

11. This experiment shows that flies respond to (respond means move to or away from):
   a. red light but not gravity
   b. gravity but not red light
   c. both red light and gravity
   d. neither red light nor gravity

12. because
   a. most flies are in the upper end of Tube III but spread about evenly in Tube II
   b. most flies did not go to the bottom of Tubes I and III.
   c. the flies need light to see and must fly against gravity.
   d. the majority of flies are in the upper ends and in the lighted ends of the tubes.
   e. some flies are in both ends of each tube.
13. In a second experiment, a different kind of fly and blue light was used. The results are shown in the drawing.

   a. blue light but not gravity
   b. gravity but not blue light
   c. both blue light and gravity
   d. neither blue light nor gravity

14. *because*

   a. some flies are in both ends of each tube.
   b. the flies need light to see and must fly against gravity.
   c. the flies are spread about evenly in Tube IV and in the upper end of tube III.
   d. most flies are in the lighted end of Tube II but do not go down in Tubes I and III.
   e. some flies are in both ends of each tube

The assessment of results in Table 1 indicated that these two question pairs were rather related. They were both highly concentrated on the (0,1) and (1,0) patterns, and they all had almost one-third students (27.1% and 23.4%, respectively) choosing the correct answer for the wrong reason. To further understand student reasoning on these question pairs, we analyze qualitative data from open-ended surveys and interviews.
Case 1 (Q11 & Q12 pair):

In this case, the student correctly noted that the flies tend to move to the top of the tubes when they were upright (I and III), but did not realize that one needed to use tubes II (level and with dark paper over one end to block light) and III (upright and without paper) to test both hypotheses – gravity and red light. It was these tubes that needed to be used to act as controls since tube II is level and independent of a gravitational effect. Tube III had no paper so it only tested the effects of gravity when compared with tube IV.

Case 2 (Q11 & Q12 pair):
In this case, the student answered the first question of the pair incorrectly, but picked the correct reason in the second question. We can also see a contradictory situation here: in the first question the student’s reasoning and explanations contradict his/her answer; the reasoning stated that the black paper had no significant effect, however, the answer picked (A) clearly stated that light did have an effect. The more curious situation is the reasoning portion where the student gave the correct answer (choice A) with the reasoning that “it represents the results”. This is true, but it is also true for answer “b”. Moreover, the student also wanted to compare tubes (tube I) that were not needed for controlling variables (so did the student in case 1 student). In addition, the poor graphical representation of the original Lawson’s test was also criticized by students.

Here we can see that the reasoning statements in question 12 need improvement. The current statements give some kind of incomplete descriptions which leave the students with a sense of uncertainty and their answers are usually based on guessing of the one that reads best. More clear and explicit comparison of results among different tubes need to be included in the choices so that the sense of controlling variables with different conditions can then be clearly manifested. Or the reasoning will still be implicit at an intuitive gut feeling level without formally recognized by students’ thought processes.

In interviews, similar situations were also found. For example, one student wanted the answer for Q12 to be a combination of “c” and “e”, though the former choice expresses the student’s belief of how flies ought to respond rather than something consistent with the data from the experiment. In follow-up questions, that student indicated they compared I to III and II to IV.
From this data, it appears that students did compare multiple tubes, rather than looking at a single tube, to test a hypothesis. Perhaps this tendency exists because the tubes they compare have only one change between them; tubes I and III are both upright and their only difference is the black paper on the bottom section of the tube or the lack thereof. Since it would seem a single variable has changed between them, then COV skills would suggest they are appropriate for deriving valid conclusions. Similarly, tubes II and IV are both level and only differ with the use of black paper over one end. The student in case 2 also indicated that a comparison between tubes I and IV and between II and IV could be of use, further validating this hypothesis. The problem is that tube I is frequently used by students in making comparisons, which lead us to believe that at certain developmental level, students tend to build their reasoning on situations that all variables are varied (tube I), which is in fact a confounded condition not useful for hypothesis testing. This further indicates that students at this level lack the formal reasoning on the meaning and utility of COV, but they started to understand the needs for co-variation of variables. It can be a possible intermediate level in the developmental progression of COV reasoning.

The question pair of 13 and 14 has a lot of similar issues to that of 11 and 12. For example, the correct answer (d) to Q14 included examining tube I even though it was not a proper implementation of COV. This might have led to the results for Q14, which is very similar to Q12, to be answered correctly more often than Q12.

In summary, based on the qualitative and quantitative results, we have reason to believe that the choices of the reasoning portion of the fly-COV questions (Q11-Q14) can
cause significant uncertainties among students and need to be reworked. In addition, the graphical representations of the questions also need to be improved. These issues are carefully revised in the similar questions used in iSTAR.

3.7 Consideration on Two-Tier Question Structure

While we have so far concentrated on the problems of incorrectly answering a question due to issues of representation and wording of the question, there is also the concern that the format of the test may allow students to answer a question correctly without real understanding of what the Lawson is meant to measure. Because of the two-question format of the exam, students can identify matched pairs of answers for the two questions using simple logic. Therefore, students can use certain cues in the wording or choices of one question pair to help answer the other.

This hypothesis was further supported in interviews with some of the students. For example, about 10% (6 out of 66) changed their answer to Q11 after they had read Q12. Apparently, something in their reading of the latter question helped them realize some issue with the prior answer. About an equal number of students (5 out of 66) changed their answers to both Q11 & Q12 after reading the Q13 & Q14 pair which is very similar in design. This again indicates that something in the questions cues the test subject into finding the answer to other questions. In the last question pair (Q23 & Q24), some interviewees did not understand the first question until reading the second, which further indicates that some (but not all) students use information from the pair of questions to help them understand and answer the questions.
To avoid this problem, one way is to eliminate the two-tier structure by either combining the two questions into a single question that addresses both content and reasoning or retaining only one part of the pair and converting the other into as part of the question in terms of propositions, presumptions, conditions, or extended information that students would need to use or consider in answering or explaining the question. The other way is to retain the use of the two-tier structure but design better choices for both questions so that multiple combination of two choices can be the correct answer pairs – this will significantly decrease the success rate of simple ruling-out strategy that students often use in solving MC questions. However, optimal design of such questions and choices is often challenging and needs many cycles of research to validate.

In conclusion here, we see that the two-tier structure of the Lawson’s test does impose an additional validity issue. In our new iSTAR questions, we have revised the fly-COV pairs, which retains the two-tier structure but builds up as a two stage question so that the second question is more independent to student answers in the first questions. This allows more varieties of reasonable two choice combinations as possible answers to address the interference between the two questions. More details on the related iSTAR questions will be discussed in later chapters of this thesis.

3.8 The Ceiling Effect and Measurement Saturation of the Lawson’s Test

The Lawson’s test is in general a simple test that measures learners’ fundamental reasoning components, which show non-trivial ceiling effect with students at college levels. The average scores of both freshmen US and Chinese college students in STEM
fields are similar, around 75% (US N=1046, China N=332). For non-science majors (US N=1061, China N=175), the average scores of both populations are also similar, around 60% (Bao et al., 2009).

As part of the research outlined in Bao et al. (2009), to understand the usability of the Lawson’s test at different age groups, we conducted further research to measure how the scientific reasoning ability is developed through the school and college years. We collected data with Chinese students from 3rd grade to graduate level (N_{Total} = 6357). The students are from 141 classes in 20 schools from 8 regions around China, thus they form a more representative population. The results are plotted in Figure 3.3, which shows the general developmental trend of the Lawson’s test scores of Chinese and U.S. students expanding from 3rd grade to graduate-level students. The blue dots are grade-average Lawson’s test scores. The red line is referred as a “Learning Evolution Index Curve”, which is obtained by fitting the data with a logistic function similar to the one used in item response theory (see Chapter 4 for more detail on model fitting).

The error bars shown in the graph are standard deviations of the class mean scores, which give the range of variance if one were to compare the mean scores of different classes. The U.S. data was collected in three mid-western states from 34 schools (N_{Total} = 3010). We plotted the mean scores of the U.S. data in red circles on top of the Chinese data. We can see that from 5th grade to first-year college level, the U.S. and Chinese data are within one standard deviation of each other, showing a similar developmental scale. On the measurement features, the Lawson’s test will work great with 9th graders but there will be significant ceiling effect for students from senior high school level and up.
The developmental data shows that students start to fully develop the basic reasoning abilities around their college years. Therefore, in order to assess the reasoning abilities of senior high school students, college students, and graduate students in STEM fields, we need to develop questions that involve more advanced reasoning components.

Figure 3.3. The developmental trend of Chinese and U.S. students’ Lawson’s test scores.

The problem of this ceiling is two-fold. One is that the saturation level is significantly below the 100% mark, including the graduate level students, with the maximum average
at about 80%. As the Lawson test is relatively simple, an advanced student should have reached a ceiling due to the maximum possible score. This lower ceiling strongly indicates that the presentation and wording of the questions can be an issue contributing to controversies among well-educated individuals regarding their interpretation and reasoning of the questions. This ceiling effect has the secondary effect of making it difficult to differentiate a senior in high school from a graduate student when it comes to reasoning abilities. This is unexpected, as people often expect that after four years of college, students would have fully developed their basic level reasoning skills. The limitation described here for the Lawson’s test calls for the needs to develop new instruments on scientific reasoning for the assessment of more advanced individuals.

3.9 Conclusions

In this chapter, I reviewed existing work and discussed our own research regarding the validity of the 2000 version of the Lawson Classroom Test of Scientific Reasoning. While the test has been a widely used tool in the assessment of students in their abilities as formal prisoners, multiple validity issues have been identified which include item/choice design issues, item context issues, item structure and wording issues (e.g. two-tier design), the limited scale of measurement range, and the ceiling effect for advanced students. All these call for research on revising the Lawson’s test and further development of new instruments that measure scientific reasoning.
4.1 Context of the Study

Research on student scientific reasoning ability has been gaining popularity in recent years. An important area is the assessment of the reasoning ability. The Lawson’s Classroom Test of Scientific Reasoning (LCTSR) is a readily available instrument and has been used in several recent studies. However, there isn’t established knowledge about the basic measurement parameters of the Lawson test such as the performance baselines for students at different ages and of different backgrounds. Using the Lawson test, we have collected data from students in both U.S. and China, which are analyzed to determine the developmental metric of student reasoning ability spanning from elementary school level to graduate level. The results provide quantitative measures on key assessment parameters of the Lawson test including flooring and ceiling baselines, average variations due to population differences, and developmental gains at different grade levels. These results will provide important information for measurement calibration and validation in future studies on student reasoning ability.

Although the validity of the Lawson’s test hasn’t been well established, it is the only readily available quantitative instrument on scientific reasoning and therefore, it has been
widely used. From the assessment point of view, it is always important to determine the
class measurement features of this instrument with large scale data, which will also help
further establish the validity of this instrument and provide baseline results for
researchers and teachers to properly interpret their assessment outcomes.

4.2 Data Collection

Using the Lawson’s Classroom Test of Scientific Reasoning (LCTSR), we collected
data in U.S. (N=3010) and China (N=6357) with students starting from the third grade to
graduate school. The Chinese students are from 141 classes in 20 schools from eight
regions around China; thus, they form a more representative population. The US data
were collected from three Midwestern states from 34 private and public schools. The
Chinese data from grades 3 to 12 were taken between 2007 and 2009. The US data from
grades 5 to 12 were taken between 2007 and 2010.

The college student data are from first and second year college students of science
and engineering majors enrolled in entry level calculus-based physics courses. These
groups of students form the main body of the next generation technology workforce in
both U.S.A. and China.

Data from four U.S. universities and three Chinese universities are used in this study.
The four U.S. universities are labeled U1, U2, U3, and U4. University ranking and
backgrounds are given below (based on 2007 U.S. News and World Report Ranking):

- U1 is a large research-1 state university, U.S. ranking top 50, acceptance rate
  59%.
• U2 is a large research-1 state university, U.S. ranking top 60, acceptance rate 60%.

• U3 is a large tier-4 state university with an acceptance rate of 84%.

• U4 is a large tier-3 state university with an acceptance rate of 69%.

The three Chinese universities are labeled with C1, C2, and C3. Their national rankings are also given below (based on 2007 ranking from Sina Educatoin News, http://edu.sina.com.cn). (A national university is one that is under direct control of the department of education.)

• C1 is a top 30 national university.

• C2 is a top 60 national university.

• C3 is a top 130 national university.

In the selection of universities, we targeted the ones with medium ranking in order to make a more representative pool of the population. All data from the college students were taken between 2007 and 2009. A small group of second year graduate students majoring in physics engineering from C1 and U1 respectively also took the Lawson’s test as anchoring mark for fully developed students in the formal education settings.

A summary of student performances on the Lawson’s test is given in Table 4.1. The data of the Chinese students span from 3rd grade to graduate school (grades 3 to 17), while the data of the US students span from 5th grade to sophomore year of college (grades 5 to 14). The summarized results are population means of students from each grade level and include the mean scores on the entire Lawson’s test as well as the mean scores on the six individual skill dimensions of the Lawson’s test.
<table>
<thead>
<tr>
<th>Grade</th>
<th>USA</th>
<th>N</th>
<th>S</th>
<th>SD</th>
<th>Conservation</th>
<th>Proportion</th>
<th>COV</th>
<th>Probability</th>
<th>Correlation</th>
<th>HD</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>83</td>
<td>0.274</td>
<td>0.119</td>
<td>0.360</td>
<td>0.288</td>
<td>0.123</td>
<td>0.187</td>
<td>0.260</td>
<td>0.200</td>
<td>0.269</td>
</tr>
<tr>
<td>6</td>
<td>256</td>
<td>0.371</td>
<td>0.138</td>
<td>0.575</td>
<td>0.296</td>
<td>0.222</td>
<td>0.210</td>
<td>0.292</td>
<td>0.203</td>
<td>0.466</td>
</tr>
<tr>
<td>7</td>
<td>256</td>
<td>0.365</td>
<td>0.162</td>
<td>0.570</td>
<td>0.314</td>
<td>0.221</td>
<td>0.126</td>
<td>0.275</td>
<td>0.232</td>
<td>0.522</td>
</tr>
<tr>
<td>8</td>
<td>44</td>
<td>0.372</td>
<td>0.162</td>
<td>0.586</td>
<td>0.392</td>
<td>0.215</td>
<td>0.165</td>
<td>0.293</td>
<td>0.171</td>
<td>0.536</td>
</tr>
<tr>
<td>9</td>
<td>507</td>
<td>0.453</td>
<td>0.172</td>
<td>0.646</td>
<td>0.318</td>
<td>0.246</td>
<td>0.171</td>
<td>0.358</td>
<td>0.232</td>
<td>0.664</td>
</tr>
<tr>
<td>10</td>
<td>182</td>
<td>0.488</td>
<td>0.199</td>
<td>0.679</td>
<td>0.314</td>
<td>0.264</td>
<td>0.169</td>
<td>0.424</td>
<td>0.251</td>
<td>0.656</td>
</tr>
<tr>
<td>11</td>
<td>483</td>
<td>0.604</td>
<td>0.161</td>
<td>0.792</td>
<td>0.265</td>
<td>0.333</td>
<td>0.179</td>
<td>0.498</td>
<td>0.239</td>
<td>0.875</td>
</tr>
<tr>
<td>12</td>
<td>210</td>
<td>0.642</td>
<td>0.195</td>
<td>0.829</td>
<td>0.269</td>
<td>0.395</td>
<td>0.242</td>
<td>0.575</td>
<td>0.247</td>
<td>0.810</td>
</tr>
<tr>
<td>13</td>
<td>782</td>
<td>0.708</td>
<td>0.164</td>
<td>0.864</td>
<td>0.212</td>
<td>0.629</td>
<td>0.328</td>
<td>0.590</td>
<td>0.256</td>
<td>0.867</td>
</tr>
<tr>
<td>14</td>
<td>207</td>
<td>0.758</td>
<td>0.162</td>
<td>0.920</td>
<td>0.187</td>
<td>0.739</td>
<td>0.322</td>
<td>0.666</td>
<td>0.243</td>
<td>0.930</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade</th>
<th>China</th>
<th>N</th>
<th>S</th>
<th>SD</th>
<th>Conservation</th>
<th>Proportion</th>
<th>COV</th>
<th>Probability</th>
<th>Correlation</th>
<th>HD</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>102</td>
<td>0.212</td>
<td>0.086</td>
<td>0.252</td>
<td>0.245</td>
<td>0.107</td>
<td>0.164</td>
<td>0.214</td>
<td>0.193</td>
<td>0.196</td>
</tr>
<tr>
<td>4</td>
<td>336</td>
<td>0.239</td>
<td>0.103</td>
<td>0.475</td>
<td>0.330</td>
<td>0.175</td>
<td>0.208</td>
<td>0.161</td>
<td>0.170</td>
<td>0.228</td>
</tr>
<tr>
<td>5</td>
<td>464</td>
<td>0.260</td>
<td>0.112</td>
<td>0.435</td>
<td>0.329</td>
<td>0.189</td>
<td>0.221</td>
<td>0.216</td>
<td>0.193</td>
<td>0.190</td>
</tr>
<tr>
<td>6</td>
<td>333</td>
<td>0.310</td>
<td>0.144</td>
<td>0.520</td>
<td>0.351</td>
<td>0.263</td>
<td>0.304</td>
<td>0.226</td>
<td>0.196</td>
<td>0.298</td>
</tr>
<tr>
<td>7</td>
<td>529</td>
<td>0.338</td>
<td>0.143</td>
<td>0.594</td>
<td>0.363</td>
<td>0.347</td>
<td>0.334</td>
<td>0.252</td>
<td>0.222</td>
<td>0.259</td>
</tr>
<tr>
<td>8</td>
<td>378</td>
<td>0.409</td>
<td>0.177</td>
<td>0.704</td>
<td>0.330</td>
<td>0.491</td>
<td>0.370</td>
<td>0.318</td>
<td>0.253</td>
<td>0.300</td>
</tr>
<tr>
<td>9</td>
<td>765</td>
<td>0.468</td>
<td>0.169</td>
<td>0.802</td>
<td>0.292</td>
<td>0.490</td>
<td>0.364</td>
<td>0.363</td>
<td>0.247</td>
<td>0.427</td>
</tr>
<tr>
<td>10</td>
<td>1195</td>
<td>0.533</td>
<td>0.193</td>
<td>0.834</td>
<td>0.278</td>
<td>0.585</td>
<td>0.374</td>
<td>0.453</td>
<td>0.284</td>
<td>0.424</td>
</tr>
<tr>
<td>11</td>
<td>1328</td>
<td>0.600</td>
<td>0.176</td>
<td>0.893</td>
<td>0.232</td>
<td>0.701</td>
<td>0.339</td>
<td>0.493</td>
<td>0.266</td>
<td>0.474</td>
</tr>
<tr>
<td>12</td>
<td>458</td>
<td>0.702</td>
<td>0.169</td>
<td>0.920</td>
<td>0.212</td>
<td>0.784</td>
<td>0.279</td>
<td>0.612</td>
<td>0.271</td>
<td>0.619</td>
</tr>
<tr>
<td>13</td>
<td>248</td>
<td>0.738</td>
<td>0.159</td>
<td>0.851</td>
<td>0.292</td>
<td>0.773</td>
<td>0.292</td>
<td>0.691</td>
<td>0.227</td>
<td>0.673</td>
</tr>
<tr>
<td>14</td>
<td>122</td>
<td>0.765</td>
<td>0.156</td>
<td>0.895</td>
<td>0.222</td>
<td>0.777</td>
<td>0.305</td>
<td>0.728</td>
<td>0.231</td>
<td>0.721</td>
</tr>
<tr>
<td>17</td>
<td>99</td>
<td>0.783</td>
<td>0.136</td>
<td>0.937</td>
<td>0.193</td>
<td>0.851</td>
<td>0.228</td>
<td>0.652</td>
<td>0.260</td>
<td>0.770</td>
</tr>
</tbody>
</table>

Table 4.1. Summary of Lawson’s Test Data from USA and China. COV (Control of Variables), HD (Hypothetical Deductive)
4.3 The Developmental Scales of the Lawson’s Test Scores

The results in Table 4.1 show a steady developmental trend from young children to fully developed learners. The results on the total scores from Chinese and US students are also comparable. However, large differences between the two populations are also observed on several skill dimensions. To make the comparisons visually straightforward and quantitatively accurate, we model the developmental trend using a logistic function motivated by item response theory (Hambleton & Swaminathan, 1985).

4.3.1 The Learning Evolution Index Curve of the Lawson’s Test

To quantitatively determine the features of the developmental trends, we fit the Lawson’s test scores with a logistic function shown in Eq. (4.1):

\[
y = F + \frac{C - F}{1 + e^{-\alpha(x-b)}}
\]  

(4.1)

where

- \(x\) – Student grade level
- \(y\) – Student score
- \(F\) – Floor, the model predicted lowest score based on student data
- \(C\) – Ceiling, the model predicted highest score based on student data
- \(\alpha\) – Discrimination factor, which determines the steepness of the central part of the curve.
- \(b\) – Grade-equivalent difficulty level, which controls the center of the curve.
Since the Chinese data is in large quantity and covers a wide spectrum of social economic backgrounds, we fit the model described in Eq. 4.1 with the Chinese data and obtained the developmental curve shown in Figure 4.1. The mean LCTSR scores of the U.S. data are scatter plotted and overlaid on top of the curve. The error bars are the standard deviations of the student scores. We can see that from 5th grade to sophomore
year in college, the U.S. and Chinese data follow each other very well, showing a similar developmental scale.

The results shown in Figure 4.1 suggest that the U.S. and Chinese students have a similar developmental path during school years. The development occurs most rapidly during 8-11 grades. For college students, the results are affected by the selection process, especially in China where only the top 10% high school graduates can enter the tier-1 universities. It is interesting to observe that the average score of 12th grade Chinese students is almost identical to that of the U.S. students in non-selective universities.

During the college years, the reasoning ability doesn’t change much. This result is consistent with existing studies which showed that the current education in STEM disciplines often train students on content understanding but have little impact on reasoning (Bao et al., 2009). Our research has shown that traditional physics courses make little changes on students’ pre and post LCTSR scores (pre-post test score effect size ~0.1), while inquiry based courses often make a sizeable impact (pre-post test score effect size ~0.6) (Bao et al., 2009). Since the components of scientific reasoning such as control of variables, correlation analysis, and hypothesis testing are explicitly and repeatedly emphasized in inquiry-based learning, it is not surprising that this type of instruction can help develop students’ abilities in these areas.

For quantitative evaluations of the model fitting, the fitting parameters are summarized in Table 4.2, which also includes the fits of the six individual skill dimensions. The performances of the fits are evaluated in terms of the root-mean-square deviations (RMSD) between the actually measured scores and the model predicted
scores. Two types of RMSDs are given. One is the RMSD of the mean scores of each grade level and the other is the weighted population RMSD calculated with each individual student’s measured and predicted scores. As expected, the mean score RMSD is much smaller than the population RMSD, since the calculation of mean scores will remove a significant amount of variances from the data. Comparing to the average standard deviations (SD) of individual student scores, the population RMSD is typically 1/4 to 1/3 of the SD and the mean score RMSD is about 10% of the SD, which suggests that the model fits the data well.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Conservation</th>
<th>Proportion</th>
<th>COV</th>
<th>Probability</th>
<th>Correlation</th>
<th>HD</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.829</td>
<td>0.940</td>
<td>0.860</td>
<td>0.770</td>
<td>0.790</td>
<td>0.860</td>
<td>0.820</td>
</tr>
<tr>
<td>F</td>
<td>0.201</td>
<td>0.280</td>
<td>0.050</td>
<td>0.200</td>
<td>0.190</td>
<td>0.160</td>
<td>0.240</td>
</tr>
<tr>
<td>α</td>
<td>0.474</td>
<td>0.576</td>
<td>0.504</td>
<td>0.693</td>
<td>0.499</td>
<td>0.334</td>
<td>0.819</td>
</tr>
<tr>
<td>Mean Score RMSD</td>
<td>0.014</td>
<td>0.046</td>
<td>0.030</td>
<td>0.025</td>
<td>0.032</td>
<td>0.047</td>
<td>0.023</td>
</tr>
<tr>
<td>Weighted Population RMSD</td>
<td>0.106</td>
<td>0.079</td>
<td>0.106</td>
<td>0.062</td>
<td>0.092</td>
<td>0.093</td>
<td>0.074</td>
</tr>
<tr>
<td>Average SD</td>
<td>0.148</td>
<td>0.282</td>
<td>0.291</td>
<td>0.232</td>
<td>0.327</td>
<td>0.405</td>
<td>0.255</td>
</tr>
</tbody>
</table>

Table 4.2. The model fit parameters and the root-mean-square deviations (RMSD) of the fit for the mean scores and population.

The fitting parameters provide quantitative scales of the measurement features of the instrument, namely the Lawson’s test. From the fit of the total score, the model predicts a nominal ceiling of 83% and a floor of 20%. The floor is consistent with the multiple
choice design of 5-answer questions. The ceiling is only 83%, instead of 100%, which gives a difference of approximately 5 questions (out of the total 24 questions). This is consistent with our analysis in Chapter 3, which shows that a number of the Lawson’s test questions have design and wording issues that contribute to a variety of different interpretations of the questions leading to large uncertainties in answering the questions even with fully developed learners.

The “b” parameter of the fit of the total test score, which gives the grade-level difficulty, is found to be 9.68. This means that students from grades between 9 and 10 would score in the middle between the floor and the ceiling. Since the Lawson’s test was indeed designed to measure the high school students’ scientific reasoning, the observed difficulty level matches well with the targeted population.

The “α” parameter, often called the discrimination parameter in the original IRT model, describes how quickly students’ scores change with grade levels. For the logistic model, students’ scores change most rapidly when the grade level equals the difficulty. Taking the derivative of Eq. 4.1 and for $\alpha=0.47$ at $x=b$, we get

$$\frac{dy}{dx} \bigg|_{x=b} = \frac{\alpha(C-F)e^{-\alpha(x-b)}}{(1+e^{-\alpha(x-b)})^2} \bigg|_{x=b} \approx 7.5\%.$$  \hspace{1cm} (4.2)

This shows that the most rapid change of the score is approximately 7.5% per year of development, which occurs between grades 9 and 10. Comparing to the measured outcomes summarized in Table 4.1, this change is about half of the standard deviation of student scores. Therefore, we can see that the baseline for highest development the
Lawson’s test scores happen at the freshman high school year with an yearly effect size of approximately 0.5.

The quantitative results from the model fitting can help researchers and teachers to compare and align their particular measures of student gains on scientific reasoning. The results also show that the scientific reasoning ability is a slow-changing ability without targeted instruction in formal education settings.

Summarizing our results, it is clear that current style of STEM education, even carried out at a very demanding level, has little impact on developing students’ scientific reasoning abilities. It is not what we teach but how we teach will make a difference on what students can develop. Ideally, we want students to gain on both content and reasoning; therefore, we need to put in more research and effort on developing a balanced method of education, incorporating more inquiry based learning into the education settings.

4.3.2 The Developmental Scales of the Six Skill Dimensions of the Lawson’s Test

The Lawson’s test measures six dimensions of the scientific reasoning. These include conservation of mass and volume, proportional thinking, control of variables, probabilistic thinking, correlational thinking, and hypothetical deductive reasoning. The mapping of the individual test items and the six skill dimensions are listed in Table 4.3.

A quick glance of the results shown in Table 4.1 shows that although the total Lawson’s test scores and the developmental trends of US and Chinese students are similar, large differences exist in student scores on the several skill dimensions.
To facilitate the comparisons of students’ performances on the individual skill dimensions, the developmental model described in Eq. 4.1 is fitted with student data on the different skill dimensions. The results are discussed in the sections below.

4.3.3 The Developmental Curve of the Conservation of Mass and Volume

In the Lawson’s test, questions 1-4 measure students’ understanding of conservation of mass and volumes. This is a very basic reasoning skill that children develop at young age and is often considered as a conceptual construct in the formation of more complicated mental models of the (Halford, 1993). Therefore, it is expected that students should have high scores on this dimension at early grade levels. The result of the model fit is plotted in Figure 4.2. Similar to the method used in fitting the total scores, the model is fitted with the Chinese data only and the US mean scores at each grade level are scatter plotted on top of the fitting curve.
From Figure 4.2, we can see that students start to fully develop this skill early on. The fitting parameters indicate high ceiling (0.94) and floor (0.28), highest among all the skill dimensions (see Table 4.2), while the difficulty level is the lowest (6.74). The discrimination parameter is also relatively high among the different skill dimensions, indicating a moderately fast developing skill.

It is worth noting that although young children will develop the fundamental ideas of conservation before kindergarten age, the Lawson’s test questions require not only the
basic conservation ideas but also the ability to read and comprehend the question narratives and the underlying contexts and to apply the ideas into the contexts. Therefore, students need to establish their reading and comprehension skills before they can correctly answer the Lawson’s test questions, which can be one of the major contributing factors for the conservation dimension to have a difficulty level equivalent to grade 6 to 7 but not earlier.

Comparing the US and Chinese data, we can see that the Chinese students are slightly ahead of the US students on this skill dimension during the middle school and high school years. The two populations start to split around 8th grade and re-converge during the college years. The possible causes for this observation are subject to many possible factors including sample selection, science curriculum in elementary schools, etc. With the current data size and sample distributions, it is difficult to identify conclusive causes to explain the differences, which will not be pursued in the scope of this thesis.

4.3.4 The Developmental Curve of the Proportional Reasoning

The ability of handling ratios and proportionality in problem solving is a highly emphasized skill, which is often considered as the fundamental requirement for STEM learners (McLaughlin, 2003). (See Chapter 2 for a detailed review of research on proportions and ratios.) Using a similar method, the development model is fitted with the student data on dimension of proportional reasoning and fitting result is shown in Figure 4.3.
From Figure 4.3, we can see a quite low floor (0.05) which below chance. Since the proportional reasoning is highly dependent on basic math computation skills typically taught grades 5-7, low scores at lower grade levels (3-5) are often expected. The fact the floor is below chance may be caused by the mathematical nature of the questions – students will engage certain types of math computation, rather than guessing, to generate answers to the questions. Using guessing will create a probability for correct answers at the chance level, however, using computation can cause the probability for correct
answers to be lower than chance as the lower grade level students will more frequently encounter computation errors and produce wrong answers.

It is noticed that the ceiling of the fit is only 0.86. It can be surprising to see that college and graduate students are not able to score perfectly on these simple math questions. As discussed in Chapter 3, it is believed that the major contributor to this lower than expected ceiling is due to issues of the question designs. As detailed in Chapter 3, Question 8, which is the reasoning part of the second two-tiered questions on proportional reasoning has two choices that experts would consider equivalent. The Chinese students overwhelmingly scored low on this question as they hesitate to pick the generally true choice as the correct answer.

From the fit, the difficulty level turns out to be 8.2, which is the second easiest among the six dimensions. This is consistent with the Chinese curriculum, which emphasizes math drilling. Comparing the Chinese data with the US data, we can see a clear gap: starting from the 7th grade, the Chinese students outperform the US students by 1 to 2 standard deviations. The gap starts to diminish at college ages as the US students become more fluent in basic math computations and the Chinese students hit the ceiling on the Lawson’s test.

4.3.5 The Developmental Curve of the Control of Variables

The control of variables is a major category in scientific reasoning. As detailed in Chapter 2, many of the existing studies on scientific reasoning did research solely in this particular area. The Lawson’s test contains six questions on control of variables, showing
its emphasis on this skill. The contexts of the questions have a strong influence of biology course background. However, the contexts are simple enough so that learners without established knowledge of biology are expected to be able to understand and reason through the questions.

![Control of Variables](image)

**Figure 4.4.** The developmental trends on control of variables.

The fitting results on control of variables are plotted in Figure 4.4. For this skill, the ceiling is the lowest among all skill dimensions (see Table 4.2). This is likely the result of
the item design issues discussed in Chapter 3, where the design of choices of question 12, which is the reasoning part of the first pair of fly question, was found to be controversial.

The floor of student scores is found to be right around the chance level (0.2), indicating that students without developed understanding of control of variables often guess for the answers.

The difficulty level of this skill dimension is between grades 10 and 11, slightly higher than that of the Lawson’s test. The discrimination is the second highest among all skill dimensions, suggesting that COV is more rapidly developed around sophomore years in high school.

Comparing the Chinese and US data, it appears the both populations are similar on this skill dimension – the differences are typically less than a small fraction of the standard deviations. This implies that any culturally embedded education factors within the two countries didn’t make a significant impact on students’ development of COV skills.

4.3.6 The Developmental Curve of the Probabilistic Reasoning

Understanding probability is an important and fundamental ability for students to correctly interpret scientific data and conduct data analysis as most scientific experiments involve uncertainties of many kinds, some are systematic while others are stochastic. In the Lawson’s test, four items are devoted to the measurement of basic probabilistic reasoning. The contexts are straightforward involving simple scenarios of counting objects of specific features and finding the likelihood for certain combinational patterns
to occur. Typically, these skills are addressed in math courses in middle schools. The results of the model fitting for this skill dimension are shown in Figure 4.5.

![Figure 4.5. The developmental trends on probabilistic reasoning.](image)

From Figure 4.5, the most striking result is that the US students outperform the Chinese students by a big step, at least one standard deviation ahead. The gap started immediately at the beginning of the middle school (6th grade) and sustained ever since all
the way through college. Based on our validity evaluations discussed in Chapter 3, there are no known issues concerning the probability questions in the Lawson’s test, and therefore, the gap must be of an origin residing in the educational and or cultural settings of the two populations.

Discussions with teachers from both countries have revealed traces of evidence that may explain the differences. In the US, probability is a standard component in middle school math curriculum. It is also emphasized by teachers and taught with hands on activities very similar to the contexts of the questions in the Lawson’s test.

While in China, probability is only slightly touched in the curriculum partially due to its simplistic nature in computation. Teachers often assume students would understand the counting based frequency calculation very quickly, spending only a few classes on it, and jump into more complicated topics such as combinations and conditional probabilities. This makes it difficult for students to grasp the main underlying concepts such as randomness and independence. The exposures to students are often in the form of narrowly structured questions that require more complicated calculation but less in conceptual modeling. These are possible causes from the formal education settings that might have contributed to the lower performance of Chinese students on probabilistic reasoning, especially when facing a real life context.

From a more culturally based perspective, in both real life and the exposed formal education, students in China are always heavily guided in learning and all the problems they have encountered have precisely defined correct answers. Years of such training
often lead students to develop a preference for certainty and in the meantime a deficit in proper understanding of uncertainty.

The possible causes of this gap are based on a small number of unstructured discussions with teachers from USA and China. The results may not be representative but are recognized by the involved educators for being plausible hypotheses. A common consensus is that this area of reasoning is definitely a topic warrants more in depth studies, which are being pursued in our current research.

4.3.7 The Developmental Curve of the Correlation Thinking

The correlation thinking bears a lot of similarities to probabilistic reasoning. One way to interpret a correlation is the conditional probability for a pair of events to co-exist. For example, one can easily rephrase a statement to explain a correlation between two events in terms of the likelihood to observe one of the events should the other occurs or not. Likewise an established understanding of correlation is fundamental to students’ capacity in analyzing experimental data and drawing conclusions.

On the Lawson’s test, there are only two questions on correlation, which also use a simple context of counting mice of various forms from a defined area. The model fitting results on the correlation dimension are plotted in Figure 4.6. Due to the small number of questions, the variance of the data is quite large causing an average standard deviation in the order of 40%.
The results on the dimension of correlation are in general similar to that of probabilistic reasoning. The US students also outperform Chinese students but the gap is smaller, averaging at half of a standard deviation. The possible causes of this gap is also similar. In China, statistical measure such as correlation are not emphasized in the K-12 curriculum and even when it is introduced, students often learn to calculate a well-defined problem using the summation equation for calculating correlation. Seldom would they encounter a problem that gives a real world experimental setting and requires...
students to both identify the related variables and come up with a possible model behind the identified variables.

The fitting parameters also suggest that for the Chinese students the difficulty level is the highest among the six skills dimensions, and it is also the slowest changing ability.

4.3.8 The Developmental Curve of the Hypothetical-deductive Reasoning

The hypothetical-deductive reasoning is thought to be most complicated ability in the Lawson’s test (and therefore put at last), representing the last stage of formal reasoning. Nevertheless, it is the most important core skill in scientific reasoning as hypothesis testing is always the goal of scientific inquiries and applications of scientific methods.

On the Lawson’s test, the last four questions are reserved for this skill dimension. These are fairly long questions requiring a whole page of reading and parsing. The two questions in a pair are not structured as answer-explanation anymore. In the first pair (questions 21-22), the first question provides a number of experimental design while the second question gives a set of possible experimental outcomes. The two questions need to be coordinated in order to form a consistent pair of design and outcome that also can be used to test the provided hypothesis. In the second pair of the questions (questions 23-24), the narratives of the questions present an experimental setting and two possible hypotheses. The first question asks for a selection of experimental outcomes that would prove the first hypothesis wrong, while the second question asks for experimental outcomes that would negate the second hypothesis. As a result, in order to respond correctly to these questions, students need to have a well-established reading and
comprehension capacity as well as information processing skills to parse out the useful information from an abundant collection of co-existing but not relevant features. The results of the model fitting are given in Figure 4.7.

![Figure 4.7. The developmental trends on hypothetical-deductive reasoning.](image)

The model fitting results show that hypothesis testing is a more advanced ability that students start to develop in high school years. It is also the most rapidly changing (discriminating) ability among all six skill dimensions. On this skill dimension, the
ceiling of the Lawson’s test questions is a little over 80%. Two potential causes of this low ceiling have been observed in research. One is the length of the reading, which often causes students to lose track of the relevant experimental structure and variables and misinterpret the questions. The other is related to the contextual elements of the questions. Students often tried to use their prior knowledge about red cells in question pair of 21-22 to answer the questions rather than using reasoning. In question pair 23-24, it uses a plastic bag that is semi-permeable. This is often considered against common sense as most plastic bags encountered in real life are water proof. Therefore, some students thought the designs were implausible, which prevented them from reasoning further through.

Comparing the Chinese and US data, we see that both populations are similar all the way up to 9\textsuperscript{th} grade. The Chinese students start to outperform US students from 10\textsuperscript{th} grade and beyond. This might be the result of population selection. In China, the period of compulsory secondary education is grades 1-9. The high school starts at 10\textsuperscript{th} grade and about 50\% of the middle school students move on to attend high school. While in the school communities tested in USA, almost all middle school students move on to high school and the high school starts at 9\textsuperscript{th} grade. Therefore, it is possible that the gap between US and Chinese students after 10\textsuperscript{th} grade is due to the population selection at the onset of high school.
4.4 Summary

This chapter presents a detailed analysis of the developmental trends of scientific reasoning abilities for US and Chinese students measured with the Lawson’s test. A logistic model is developed and fitted to the data. The model fits the data reasonably well with a mean score RMSD at the 10% level of the average standard deviations.

The parameters obtained from the model fitting provide quantitative measures to compare the overall scientific reasoning ability and the individual skill dimensions. The results show that the Chinese and US students are similar in their total scores on the Lawson’s test but diverge in five out of the six skill dimensions. The actual causes for the differences are still under investigation; however, initial clues suggest that cultural and educational settings of the two countries have a lot to contribute to the differences.

The analysis also provides quantitative metrics for the difficulty levels of the Lawson’s test and the individual skill dimensions. The results show that this test is optimized for assessing high school students.

From the model fitting, we also see the common issue of low ceiling, which is typically in the 80% level. This is consistent with the test design issues identified in the validity evaluation studies discussed in Chapter 3.
5.1 Context of the Study

Scientific reasoning skills have been widely researched, in which the Lawson Classroom Test of Scientific Reasoning is widely used. The study discussed in this chapter aims at developing a method of mining data from the Lawson’s Test. By analyzing the patterns of responses to four items from the Lawson’s test, we have determined that students providing the correct answer without the correct reasoning are at an intermediate level of understanding, a construct that is overlooked by traditional scoring of two-tier test items. From this, we were able to identify six levels of performance based on student responses to the Lawson’s Test. Based on the analysis, a new scoring method for the Lawson’s Test is also proposed.

5.2 Review of Learning Progression in the Context of Scientific Reasoning

Scientific reasoning skills are vital to a student’s education. In her extensive review of scientific reasoning literature, Zimmerman (2007) claims that investigation skills and content knowledge bootstrap one another, creating a relationship that underlies the development of scientific thinking. Research has been conducted to determine how these
scientific thinking skills can best be fostered and which teaching strategies contribute most to learning, retention, and transfer of these skills. Zimmerman found that children are more capable in scientific thinking than was originally thought, whereas, and that adults are less so. Additionally, there she found that there is a long developmental path followed by those acquiring scientific thinking skills, student performance varies as they progress along this path. Zimmerman (2007) further stated that scientific thinking requires a complex set of cognitive skills, the development of which requires much practice and patience. It is important, then, for educators to understand how scientific reasoning abilities develop.

The idea that students build their knowledge progressively is well established; Lawson (1979) states that the developmentalist’s view of intelligence is not that of innate ability but rather how a student’s abilities have progressed over time. Lawson also poses important questions: how does reasoning develop, and does every student develop skills in the same order and at the same rate? Learning progressions, a relatively recent focus for researchers (Duncan & Hmelo-Silver, 2009; Steedle & Shavelson, 2009), are excellent tools that can be used in answering these questions.

Generally defined, a learning progression is a way of describing how students build their knowledge of a certain concept over time (Alonzo & Steedle, 2009). Duncan and Hmelo-Silver (2009) offer a four-part, comprehensive definition. (1) Learning progressions are focused on a few basic ideas and practices that are developed over time. (2) Learning progressions have an upper and lower bound. The upper bound describes what students are expected to know and is determined by academic standards (Alonzo &
Steedle) and research in the content area of the learning progression. The lower bound is
determined by the prior knowledge and skills students have as they enter the progression.
(3) Learning progressions are comprised of levels that describe the steps students take
between the lower and upper bounds. The levels are determined by examining existing
research and by empirical studies of the progression. (4) Learning progressions do not
describe learning as it would occur naturally; instruction is required. Learning
progressions may seem linear, but they do not assume that all students follow a single
path through the progression.

Other researchers have described the traits of a learning progression. The period of
time covered by a learning progression can vary from one instructional unit to several
years (Alonzo & Steedle, 2009), and the number and size of levels can differ between
progressions (Duncan & Hmelo-Silver; 2009). Additionally, learning progressions are at
least partly hypothetical in nature, so research is needed to verify the learning progression
(Duncan & Hmelo-Silver, 2009; Steedle & Shavelson, 2009). Alonzo & Steedle (2009)
argue that longitudinal studies are necessary in validating a learning progression.
Learning progressions can be informed by new research and should be modified
accordingly. This leads to an iterative process: define a learning progression, use it to
design assessment instruments, use these assessments to modify the learning progression,
and so on.

There are multiple well-defined learning progressions already, such as force and
motion (Alonzo & Steedle, 2009) and biodiversity (Songer, Kelcey, & Gotwals, 2009),
but there are limited studies relating to learning progressions and scientific reasoning. In
the field of biology, Lawson, Alkhoury, Benford, Clark, and Falconer (2000) proposed a
progression describing student reasoning. The levels they named were (1) sensory-motor
stage, (2) preoperational stage, (3) descriptive concepts, (4) hypothetical concepts, and
(5) theoretical concepts. They ordered the levels based on the difficulty of the tasks
associated with each level; descriptive concepts are based on experience, hypothetical
concepts require imagining past or future events, and theoretical concepts cannot be
derived from observation. A student at a particular level should be able to perform at all
lower levels as well. Their study used a modified form (Lawson, 2000) of the original
Lawson’s Classroom Test of Scientific Reasoning (Lawson, 1978), so it is highly relevant
to our focus of scientific reasoning skills. The progression that was developed serves
biology concepts well as it categorizes these concepts according to difficulty. However,
the scientific reasoning aspects of the progression are not fine-grained.

Since national standards indicate that students are expected to learn scientific
reasoning skills (National Research Council, 1996), educators need a way to determine
what their students should know and at what point in their schooling they should know it.
Lawson (1979) states that only material that is appropriate to the developmental level of
the students should be utilized. By defining a scientific reasoning learning progression,
curricula could be appropriately tailored to a given age group. To define a scientific
reasoning learning progression, more work is needed.

Defining a learning progression is not an easy task, particularly because each student
learns in a different way (Alonzo & Steedle, 2009), but a common trait among learning
progressions is that they are designed based on research evidence (Duncan & Hmelo-
Silver, 2009). Typical methodology in researching a learning progression involves both qualitative (interviews, open-ended questions) and quantitative (multiple choice) data. One such strategy is ordered multiple choice (OMC) (Briggs, Alonzo, Schwab, & Wilson, 2006). OMC items are unique in that they are based on what Briggs et al. call a construct map—a model of student cognitive development. Each answer option depicts a different level of understanding within the construct map. This allows for straightforward diagnosis of the student’s performance. Briggs et al. state OMC items are useful because they provide more diagnostic information than traditional multiple choice items while remaining efficient.

Briggs et al. (2006) state that the key to writing effective OMC items is carefully defining a construct map for the concept at hand. Distracters are written to represent common misconceptions or errors that students would be expected to make. These errors represent different levels of the construct map. To validate OMC items, Briggs et al. used open-ended questions in the same content area.

OMC items have great diagnostic power (Briggs et al., 2006). Briggs et al. describe a chain of reasoning to support this claim: OMC items are based on concept maps; concept maps are based on student cognitive development; cognitive development is based on national standards; thus, OMC items can indicate where students are in a larger content domain, which cannot be achieved with traditional multiple choice.

Alonzo and Steedle (2009) employed the OMC methodology described above in a study of the force and motion learning progression. One of their research goals was to determine the advantages and disadvantages of OMC and how OMC compares to more
Alonzo and Steedle’s (2009) study used a learning progression that was already well-defined, but not all learning progression research is done this way. Songer, Kelcey, and Gotwals (2009) developed a biodiversity learning progression from scratch. They defined both content and inquiry reasoning progressions that combine to form the learning progression. Their development process can be described by five steps.

The first step was working with experts (in this case, zoologists) to determine the focal points of the learning progression (Songer, Kelcey, & Gotwals, 2009). The goal
was to make content “simply complex” (i.e. maintain rigor but make it accessible to fourth- through sixth-graders). In the second step, the focal points were translated into curricular activities. This was done by referencing previous work in scaffolding as well as working with current teachers. The third step involved developing assessment items that corresponded to the content and inquiry reasoning progressions. Pre-tests, embedded assessments, and post-tests were developed using both forward and reverse engineering. The purpose of the test items was to indicate how the students’ knowledge development connected to the progressions. The fourth step was to empirically evaluate the content and inquiry reasoning progressions using the assessment instruments from step three. This was done using both cross-sectional and growth curve analyses. The growth curve was done in a piecewise fashion since the researchers did not want to assume linear growth. The fifth and final step was to take the results of step four, as well as national and state education standards, and develop a three year learning progression.

The biodiversity learning progression (Songer, Kelcey, & Gotwals, 2009) is fully developed (Duncan & Hmelo-Silver, 2009), but other learning progressions are still in their early stages of development. One example is Duncan, Rogat, and Yarden’s (2009) modern genetics learning progression. The progression, which was developed based on existing research of student understanding in genetics as well as national education standards, has well-defined levels, but it has not yet been validated. There are eight big ideas within the progression that fall into three models (genetic, meiotic, and molecular). These three models were previously established, but Duncan, Rogat, and Yarden expanded them to include what it means to have understanding at those levels and how
that understanding might develop. The learning progression was further organized around two domain-specific questions: how do genes influence how organisms look and function and why do organisms vary in how they look and function? These two questions provide a meaningful way to categorize important ideas. The learning progression spans grades five through ten. Fifth grade was chosen as the starting level because students have been exposed to some ideas in genetics at this point.

An important point made by Duncan, Rogat, and Yarden (2009) is that the three models named above should all be introduced at the same time. They state that progress is not defined by simply learning what the models are. Rather, developing more sophisticated versions of the models and the relationships between them indicates mental growth. This growth falls into three bands, each of which loosely encompasses two grade levels. The expectations of student performance at each band were developed based on the theoretical framework of the learning progression as well as research done in student learning of genetics. However, these expectations have yet to be tested, and empirical evidence will be necessary to refine and validate the learning progression.

One way Duncan, Rogat, and Yarden (2009) plan on assessing their learning progression is through the use of learning performances. Learning performances combine scientific inquiry practices with scientific concepts to describe the ways in which students should be able to use their knowledge. Duncan, Rogat, and Yarden have developed a set of these learning progressions that reflect the big ideas in genetics at each grade level. Variation in student responses will indicate different positions within the learning progression. They acknowledge that the learning progressions will need to have
psychometric properties in order to be reliable. A possible downside to the learning performances is that they may not cover all the levels within the learning progression; that is, an item may reveal placement in levels two and three but not level one.

Described above were three learning progressions in different stages of development. Any learning progression, regardless of its stage of development, needs to be refined and verified using empirical data. Qualitative data has the advantage of being full of information, but it is time-consuming to obtain. Quantitative data is readily available, but the methods with which it can be used are still lacking. There are vast amounts of existing data (from standardized tests, for example) that we can not yet utilize.

One test that has potential in applying to defining a scientific reasoning learning progression is Lawson's Classroom Test of Scientific Reasoning (Lawson, 1978). The Lawson Test has been widely used by the education research community (Bao et al., 2009). Additionally, in validating his test, Lawson used both quantitative and qualitative methods. As a result, there is a complete and rich existing data set available, but how can we use it? Is there a way we can extract learning progression information from these resources? Data we have collected in a large scale assessment using the Lawson Test shows a developmental progression through the grade levels, which suggests that the existing data has great potential to be useful. Data mining could provide an opportunity to define a scientific reasoning learning progression and would renew existing data.

Data mining is not a new technique. It is part of a larger effort referred to as knowledge discovery in databases, which is centered on the investigation of processes, algorithms, and mechanisms for retrieving potential knowledge from data collections
At present, there is not an established method for renewing data from the Lawson Test to define a scientific reasoning learning progression. While there are many methodologies available to extract information from data (Norton, 1999), the one that we believe will be most applicable is pattern recognition. We will need to determine how to utilize this method with the existing data to extract information and define learning progression levels. Doing so could be extremely valuable as data that is no longer in use would become renewable. In addition, we would be able to define a scientific reasoning learning progression based on a valid assessment instrument.

5.3 Research Design to Study Learning Progress in Scientific Reasoning

The overarching goal of this research is to develop a better method for categorizing student scientific reasoning ability. The result will help identify possible learning progression levels based on responses to Lawson’s Test, which will shed light on understanding and revising the scoring method of the Lawson’s test.

Lawson (1978) originally designed his test of formal reasoning to address the need for a reliable, convenient assessment tool that would allow for diagnosis of a student’s developmental level. A valid form of measurement prior to the Lawson Test was the administration of Piagetian tasks. This method, however, is time-consuming and requires experienced interviewers, special materials, and equipment. A paper and pencil test would be more practical for classroom use, but there are also problems with this method. Paper and pencil tests require reading and writing ability, test takers have no motivation from materials or equipment to use, and it is not as relaxed as a clinical interview setting.
In the development of his test, Lawson (1978) aimed for a balance between the convenience of paper and pencil tests and the positive factors of interview tasks. He studied eighth- through tenth-grade students to determine their scientific reasoning skill level. Lawson breaks scientific reasoning into several categories: isolation and control of variables, combinatorial reasoning, correlational reasoning, probabilistic reasoning, and proportional reasoning. Test items were based on these dimensions. The original format of the test had an instructor perform a demonstration in front of a class, after which the instructor would pose a question to the entire class and the students would mark their answers in a test booklet. The booklet contained the questions followed by several answer choices. For each of the 15 test items, students had to choose the correct answer and provide a reasonable explanation in order to receive credit for that item.

To establish the validity of his test, Lawson (1978) compared test scores to responses to interview tasks, which were known to reflect the three established levels of reasoning (concrete, transitional, formal-level). He found that the majority of students were classified at the same level by both the test and interview tasks but that the classroom test may slightly underestimate student abilities. Validity was further established by referencing previous research on what the test items were supposed to measure as well as performing item analysis and principal-components analysis.

The context for our study is the modified form of Lawson's Classroom Test of Scientific Reasoning. It is a 24 item, two-tier, multiple choice test. Treagust (1995) describes a two-tier item as a question with some possible answers followed by a second question giving possible reasons for the response to the first question. The reasoning
options are based on student misconceptions that are discovered via free response tests, interviews, and the literature.

The traditional scoring method for two-tier items, such as those on the Lawson Test, is described by Table 5.1; both the answer to the question and the reasoning need to be correct in order for the student to receive credit (Lawson, 1978; Treagust, 1995).

<table>
<thead>
<tr>
<th>Answer</th>
<th>Reasoning</th>
<th>Total score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect</td>
<td>Incorrect</td>
<td>0</td>
</tr>
<tr>
<td>Incorrect</td>
<td>Correct</td>
<td>0</td>
</tr>
<tr>
<td>Correct</td>
<td>Incorrect</td>
<td>0</td>
</tr>
<tr>
<td>Correct</td>
<td>Correct</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5.1. Traditional scoring on a two-tier item from the Lawson Test.

According to the traditionally used scoring method for two-tier type questions, the first three rows of the table represent equivalent skill levels. This leads to a step-function in the scoring of a particular problem. We feel we can identify skill levels at a finer grain size. We believe each row of Table 1 represents a different level of understanding. Getting both the answer and the reasoning incorrect certainly indicates the lowest skill level while getting both correct indicates the highest, but the skill level when only the answer or the reasoning is correct is unclear. One goal of our research is to identify if getting just the answer or just the reasoning correct indicate different skill levels, and
which represents a higher skill level. Once this is accomplished, student scores will resemble a ramp from low to high skill levels.

We hypothesize that students may understand the answer to a question before they can fully articulate the reasoning behind their response. This is based on teaching experience; we have seen that students can often recite an answer without being able to describe the reasoning that led them to that answer. There is also existing research proposing that reasoning is preceded by a subconscious bias toward the right answer (Bechara, Damasio, Tranel, & Damasio, 1997). Bechara et al. studied risk-taking behavior by having a control group and a test group (patients with prefrontal cortex damage and decision-making deficiencies) perform a gambling task. They found that the control group began to choose advantageously even before they had realized the correct strategy. This behavior was not seen in the test group at all; in fact, they continued to choose disadvantageously even when they knew the correct strategy. This suggests that there is a subconscious influence on decision-making that develops before reasoning. Bechara et al. proposed that this subconscious bias calls on the individual’s previous experience.

This research suggests to us that students who answer the question correctly but the reasoning incorrectly are at a higher level of skill that those who answer both incorrectly or just the reasoning correctly. To study this, we will examine student performance on several questions from the Lawson Test.

When Lawson (1978) first designed his test, he found that the questions typically fell into three levels of difficulty. Similarly, we have chosen items that appear to be easy or
difficult for students based on the overall performance. The items we have chosen are shown in Figure 5.1 and will be referred to as P1 (pendulum answer, easy), P2 (pendulum reasoning, easy), F1 (flies in a tube answer, difficult), and F2 (flies in a tube reasoning, difficult). See Table 5.2 for a sampling of student performance on these questions. We believe students will understand P1 and P2 before they understand F1 and F2.

<table>
<thead>
<tr>
<th>Item</th>
<th>Context</th>
<th>Percent correct</th>
<th>Difficulty level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Grades 6-7</td>
<td>Grades 9-10</td>
</tr>
<tr>
<td>P1</td>
<td>Pendulum – answer</td>
<td>34%</td>
<td>66%</td>
</tr>
<tr>
<td>P2</td>
<td>Pendulum – reasoning</td>
<td>29%</td>
<td>57%</td>
</tr>
<tr>
<td>F1</td>
<td>Flies – answer</td>
<td>16%</td>
<td>29%</td>
</tr>
<tr>
<td>F2</td>
<td>Flies – reasoning</td>
<td>16%</td>
<td>17%</td>
</tr>
</tbody>
</table>

Table 5.2. Student performance on two easy and two difficult questions.

One method that can be used to analyze data in studies such as this one is Item Response Theory (IRT), which operates under several assumptions, the most relevant of which is local independence (Hambleton, Swaminathan, & Rogers, 1991). Local independence means that responses to any two items on a given test are statistically independent. Clearly IRT will not apply to an analysis of the Lawson Test as it has a two-tier design and responses to consecutive items are highly dependent. If fact, we rely on the dependency between questions to extract information about student reasoning. Thus, we need to develop a new method to utilize the existing Lawson Test data. We
believe that by analyzing the patterns in student responses, we can identify a learning progression.

Figure 5.1. Items from the Lawson’s Test used in this study.

Any learning progression needs to be verified after it has been developed (Alonzo & Steedle, 2009). Steedle & Shavelson (2009) used latent class analysis to verify a force and motion learning progression. Latent class analysis states that observed variables are dependent on an unobservable (latent) variable (Lazarsfeld & Henry, 1968). For example, items responses are observed variables while ability is the latent variable. According to Lazarsfeld and Henry, some assumptions need to be made about the latent
variable since there is no way that it can be directly measured. A latent variable is
defined, then, by the effects it has on certain indicators. Steedle & Shavelson argue that
this method is appropriate because it assumes each student belongs to a particular latent
class that accounts for that student’s performance patterns. It is also useful because it
provides information about individuals within classes (i.e. ability groupings) and does not
make any assumptions about an existing learning progression. Use of this method relies
on Bayes’ Theorem, so Steedle and Shavelson developed two models: exploratory (made
no learning progression assumptions) and confirmatory (based on their proposed learning
progression). They found that the number of latent classes in the confirmatory model
was fixed by the number of levels in the proposed learning progression. To determine the
effectiveness of their latent class model, they compared the item response learning
progression levels (for example, option “A” indicates level 3, etc.) to the latent classes. If
a lower latent class had a high probability of selecting a low level response, it signaled
that the latent class model was correct.

While latent class analysis is a useful tool in developing a learning progression, there
are multiple reasons why it was not used in this study. First, latent class analysis is not
commonly used for two-tier items or grouped items, which is what our study entails.
Lazarsfeld and Henry (1968) state that latent class analysis is primarily used for
dichotomous (two-response) items. If this method could be used with two-tier items, and
to our knowledge no one has attempted to do so, the process would likely be very
complicated. The statistical work that is done with individual questions is already
complicated, which is a second reason why we are choosing to not use latent class
analysis—it is not easily accessible. The algorithm used in latent class analysis is run through a computer program that provides results; Steedle and Shavelson (2009), for example, estimated their latent class parameters using Markov Chain Monte Carlo methods carried out with Gibbs sampling implemented by WinBUGS version 1.4. The program is something of a black box; the processes that go on are unclear. The method we are using, pattern analysis, is much more intuitive. It is a straightforward method that can be easily implemented since all data organization is done using ordinary spreadsheets. Furthermore, it connects the data to the learning progression more directly; rather than having to interpret latent class analysis matrices, any pattern that is seen is literally the learning progression result.

We will utilize pattern analysis in three different ways by examining (1) a cross-section of all the data with multiple grade levels, (2) transitional behavior with pre- and post-tests for a particular grade level, and (3) the distribution of scores within a given population.

5.4 Data Collection

Data for this study was based on the first data set collected from 2007 to 2009 with students in grades three through twelve in both China and the United States as well as college students from a large Midwestern university in the United States. The results showed that both Chinese and US students have similar reasoning abilities (Bao et al., 2009). Therefore, the data from both countries are combined. The distribution of collected data across grade levels is given in Table 5.3. All students were given enough
time to finish the test. Younger students took 45 to 50 minutes while college students needed about 30 minutes. The Chinese student used a translated Chinese version of the Lawson’s test.

<table>
<thead>
<tr>
<th>Grade</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>College</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>102</td>
<td>336</td>
<td>547</td>
<td>588</td>
<td>868</td>
<td>606</td>
<td>1489</td>
<td>1520</td>
<td>2083</td>
<td>847</td>
<td>1823</td>
</tr>
</tbody>
</table>

Table 5.3. Distribution of collected student data across different grade levels.

5.5 Data Analysis and Results

It is hypothesized that students will be able to provide a correct answer before they can provide correct reasoning. This suggests to us possible levels of student performance when analyzing the four items chosen from the Lawson Test. Additional motivation for defining levels comes from item difficulty; since P1 and P2 are easy questions, students will likely answer those correctly before answering F1 and F2 correctly. Since each of the four items can be answered correctly or incorrectly, there are 16 possible responses total. We group these into six levels of performance based on the ideas above. The groupings are shown in Table 5.4.

Responses are coded using “0” and “1”. There are four items, which corresponds to two groups of two, each group having an answer and a reason component. The two responses for P1 and P2 are listed as the first pair, while those for F1 and F2 are listed as
the second pair. The answer is the first digit in each pair, and the reasoning is the second digit. Thus, the code 00-00 means all responses were incorrect, while a code of 11-11 means all responses were correct. A code of 11-10 would mean that P1 and P2 were correct, F1 (the answer) was correct, and F2 (the reasoning) was incorrect.

Level 1 represents entirely incorrect responses and serves as the lower anchor of a possible learning progression. Level 2 includes all responses with P1 incorrect and P2 correct. If a student cannot answer P1 correctly, we believe any other items answered correctly have a large chance being the results of guessing. These responses are not included in Level 1 for reasons that will be discussed shortly. Level 3 includes responses where P1 is correct and P2 is incorrect. That is, students correctly answer the first item but miss the reasoning. A 10-01 response is included in this level as F2 was likely guessed correctly. A 10-11 response is grouped into Level 4 rather than Level 3 because there is a chance that students can miss P2 while still fully understanding F1 and F2. Also, it would be unlikely that a student could guess correctly on both F1 and F2. Other Level 4 responses involve getting both P1 and F1 correct or getting P1 and P2 correct; it is unclear which of these responses indicates a higher ability, which is why they are included in the same level. Level 5 requires that students answer P1 and P2 correctly and F1 correctly; that is, they fully understand the easy items and are partway to understanding the difficult items. Finally, Level 6 represents entirely correct responses and serves as the upper anchor for our potential learning progression.

Level 2 is comprised mostly of guessing responses. This is not combined with Level 1 because a correctly guessed response may indicate some amount of understanding. For
instance, a student could have eliminated some responses knowing that they are incorrect but still had to guess from those remaining. At the same time, “00” responses may indicate a misconception; students may have some ideas relating to the problem, but their ideas are misconceptions, which is separate from pure guessing. These considerations make it difficult to determine the relative skill requirements of Levels 1 and 2. Future research will be aimed at better defining these levels. For the present study, the levels shown in Table 5.4 will be used.

We can make some rough predictions of the patterns that will be seen in the six levels. Since Levels 1 and 2 indicate relatively low skill levels, the number of students performing at these levels should decrease with age. Level 3 is an intermediate level that should stay about the same for all ages. Levels 4 through 6 indicate meaningful learning, so we expect them to increase with age. Levels 4 should increase more rapidly than Levels 5 or 6.

Table 5.4 shows all responses to Lawson Test items P1, P2, F1, and F2. Scores from these four items were separated from the rest of the items. Students were then grouped by grade level and performance within that grade level. This performance division is based on Lawson’s (1978) results with his original test. When comparing classification of ability based on interview results to scores on the test, he found that those at the lowest level (concrete reasoning) generally scored 0 to 5 points out of 15. Those at the middle level (transitional reasoning) generally scored 6 to 11 points. Those at the highest level (formal reasoning) generally scored 12 to 15 points. These point values represent the lowest 30%, middle 40%, and highest 30% of scores. We divide our populations into
similar percentages while making sure to include all the same scores in the same category. That is, if the lowest 30% includes scores of 0 to 5 but a student with a score of 5 falls into the middle 40%, that student will be grouped with the lowest 30%. For example, in grades 6-7, the division is as follows: lowest 30% (score of 0-5 of 20), middle 42% (score of 6-9), and highest 28% (score of 10-19). The percentages vary slightly in each grade, but we aim at dividing into lower 30%, middle 40%, and upper 30%.

Our goal in looking at Table 5.4 is to compare how students of different ages and abilities respond to the test items in order to find patterns within the responses. We chose to examine grades 6-7, 9-10, and college. By having two years separating each group, any significant changes in performance will be clearer. Since the college-level data came from American students, American student responses are the only ones presented in Table 5.4.

There are three methods of analysis we will use in this study. First, a single population can be divided into performance-based levels (as in Table 5.4) to see how ability affects responses. Second, a cross-sectional study of all students can be examined for changes in responses with age. Third, pre- and post-tests can be examine to see how one year of learning impacts responses. These methods revealed three main results.
### Table 5.4. Responses to Lawson Test items P1, P2, F1, and F2 from grades 6-7, 9-10, and college.

<table>
<thead>
<tr>
<th>Score</th>
<th>N</th>
<th>Level 1</th>
<th></th>
<th>Level 2</th>
<th></th>
<th>Level 3</th>
<th></th>
<th>Level 4</th>
<th></th>
<th>Level 5</th>
<th></th>
<th>Level 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>00-00</td>
<td>00-01</td>
<td>01-00</td>
<td>01-01</td>
<td>00-10</td>
<td>01-10</td>
<td>00-01</td>
<td>01-01</td>
<td>01-10</td>
<td>01-11</td>
<td>10-10</td>
</tr>
<tr>
<td><strong>Grades 6-7</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low 30%</td>
<td>0-5</td>
<td>433</td>
<td><strong>44.5%</strong></td>
<td>11.1%</td>
<td>6.2%</td>
<td>1.8%</td>
<td>10.8%</td>
<td>1.8%</td>
<td>2.8%</td>
<td>0.2%</td>
<td><strong>7.8%</strong></td>
<td>1.6%</td>
</tr>
<tr>
<td>Mid 42%</td>
<td>6-9</td>
<td>616</td>
<td><strong>47.8%</strong></td>
<td>7.3%</td>
<td>5.8%</td>
<td>1.0%</td>
<td>6.6%</td>
<td>0.8%</td>
<td>1.8%</td>
<td>0.0%</td>
<td><strong>8.6%</strong></td>
<td>1.5%</td>
</tr>
<tr>
<td>High 28%</td>
<td>10-19</td>
<td>404</td>
<td><strong>27.2%</strong></td>
<td>5.4%</td>
<td>4.0%</td>
<td>0.0%</td>
<td>4.7%</td>
<td>0.5%</td>
<td>1.2%</td>
<td>0.2%</td>
<td><strong>11.6%</strong></td>
<td>1.5%</td>
</tr>
<tr>
<td><strong>Grades 9-10</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low 33%</td>
<td>0-8</td>
<td>991</td>
<td><strong>32.3%</strong></td>
<td>4.6%</td>
<td>7.8%</td>
<td>1.1%</td>
<td>5.5%</td>
<td>1.9%</td>
<td>1.4%</td>
<td>0.6%</td>
<td><strong>11.3%</strong></td>
<td>2.0%</td>
</tr>
<tr>
<td>Mid 40%</td>
<td>9-13</td>
<td>1211</td>
<td><strong>16.4%</strong></td>
<td>2.4%</td>
<td>2.6%</td>
<td>0.6%</td>
<td>4.5%</td>
<td>0.9%</td>
<td>1.3%</td>
<td>0.9%</td>
<td><strong>11.8%</strong></td>
<td>1.4%</td>
</tr>
<tr>
<td>High 27%</td>
<td>14-20</td>
<td>804</td>
<td><strong>6.0%</strong></td>
<td>0.4%</td>
<td>2.2%</td>
<td>0.7%</td>
<td>2.5%</td>
<td>0.5%</td>
<td>0.4%</td>
<td>0.7%</td>
<td><strong>8.4%</strong></td>
<td>0.6%</td>
</tr>
<tr>
<td><strong>College</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low 29%</td>
<td>0-11</td>
<td>523</td>
<td><strong>27.7%</strong></td>
<td>4.6%</td>
<td>1.1%</td>
<td>0.6%</td>
<td>6.1%</td>
<td>1.9%</td>
<td>2.1%</td>
<td>0.0%</td>
<td><strong>3.4%</strong></td>
<td>0.2%</td>
</tr>
<tr>
<td>Mid 40%</td>
<td>12-16</td>
<td>724</td>
<td><strong>8.8%</strong></td>
<td>1.8%</td>
<td>0.1%</td>
<td>0.3%</td>
<td>3.0%</td>
<td>0.7%</td>
<td>2.3%</td>
<td>0.4%</td>
<td><strong>1.4%</strong></td>
<td>0.7%</td>
</tr>
<tr>
<td>High 31%</td>
<td>17-20</td>
<td>573</td>
<td><strong>1.2%</strong></td>
<td>0.0%</td>
<td>0.2%</td>
<td>0.0%</td>
<td>1.9%</td>
<td>0.3%</td>
<td>1.2%</td>
<td>0.7%</td>
<td><strong>0.2%</strong></td>
<td>0.2%</td>
</tr>
</tbody>
</table>
5.5.1 Result 1: Defining a new level in the scoring of the Lawson Test

As described in the research design section, traditional scoring of the Lawson Test allows for only two levels of performance—both the answer and reasoning need to be correct or no credit is given. We believe this does not accurately reflect the possible levels of student understanding. Students who get just the answer or just the reasoning correct may be at a higher level of understanding than those who get both incorrect. We want to determine if a “10” (answer correct) response is at a higher level of understanding than a “01” (reasoning correct) response. To do so, we examine responses to Lawson Test items P1, P2, F1, and F2.

First, we examine college students’ responses and divide the population into performance-based levels as described above. The performance of the three groups within the college student population is shown in the bottom section of Table 4. We can narrow our focus to compare “01” and “10” responses by looking specifically at two pairs of columns: 01-00 with 10-00 and 11-01 with 11-10. In both pairs, we see that many more students respond “10”, which leads us to believe that responses of “01” could be due to random guessing. We also note when comparing 11-01 to 11-10 that as ability level increases, the number of “10” responses increases while the number of “01” responses decreases. This suggests that “10” responses indicate a higher level of ability than “01” responses. This is an important result because a “10” response is traditionally worth zero points. These results appear to indicate that “10” actually represents a higher level of reasoning and should therefore be worth some credit.
Next, we look at a cross-section of grades 3 through 12. Table 5.5 shows responses to P1, P2, F1, and F2. In comparing the right pair of shaded columns, we see that students answer “10” more frequently as they get older, which corroborates with our result that “10” indicates a higher level of reasoning than “00” or “01”.

Table 5.5. Student performance on P1, P2, F1, and F2 from grades 3 to 12.

<table>
<thead>
<tr>
<th>Grade</th>
<th>00-00</th>
<th>01-00</th>
<th>10-00</th>
<th>11-00</th>
<th>11-01</th>
<th>11-10</th>
<th>11-11</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>38.2%</td>
<td>5.9%</td>
<td>14.7%</td>
<td>3.9%</td>
<td>2.0%</td>
<td>1.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>4</td>
<td>51.5%</td>
<td>9.5%</td>
<td>6.5%</td>
<td>1.8%</td>
<td>0.6%</td>
<td>0.3%</td>
<td>0.0%</td>
</tr>
<tr>
<td>5</td>
<td>43.0%</td>
<td>6.4%</td>
<td>9.0%</td>
<td>7.1%</td>
<td>0.9%</td>
<td>1.1%</td>
<td>0.2%</td>
</tr>
<tr>
<td>6</td>
<td>39.5%</td>
<td>6.0%</td>
<td>10.7%</td>
<td>15.6%</td>
<td>2.4%</td>
<td>1.5%</td>
<td>0.5%</td>
</tr>
<tr>
<td>7</td>
<td>42.2%</td>
<td>5.1%</td>
<td>8.2%</td>
<td>15.7%</td>
<td>2.0%</td>
<td>3.1%</td>
<td>1.3%</td>
</tr>
<tr>
<td>8</td>
<td>25.4%</td>
<td>2.8%</td>
<td>17.0%</td>
<td>25.6%</td>
<td>1.5%</td>
<td>10.1%</td>
<td>2.6%</td>
</tr>
<tr>
<td>9</td>
<td>22.6%</td>
<td>3.8%</td>
<td>12.0%</td>
<td>27.3%</td>
<td>3.2%</td>
<td>9.9%</td>
<td>4.8%</td>
</tr>
<tr>
<td>10</td>
<td>15.2%</td>
<td>4.6%</td>
<td>9.5%</td>
<td>31.1%</td>
<td>2.9%</td>
<td>12.8%</td>
<td>7.4%</td>
</tr>
<tr>
<td>11</td>
<td>12.0%</td>
<td>3.5%</td>
<td>8.5%</td>
<td>32.6%</td>
<td>4.6%</td>
<td>14.5%</td>
<td>10.9%</td>
</tr>
<tr>
<td>12</td>
<td>8.5%</td>
<td>1.8%</td>
<td>4.4%</td>
<td>25.4%</td>
<td>4.1%</td>
<td>23.1%</td>
<td>18.5%</td>
</tr>
</tbody>
</table>

Table 5.6. College student responses to P1, P2, F1, and F2 on pre- and post-tests.

<table>
<thead>
<tr>
<th>Pre-test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.6%</td>
<td>00-00</td>
</tr>
<tr>
<td></td>
<td>to 00-00</td>
</tr>
</tbody>
</table>

|          | to 01-00  |
|          | 01-00     |
|          | 10-00     |
|          | 0.6%      |

|          | to 11-00  |
| 40.0%    | 11-00     |

|          | to 11-01  |
|          | 11-01     |
|          | 1.7%      |

|          | to 11-10  |
|          | 11-10     |
|          | 8.6%      |

|          | to 11-11  |
|          | 11-11     |
|          | 5.7%      |
The left pair of shaded columns in Table 5.5 also provides valuable information. Consider the “01-00” column; there is no pattern to the responses other than a slight decrease in the older grades. This suggests that a “01” response is likely a guess. We can also compare the ratio of percentages in the “10-00” column to the “01-00” column. Between third and seventh grade, the average ratio is 1.6. At the eighth grade, however, there is a dramatic shift in responses, and the ratio jumps to 6.1. This indicates a major learning shift and a possible step in a scientific reasoning learning progression. After the ninth grade (ratio 3.2), the ratio decreases (though it is still higher than in the lower grades). This is due to older students moving to a higher level (i.e. answering P1 and P2 correctly).

Finally, we examine pre- and post-test data from college students. Some students (N=175) were given the Lawson Test before and after taking a college physics course. Results are given in Table 5.6 with the same coding as above. The post-test responses were only taken from the 12.6% and 40.0% represented in the pre-test portion of the table.

P1 and P2 tend to be very easy for college students, so there is little information to be gleaned from studying 00-00 responses on the pre-test. F1 and F2, on the other hand, give college students difficulty, so meaningful comparisons can be made by analyzing the changes in 11-00 pre-test responses. The shaded cells in Table 6 indicate that many students stayed at the 11-00 level, but if learning gains occurred, the students moved into the 11-10 or 11-11 categories more than the 11-01 category. Again it suggests that “10” responses indicate a higher skill level than “01” responses.
To summarize our first result, three forms of analysis suggest that for an item on the Lawson Test, a correct answer with incorrect reasoning indicates a higher skill level than getting both incorrect. This is a construct that is overlooked by traditional two-tier item scoring. There are important educational implications of this result. Students may be performing at better levels than teachers are realizing. By recognizing that providing a correct answer is progress toward full understanding, teachers will know what to look for in their students and will be able to offer proper amounts of encouragement and credit.

5.5.2 Result 2: Combined patterns of responses as indicators for performance levels

The first result indicates that responding to an item with a correct answer and incorrect reasoning represents a different level of understanding than responding to both incorrectly. Knowing that “00”, “01”, and “10” responses cannot be treated equally, we can divide student responses into the six levels shown in Table 4 and described above.

We can see in Table 5.4 that the total distribution is sparse; most cells have a very low percentage of students. By looking for areas with high concentrations of students, we can find meaningful patterns. Columns in Table 4 that show such concentrations have bold text. Note that many of the non-bold columns in Table 5.4 include “01” responses (answer incorrect, reasoning correct). Such responses do not yield any meaningful patterns. They are also more common at younger ages, which provides further evidence that “01” responses indicate guessing.

As described in the first part of the results section, we group responses into six ability levels as shown in Table 4. The highest level is 11-11 while the lowest is 00-00. The
middle levels are ordered based on the reasoning that students will be able to provide a
correct answer before they provide correct reasoning to the same item (our first result)
and that “01” responses are likely due to guessing.

Having grouped possible responses into six levels, scores in each level can be
summed to make the data less sparse and make patterns more visible. This essentially
condenses Table 5.4 to allow for better data analysis. Table 5.7 shows the number of
responses at each level for grades 3 through 12. Figure 5.2 shows a plot of the data
represented in Table 7 to more clearly see any patterns or trends in the data.

Each level in Table 5.7 and Figure 5.2 shows a distinct progression. Level 1 (00-00)
starts with high percentages in the third grade and continually decreases. This is to be
expected; many younger students will answer all four items incorrectly, but as students
get older, they will answer more items correctly.

Level 2 (0x-xx, 00-xx) responses decrease with age. This supports our proposition
that many of the responses in Level 2 reflect guessing. Older students no longer need to
guess, particularly on P1 and P2.

Level 3 (10-0x) shows that percentages remain relatively steady until a significant
jump between seventh and eighth grade (p<0.001, effect size = 0.68), after which we see
a continual decrease. At eighth grade, there is a peak in students getting P1 (the first
answer) right, then the decrease is due to getting P2 (the reasoning) correct as well. This
matches our first result that the answer precedes the reasoning.

Level 4 (11-0x, 10-1x) shows a rapid, steady increase from grades four to six, little
change between sixth and seventh grades, then steady increase (though less rapid) from
grades eight to ten, and finally a decrease from grades eleven to twelve. More and more students are able to correctly answer P1 and P2 as they get older, which accounts for the increases. By twelfth grade, though, students begin to answer F1 correctly as well, which accounts for the decrease.

Level 5 (11-10) shows an increase as age increases. Few young students are able to answer this many items correctly. Around eighth grade is when this level begins to take off, and there is a significant increase (p<0.001, effect size=0.50) between the eleventh and twelfth grades. We do not see a decrease in this level for older students as Level 5 is often the highest level reached.

<table>
<thead>
<tr>
<th>Grade</th>
<th>00-00</th>
<th>00-01, 01-00</th>
<th>01-01, 00-10</th>
<th>01-10, 00-11</th>
<th>01-11</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38.2%</td>
<td>32.4%</td>
<td>17.6%</td>
<td>10.8%</td>
<td>1.0%</td>
</tr>
<tr>
<td>2</td>
<td>51.5%</td>
<td>34.8%</td>
<td>8.6%</td>
<td>4.8%</td>
<td>0.3%</td>
</tr>
<tr>
<td>3</td>
<td>43.0%</td>
<td>33.3%</td>
<td>11.9%</td>
<td>10.6%</td>
<td>1.1%</td>
</tr>
<tr>
<td>4</td>
<td>39.5%</td>
<td>26.5%</td>
<td>12.2%</td>
<td>19.7%</td>
<td>1.5%</td>
</tr>
<tr>
<td>5</td>
<td>42.2%</td>
<td>23.5%</td>
<td>9.7%</td>
<td>20.3%</td>
<td>3.1%</td>
</tr>
<tr>
<td>6</td>
<td>25.4%</td>
<td>11.6%</td>
<td>18.8%</td>
<td>31.5%</td>
<td>10.1%</td>
</tr>
<tr>
<td>7</td>
<td>22.6%</td>
<td>14.1%</td>
<td>13.8%</td>
<td>34.8%</td>
<td>9.9%</td>
</tr>
<tr>
<td>8</td>
<td>15.2%</td>
<td>15.7%</td>
<td>10.5%</td>
<td>38.6%</td>
<td>12.8%</td>
</tr>
<tr>
<td>9</td>
<td>12.0%</td>
<td>11.1%</td>
<td>9.6%</td>
<td>41.9%</td>
<td>14.5%</td>
</tr>
<tr>
<td>10</td>
<td>8.5%</td>
<td>10.2%</td>
<td>5.0%</td>
<td>34.7%</td>
<td>23.1%</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.7. Percentage of grades 3-12 at the six levels of Lawson Test performance.
Level 6 (11-11) shows a slow increase in younger grades and a large increase by twelfth grade. We would expect this behavior; students need strong reasoning skills to answer all four items correctly, and these skills are not developed until the later years of school. The percentages reflect this as it is not until twelfth grade that this level really takes off. Between eleventh and twelfth grades, there is a significant jump in the percentage of students at Level 6 (p<0.001, effect size=0.63).

Figure 5.2. Percentage of grades 3-12 at the six levels of Lawson Test performance.

We can also consider all of the patterns together and the distributions of each level across the grades. In performing a Chi-Square test with Pearson Chi-Square method, we
see a significant difference between each grade level ($\chi^2=2106.448, p=0.000$) with the exception of sixth to seventh grades ($\chi^2=9.765, p=0.082$) which shows borderline significance. This tells us that scientific reasoning learning gains are made at each grade level.

There are also some broader patterns shown in Figure 5.2. There appears to be an important transition point between seventh and eighth grade. Levels 1 and 2 (lower skill) have big decreases and Levels 3, 4, and 5 (higher skill) have big increases between seventh and eighth grade. We believe this is the time when students begin to grasp the ideas in the Lawson Test items selected.

Since Levels 5 and 6 have very low percentages until eighth grade or higher, F1 and F2 appear to be difficult and discriminating questions. They serve as a kind of low-pass filter; they filter out the rapidly changing responses (only a few students with the same answer) while allowing consistent responses (many students with the same answer) to be seen as a pattern.

Table 5.7 and Figure 5.2 give very detailed information, but it may be valuable to take a more general look at the patterns in this data. By grouping students by grades 3-5, 6-7, 8-10, and 11-12 and plotting what percentage of each grade grouping is at a particular level, learning progression patterns can be observed. Grouping the grades will allow us to see more dramatic shifts in performance. We can predict what trends will be seen. Since Levels 1 and 2 represent a lower ability level, there should be a sharp decrease from the first grade grouping to the last. Level 3 is an intermediate level, so we would expect it to remain relatively steady over the years. Levels 4-6 represent high
ability levels, so there should be an increase from the lowest grade grouping to the highest. Figure 3 shows that these expectations are indeed correct.

We can also see that some levels show big jumps between grade groupings. For example, Level 1 shows a large gap between grades 6-7 and grades 8-10. Similar jumps occur between grades 6-7 and 8-10 in Levels 4 and 5 as well as between grades 8-10 and 11-12 in Level 6. The meaning of these jumps is that large learning gains occur between those grades. It is important for teachers to realize when students are making these scientific reasoning developments so that they can promote learning and expect significant changes in their students.
To summarize our second result, examining existing data from the Lawson Test reveals distinct progressions through six different performance levels with a dramatic jump in ability between seventh and eighth grades. Progression through these levels indicates that students will be able to provide a correct answer before they can provide correct reasoning.

5.5.3 Result 3: Proposing a three-level scoring system for the Lawson’s Test

The information available in Table 5.7 and Figure 5.2 is highly valuable, but teachers may not need such detailed information when assessing their students. We want to strike a balance between the complexity of the data analysis done above and the simplicity of grading a multiple choice assessment instrument. This is possible by using a new scoring method for the Lawson Test.

As previously discussed, traditional scoring of the Lawson Test allows for only one point on each pair of items (answer and reasoning). We propose that scoring should be restructured to award points to students who can provide the correct answer but not the correct reasoning. Our first result shows that such responses indicate a higher skill level, so this type of response should be recognized.

There are three ways that a two-tier item could logically be scored. First is the traditional method where both the answer and reasoning need to be correct for credit. Second, each individual item (the answer and the reasoning) could be worth one point, and credit would be awarded for getting either one correct. Third, since “01” responses appear to be due to guessing while “10” responses indicate a higher skill level, a point
could be awarded for responding with the right answer while two points would be awarded for getting both the answer and these reasoning correct. We refer to this last method as a three-level scoring system because it reflects the ability levels established in the previous result: (1) nothing correct and/or guessing, (2) answer correct, (3) answer and reasoning correct. These scoring methods are summarized in Table 5.8.

To establish a base level of validity for either of the proposed methods, we compare the traditional scoring method to the proposed methods using the data from this study. Figure 5.4 shows the three scoring methods applied to P1, P2, F1, and F2. Since individual item and three-level scoring allow for a higher point total, the scores have been scaled appropriately.

<table>
<thead>
<tr>
<th>Answer</th>
<th>Reasoning</th>
<th>Traditional</th>
<th>Individual</th>
<th>Three-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect</td>
<td>Incorrect</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Incorrect</td>
<td>Correct</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Correct</td>
<td>Incorrect</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Correct</td>
<td>Correct</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5.8. Traditional and proposed scoring methods for two-tier items on the Lawson’s Test.

Figure 5.4 shows that the proposed methods are potentially valid. The curves of the individual, three-level, and traditional scoring systems are very similarly shaped. There is, however, a gap between each of the curves that must be taken into account.
First, the gap between either of our two proposed methods and the traditional method is due to the fact that our proposed methods give credit for “01” and “10” responses. Such responses traditionally receive zero points. The gaps reflect students who provide “10” and “01” responses. Second, there is a gap between the individual and the three-level scoring methods. This is because the three-level system eliminates points for guessing (a “01” response) that the individual method includes. Note that the gap between the three-level and individual curves is greater for younger students, particularly for P1 and P2. This is when we see the most guessing. For later grades, when students no longer need to guess on P1 and P2, the gap closes. The curves for F1 and F2 have a larger gap for a longer period of time. These are harder items, so guessing continues into higher grade levels.

From these results, we believe that the three-level scoring system best reflects student ability. There are other reasons for choosing this system. The three-level method allows for better statistical analysis. The average score using traditional method is 21% while the average score using the three-level system is 34%. Having higher average scores allows for better analysis; i.e. when the average score is so low, it is difficult to distinguish between results and noise that may result from student guessing. The lowest traditional score is less than 5%, which is very small, but the lowest score for the individual method is near 20%, which seems too high. Thus, the three-level system with a low score of about 10% seems like the best choice.
Another benefit to our proposed three-level scoring method is that it rewards students who have accomplished something by providing a correct answer. Our results
have shown that these students are at a higher skill level, which is something our assessment should reflect. A three-level scoring system is a more accurate assessment, but it also lets the student know that he/she is doing something right, which can provide motivation and a sense of achievement. Traditional scoring does not recognize the accomplishment of providing a correct answer.

To summarize our third result, traditional scoring of the Lawson Test does not accurately reflect student ability. As indicated by the developmental data on the Lawson’s Test (see Chapter 4), reasoning skills develop slowly, and there is an intermediate level that traditional scoring does not recognize. A three-level scoring system more accurately reflects student ability, allows finer grained data analysis, and provides teachers with a simple way to track progression.

5.6 Conclusions

In this chapter, a data mining method is introduced and with it an analysis of four items on the Lawson Classroom Test of Scientific Reasoning revealed three results. First, we acquired some theoretical insight into student responses; students will be able to answer a question correctly before they can provide the correct reasoning. Second, we established six performance levels based on student responses to the four items. These levels revealed a learning progression with distinct patterns being seen at all levels. Third, we proposed a new scoring method for the Lawson Test. In line with our first result, we believe a three-level scoring system (where students get credit for providing
correct answers with incorrect reasoning) better reflects student understanding and is therefore more accurate in assessment.

All three results are based on a new insight into a learning progression that was obtained via pattern analysis. The primary goal of this study was to explore a method that could be applied to existing Lawson Test data. The pattern analysis method used has proved to be successful. It is an easily accessible method; all data was analyzed using spreadsheets. The data can be intuitively analyzed by looking for patterns based on groupings of questions, and these patterns can be interpreted in a straightforward manner. This method could be employed by teachers in assessing their students and by researchers in developing and verifying learning progressions.

This method of data mining and pattern analysis that we have developed has great potential as vast quantities of previously collected Lawson’s test data can be mined to identify new information about student learning. However, use of this method is context dependent. For instance, a different question design might not allow for levels to be seen as distinctly. The item content and student prior knowledge may affect the patterns. Even additional items on the Lawson Test may not show the patterns that we have seen. Each item, then, deserves a detailed analysis as many factors affect student responses, but this is beyond the scope of this paper. An important message is to not blindly apply this method to any question. No analysis method works perfectly in all situations, and this holds true here.
Future work will involve addressing the issues mentioned above. Additional items on the Lawson test as well as other scientific reasoning assessment instruments should be analyzed for their potential power in defining a scientific reasoning learning progression.
Chapter 6. A Case Study on Fine Grained Learning Progression of Control of Variables

6.1 Context of the Study

Scientific reasoning skills are more and more emphasized in science curricula. This study focuses on a particular skill of scientific reasoning, the control of variables (COV), to identify fine grained learning progression levels, which can inform teachers and researchers for developing and delivering better aligned curriculum and assessment. The main hypothesis of the research is based on observations from our previous research which suggested that when students were given experimental design and evaluation tasks in teaching and assessment of COV skills, the presence of experimental data often trigger students into different mode of reasoning. In particular, when experimental data is given, students in a transitional stage of understanding COV often have a tendency to focus on the plausibility of the experimental data which is related but not directly addressing the COV skills involved in the experimental design method. To quantitative determine these intermediate levels of understanding in manipulating COV conditions, two forms of assessment (providing and not providing experimental data) were developed to probe how students handle data and how context affects performance. The design of the assessment tool can help identify common student difficulties and reasoning patterns at a finer grain size. Results from this study show that (1) students perform better when no
experimental data is provided, (2) students perform better in physics contexts than in real-life contexts, and (3) students potentially have a tendency to equate non-influential variables to non-testable variables. Additional analysis begins to reveal a possible progression of different levels of control of variables skills. The new form of assessment design developed in this study provides a practical means for researchers and teachers to evaluate student learning progression on control of variables.

6.2 Review of Research on Control of Variables

Physics courses provide opportunities to teach scientific reasoning and the American Association of Physics Teachers has laid out goals for physics education that reflect this fact; categories include the art of experimentation, experimental and analytical skills, conceptual learning, understanding the basis of knowledge in physics, and developing collaborative learning skills (Boudreaux et al. 2008, AAPT, 1998). To better achieve these goals in physics education, an increasing number of reformed physics curricula have been designed with inquiry learning as their focus, which helps students learn both science content and scientific reasoning skills. A non-exhaustive list of such new courses includes Physics by Inquiry (McDermott & Shaffer, 1996), RealTime Physics (Sokoloff, Thornton & Laws, 2004), ISLE (Etkina & Van Heuvelen, 2007), Modeling Instruction (Wells, Hestenes & Swackhamer), and The SCALE-UP (Student-Centered Activities for Large Enrollment Undergraduate Programs) Project (Beichner, 1999; 2008). A common emphasis of these reformed curricula is to engage students in a constructive inquiry learning process, which has been shown to have positive impacts on advancing students’
problem solving abilities, improving conceptual understanding, and reducing failure rate in physics courses. Most importantly, the inquiry-based learning environment in these reformed courses offers students more opportunities to develop their reasoning skills; these opportunities are otherwise unavailable in traditionally taught courses (Etkina & Van Heuvelen, 2007; Beichner & Saul, 2003).

Since scientific reasoning is increasingly emphasized in reformed physics courses, it is important to understand how and why students may be struggling with specific skills in scientific reasoning. For both research and teaching purposes, we need a good knowledge base and assessment tools regarding student difficulties in specific aspects of scientific reasoning. Unfortunately, there has been limited research on scientific reasoning in the context of learning physics. Resources on assessment of specific scientific reasoning skills using physics contexts are also scarce. Instead, research and assessment tools have been more focused on student learning of scientific information, and teachers are often inexperienced in assessing student performance on abilities and skills underlying the surface level content knowledge (Yung, 2001; Hofstein & Lunetta, 2004).

Among the different dimensions in scientific reasoning, control of variables (COV) is a core construct supporting a wide range of higher-order scientific thinking skills. COV is also an important skill fundamental to understanding physics concepts and experiments. In a recent study, Boudreaux et al. found that college students and in-service teachers had difficulties with basic methods in COV which included failure to control variables, assuming that only one variable can influence a system’s behavior, and rejection of entire sets of data due to a few uncontrolled experiments (Boudreaux et al. 2008). Boudreaux et
al. concluded that students and teachers typically understand that it is important to control variables but often encounter difficulties in implementing the appropriate COV strategies to interpret experimental results.

As a fundamental construct in scientific reasoning, control of variables has been heavily researched by cognitive scientists for more than a decade (Chen & Klahr, 1999; Toth, Klahr & Chen, 2000; Kuhn & Dean 2005; Kuhn, 2007). Their studies have typically centered on the scientific reasoning skills (specifically COV) of elementary school students. The recent work by Boudreaux et al. focused on college students’ and in-service teachers’ understanding of COV in physics contexts (Boudreaux et al., 2008). The existing research has revealed a rich spectrum of COV skills from simple tests of COV conditions to complex tasks involving multi-variable controls and causal inferences from experimental evidence.

For example, in examinations of simple COV skills, researchers used simple experiments involving few variables (Chen & Klahr, 1999; Toth, Klahr & Chen, 2000). Second through fourth grade students were presented with a pair of pictures and asked to identify whether they showed a valid or invalid experiment to determine the effect of a particular variable. Chen and Klahr (1999) found that elementary students are capable of learning how to perform COV experiments. Students as young as second grade were able to transfer their COV knowledge when the learning task and the transfer task were the same.

To study more complex COV constructs, Chen and Klahr asked students to design experiments involving a ball rolling down a ramp to test a given variable and then state
what they could conclude from the outcomes (Chen & Klahr, 1999). With increasing complexity by involving more variables in contexts of ramps, springs, and sinking objects, Penner and Klahr (1996); Toth, Klahr, and Chen (2000) had students design and conduct experiments, justify their choices, and consolidate and summarize their findings. In the context of sinking objects, researchers also probed student understanding of multi-variable influence by asking students what combination of variables would make the fastest sinking object. They found that older students (14-year-olds) performed better than younger students (10-year-olds).

Kuhn focused more on high-end skills regarding students’ abilities in deriving multi-variable causal relations (Kuhn, 2007). Kuhn had fourth-graders use computer software to run experiments relating to earthquakes (and ocean voyages). This study used more variables than previous studies mentioned and asked students to determine whether each variable was causal, non-causal, or indeterminate. Identifying a causal variable is an intermediate level task, but identifying a non-causal variable is higher on the spectrum of skills because students do not always realize that one can test something even if it does not influence the result. Kuhn found that students made progress in learning COV skills despite lacking direct instruction. Nevertheless, students struggled when faced with handling multivariable causality.

In testing the high-end skill of understanding a multi-variable context, researchers also found inconsistencies in student reasoning, especially with transfer tasks relating to COV; students sometimes described what they thought would be the cause of an outcome using descriptors that did not match their experimental results. For example, a student
might describe a certain material as being necessary even though they did not mention it during experimentation. Chen and Klahr noted that students can learn how to do COV experiments but will often deem them unnecessary during transfer tasks (Chen & Klahr, 1999). Kuhn saw that students could correctly design experiments but did not have a good method for handling multivariable causality (Kuhn, 2007).

Turning to older and more educated subjects, Boudreaux et al. studied college students’ and in-service teachers’ understanding of COV. They observed three distinctive abilities at different levels of complexity (Boudreaux et al. 2008). The first and simplest level was the ability to design experimental trials. To test this, students were given a set-up with a specific set of variables and were asked to design an experiment that could test whether a particular variable influenced the outcome and explain their reasoning. The second level was the ability to interpret results when the data warrant a conclusion. Students were presented with a table of trials and data from a COV experiment and asked whether a given variable influenced the behavior of the system. The third level was the ability to interpret results when the experimental design and data do not warrant a conclusion. In this case, students were provided with a table of trials and data that did not represent a COV experiment and were asked if a given variable is influential. Boudreaux et al. showed that students often have more difficulty interpreting data with an inconclusive relation than with a conclusive one (Boudreaux et al. 2008).

From these studies, we can see a rich spectrum of scientific reasoning skills relating to COV in which there is also a possible developmental progression from simple control of a few variables to complex multivariable control and causal analysis. The structure of
the complexity can come from several categories of factors including the number of involved variables, structures and context features of the problem or task, the type of embedded relations (conclusive or inconclusive), and control forms (testable and non-testable).

| Low-end skills | • Identifying or recognizing a COV situation  
• Designing a COV experiment to test a possible causal relation |
| Intermediate skills | • Deciding whether an experimental design involving multiple variables (>2) is a valid COV test of selected variables  
• Deciding, given an experimental design, whether a test is NOT a valid COV test  
• Inferring from experiment results and designs that a variable, among several, is causally related to the outcome |
| High-end skills | • Inferring from experiment results and designs that a variable is testable in the design when it is non-causal  
• Being able to reason through experiments and hypotheses by manipulating an integrated network of multivariate causal relations |

Table 6.1. A summary of different levels of COV skills studied in the literature

Table 6.1 gives a compact list of several different levels of COV skills that have been commonly studied in the existing literature. However, the existing work cannot provide a complete metric to place the different skills in terms of their developmental levels. This
is because only subsets of these skills were researched in individual studies, which makes it difficult to pull together a holistic picture of the developmental progression of the different skills. In this study, we designed an experiment that can probe all the related skills in a single study, allowing us to more accurately map out the relative difficulties of the different skills and investigate how the difficulty of COV tasks is affected by task formats, contexts, and tested relations.

The results of this study can advance our understanding in one fundamental scientific reasoning element that is central to the hands-on and minds-on inquiry-based learning method. Researchers and teachers can gain insight into typical patterns of student reasoning at different developmental stages. The assessment method can also directly facilitate teaching and research in science courses.

6.3 Research Design

6.3.1 Research Questions and Goals

Built on the existing research, we have conducted a study to further investigate student difficulties regarding the understanding and application of COV in physics and real world contexts. In particular, this research aims to (1) identify at finer grain sizes common student difficulties and reasoning patterns in COV under the influence of question context and difficulty, (2) study if student difficulties reveal a developmental progression from naïve to expert learners, and (3) develop a practical assessment design for evaluating student ability in COV.
6.3.2 The Design of the Assessment Instrument

As described in the previous section, Boudreaux et al. did important work in identifying several common difficulties relating to COV (Boudreaux et al. 2008). For our study, we modified two questions from Boudreaux et al.’s study and included one additional question. These alterations aim to identify student levels of understanding, see if there is a progression through these levels, and determine how context plays a role. The instruments used in this study include two tests, each of which contains three questions on COV experiments. The questions in the two tests are identical except that in one test all questions provided experimental conditions alone while in the other test all questions provided both experimental conditions and data. Figure 6.1 shows one form of the test with three test questions all containing experimental data. The format of the test without data was the same as Figure 6.1 except the rows of data were removed. Table 6.2 outlines the details of the questions.

<table>
<thead>
<tr>
<th>Tests</th>
<th>Item</th>
<th>Context</th>
<th>Posed Question</th>
<th>Correct Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version A</td>
<td>Fishing</td>
<td>Real-life</td>
<td>Can a named variable be tested?</td>
<td>Named variable cannot be tested</td>
</tr>
<tr>
<td>(data not given)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Version B</td>
<td>Spring</td>
<td>Physics</td>
<td>Which variables can be tested?</td>
<td>Two variables can be tested; one is influential, one is not</td>
</tr>
<tr>
<td>(data given)</td>
<td>Pendulum</td>
<td>Physics</td>
<td>Which variables can be tested?</td>
<td>No variables can be tested</td>
</tr>
</tbody>
</table>

Table 6.2. Information about test items
1. Tom, Jerry and Dan are good friends and go fishing together most weekends. They often use the same type of fishing tools and have similar skills in fishing (that is, they often each catch a similar number of fish every time). On their last fishing trip, they had selections of different fishing rods and fishhooks, and they each picked a different location to fish (see the conditions given in the table below). They fished for a total of two hours and the number of fish they each caught during this period is shown below.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Tom</th>
<th>Jerry</th>
<th>Dan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fishing rods</td>
<td>Long</td>
<td>Long</td>
<td>Short</td>
</tr>
<tr>
<td>Fishhooks</td>
<td>Thick</td>
<td>Thin</td>
<td>Thin</td>
</tr>
<tr>
<td>Locations</td>
<td>Point A</td>
<td>Point A</td>
<td>Point B</td>
</tr>
</tbody>
</table>

| Numbers of fish caught during the two-hour time | 15 | 15 | 8 |

Ignoring all other variables, can the information in the table be used to test whether the thickness of fishhooks affects the number of fish caught?

a. Yes  
b. No

Please write down your reason:

2. A small bob of mass M is suspended from a spring, which can be made to oscillate up and down. The number of complete oscillations during a certain time interval can be counted. A student wants to know whether or not the number of oscillations in 10 seconds is affected by the un-stretched length of the spring, the distance the object is pulled from its balanced position at the time of release, and/or the mass of the bob.

The student carries out several different experiments to investigate what factors affect the number of oscillations in 10 seconds. The conditions and results are shown in the table.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Trial 1</th>
<th>Trial 2</th>
<th>Trial 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Un-stretched length of the spring</td>
<td>27 in</td>
<td>12 in</td>
<td>12 in</td>
</tr>
<tr>
<td>Distance the bob is pulled from its balanced position at the time of release</td>
<td>5 in</td>
<td>3 in</td>
<td>6 in</td>
</tr>
<tr>
<td>Mass of bob</td>
<td>72 g</td>
<td>72 g</td>
<td>72 g</td>
</tr>
<tr>
<td>Number of oscillations in 10 seconds</td>
<td>30</td>
<td>45</td>
<td>45</td>
</tr>
</tbody>
</table>

Ignoring all other variables, which variable or variables can be tested using the information shown in the table above?

a. The un-stretched length of the spring  
b. The distance the bob is pulled from its balanced position at the time of release  
c. The mass of bob  
d. a and b  
e. b and c  
f. a and c  
g. a, b, and c  
h. No variable can be tested using the conditions and results in the table.

Please write down your reason:

3. As shown below, a string that hangs from a bar has a ball attached to its end. The string (and the attached ball) can swing back and forth. The number of complete swings during a certain time interval can be counted. A student wants to know whether or not the number of swings in 10 seconds is affected by the length of the string, the mass of the ball, and/or the angle the string makes with vertical at the time of release.

The student carries out several experiments to investigate what factors affect the number of swings in 10 seconds. The conditions and results are shown in the table below.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Trial 1</th>
<th>Trial 2</th>
<th>Trial 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of string</td>
<td>10 cm</td>
<td>10 cm</td>
<td>40 cm</td>
</tr>
<tr>
<td>Mass of ball</td>
<td>20 g</td>
<td>30 g</td>
<td>30 g</td>
</tr>
<tr>
<td>Angle of release</td>
<td>15°</td>
<td>30°</td>
<td>15°</td>
</tr>
<tr>
<td>Number of swings in 10 seconds</td>
<td>16</td>
<td>16</td>
<td>8</td>
</tr>
</tbody>
</table>

Ignoring all other variables, which variable or variables do you think can be tested using the information shown in the table above?

a. Length of string  
b. Mass of ball  
c. Angle of release  
d. a and b  
e. a and c  
f. b and c  
g. a, b, and c  
h. No variable can be tested using the conditions and results in the table.

Please write down your reason:

Figure 6.1. Test questions on COV with experimental data. Question 1 poses a COV situation using a real-life context. Questions 2 and 3 are in physics contexts and are based on the tasks used in Boudreaux et al. (2008).
There are three main results we expect to obtain from our research design. The first is a comparison between a test where data were given and a test where there were no data provided. That is, we test how the structure of the task influences performance. When data are not given, the test question is basically probing the reasoning on experimental design (determining whether it is a COV experiment), and therefore is a low-end COV skill. On the other hand, when data are given, students can be drawn into reasoning through possible causal relations between variables and experimental outcomes on top of the COV situations. In such cases, students will be engaged in coordinating evidence with hypotheses in a multivariable COV setting that involves a network of possible causal relations, which is a higher level reasoning skill (Kuhn & Dean, 2005; Klahr & Dunbar, 1988). Therefore, we hypothesize that providing experimental data in a COV task will increase its difficulty. The results from this study will allow us to evaluate this hypothesis and develop an understanding of how task format may affect task difficulty in COV situations.

The first result will tell us how providing data influences performance. For the second result, we want to know if context influences how subjects handle data. In our instrument, the fishing question is a real-life context, while the spring and pendulum questions are physics contexts. By contrasting students’ responses and explanations to these questions, we can study if (or the extent to which) student reasoning is affected by contexts.

The third result centers on two threads of COV skills used in determining if selected variables are testable under a given COV condition and, when testable, if selected
variables are causally related to the experimental outcomes. The spring question provided a design for two testable variables (spring length and distance pulled back at release). In Test B, where data were given, one of these variables is influential (influences the experiment outcome) while the other is not.

Boudreaux et al. showed that students performed better on questions that had testable variables (both influential and non-influential) than on questions that had non-testable variables – uncontrolled experiments from which no conclusion can be drawn (Boudreaux et al. 2008). In their design, students were asked if an experiment can “be used to test” whether a variable “influences” the result. The statement is clear to an expert that it probes the testability of variables on whether a relation can be tested with the given experiments. However, novice learners may be distracted by the “influence” component of the problem and misinterpret the question. This is also evident from the results reported by Boudreaux et al. (p.164), which suggest that students tend to intertwine causal mechanisms (influential relations) with control of variables (Boudreaux et al. 2008). Therefore, we believe that the posed questions in Boudreaux et al. contain a subtlety of wording that may confound students’ interpretations of the question and does not allow a clear measure of the ability to distinguish between testability and influence. Moreover, students seem to have a real difficulty with this skill such that they may equate, either explicitly or implicitly, a non-influential variable to a non-testable variable. A careful inspection of results from Boudreaux et al. (p.166) confirms the possibility: when given non-testable conditions, half of the students failed to provide the correct response (non-testable) and most of these students stated that “the data indicate that mass
does not affect (influence) the number of swings”, which is an obvious example of conflating non-influential with non-testable (Boudreaux et al. 2008).

In our study, we make an emphasis on measuring the ability to distinguish between testability and influence, which is considered an important construct fundamental to advanced reasoning in COV situations. Since the testability of variables is solely determined by the experimental design without the need of experimental data, providing and not providing experimental data in questions constitute two stages of measurement on this particular construct. Without experimental data, the task is equivalent to an experimental design that can only test students’ ability in recognizing COV conditions which leads to the conclusion on the testability of variables in a given experiment. With experimental data, the task turns into measuring if students can distinguish between testability and influence, or, in other words, if students can resist the distraction of considering the possible casual relations instead of the testability of such relations. Based on the discussions, it is reasonable to hypothesize that tasks not showing experimental data are easier (measuring basic COV skills) than those showing the outcomes (measuring more advanced COV reasoning).

To summarize, in this study we have employed a unique design of contrasting identical COV tasks in two forms, with and without experimental data. In the tasks, we explicitly ask students which variables can be tested. In addition, the questions used in this study allow multiple answers, in which both influential and non-influential relations were embedded. Therefore, the questions can measure the COV reasoning for simultaneously handling multiple variables and relations. This allows us to probe a more
complete set of skills ranging from a low-end skill such as identifying a COV experiment, an intermediate skill such as identifying a causal variable, and a high-end skill such as identifying a non-causal variable in multi-variable conditions. The data collected from this study will allow us to quantitatively determine the relative difficulty levels of the targeted skills.

6.3.3 Data Collection

The two versions of questions discussed in the previous section were compiled into two tests: Test A contains three multiple choice questions without experimental data, and Test B contains the same three multiple choice questions in the same order but with experimental data given. To obtain more details on student reasoning, we added an open response field after each question labeled with “Please write down your reason”, in which students gave short explanations on their reasoning in solving the questions. Students’ responses to this open-ended question are coded and analyzed with the goal of determining response validity for the questions and probing student reasoning. In particular, we want to find out (1) whether a student attended to the fact that the experimental data were given or not when writing an explanation, (2) what variables a student considered in different contexts, and (3) if a student attended differently to influential and non-influential variables (in the spring question).

The subjects of this study were 314 high school students in tenth grade (with an average age of 16) from two public schools in China, one in the Beijing area with 198 students and the other in the Guangdong province with 116 students. Our test results
(mean scores) show no statistical difference between students from the two schools (p=0.32), and therefore the two groups of students were treated as equivalent and their data were mixed together in our analysis.

At both schools, the two forms of the test, Tests A and B, were equally mixed one after the other and handed out to students in random order. That is, students in a class randomly received either Test A or Test B. Students were given 15 minutes to complete the test, which seemed to be enough time to finish three questions. Most students took less than 10 minutes to complete the test.

6.4 Data Analysis and Results

6.4.1 Impact of giving experimental data in a question on student performance

As outlined in the Research Design section, there are three main results we aim to get from this study. First we examine the difference between student performance on questions with and without experimental data. The scores from all three questions were summed to determine a main effect. Figure 6.2 shows the results. There is a statistically significant difference (p=0.001, effect size=0.38) between the two groups; those taking Test A (data not given) outperformed those taking Test B (data given). The students who were not given data had a mean score of 56% while the students who were given data had a mean score of 37% (N ~ 150 for each group).

Testing how students handle data aligns with our first research goal in trying to identify at finer grain sizes student difficulties with COV. The results show that students perform better on COV tasks when data are not present. Our result is consistent with the
literature, which notes that there is a considerable difference between identifying a COV experiment (a low-end skill by our definition) and coordinating results with an experiment (an intermediate to high-end skill) (Kuhn, 2007). Identifying or designing a simple COV experiment is a basic skill that young students can learn (Chen & Klahr, 1999; Klahr & Nigam, 2004; Dean & Kuhn, 2007). When the question does not provide data and asks if the experiment is a valid COV experiment, the students only need to identify which variables are changing and which are held constant. As evident from students’ free responses, many students clearly attended to features of the variables being changing or held constant across trials and used such features in their reasoning. Over 90% of the students explicitly mentioned the COV strategy in terms of changing and controlling variables. Therefore questions probing only basic COV reasoning are relatively simple to these students.

Figure 6.2. Mean scores on Test A (data not given) and Test B (data given). The error bars (approximately ±0.04) represent the standard error.
Integrating COV into more advanced situations (such as multivariable causal reasoning) represents a higher level ability. Kuhn performed such a study where students were asked to determine the causal nature of multiple variables by using COV experimentation methods (Kuhn, 2007). In Kuhn’s study, COV supports the multivariable causal reasoning but is no longer the only thing students need to consider. The same can be said in our study where students are shown experimental data and asked if an experiment is valid. Students are no longer just deciding if an experiment is valid (even though that is what the question is asking). Rather, we have observed that students tend to go back and forth between the data and the experiment designs, trying to coordinate the two and decide if the variable is influential. In our research design, Test A provides the simple task of identifying a COV experiment, but Test B uses COV as a context for more complicated reasoning. Our data show this to be the case as we see that when students are shown identical questions, those who are also shown data (Test B) perform at a lower level than those who are not shown data (Test A).

The student written responses support this explanation as well. For all questions, we found that student explanations were consistent with their choices and the intended interpretation of the questions. It is observed that when data were absent (Test A), more students appear to reason using the COV strategy. As shown from written explanations, 94% of students taking Test A (data not given) used the COV strategy in their reasoning, although many still produced incorrect answers. One student stated in response to the spring question, “The mass of the bob, which does not change in all three trials, can’t be
tested to see whether it affects the outcome or not. However, both the remaining two variables are changing, so they can be tested.” Another student had a similar comment on the pendulum question, stating that “Every pair of two experiments does not satisfy the rule that only one variable varies while the other variables keep constant. So they cannot be tested.”

On the other hand, in Test B when data were given, many students seemed to base their reasoning on that data. For example, one student explained, “If both the mass of the bob and the distance the bob is pulled from its balanced position at the time of release are the same, and the un-stretched lengths of the spring are different, then the numbers of oscillations that occur in 10s also vary, so the un-stretched length of the spring affects the number of oscillations.” As another student noted, “When the length of the string is kept the same, the number of swings varies with changes in the mass of the ball and the angle at release.” Apparently, students were attending to the experiment data and were using the data in their reasoning, which can sidetrack into considering the casual relations between variables and experimental data rather than the COV design of the experiment. Of the students who took Test B (data given), 59% had similar reasoning in their written responses. In other words, only 41% of students taking Test B used the desired COV strategies in their reasoning, which is less than half of the number of students with correct reasoning taking Test A.

Apparently, students are attracted to the data, which distracts them into thinking about the possible causal relations (influence) associated with the variables instead of the testability of the variables. Based the literature and the results from this study, we
consider that a student is at a higher skill level if he/she can resist the distracter and still perform sound COV reasoning. Using this technique of providing distracters is similar to what has already been widely used by physics teachers and others writing physics tests. The Force Concept Inventory (FCI) for instance tries to distract test-takers by providing answer options that play to common sense (Hestenes, Wells & Swackhamer). A student who does well on the FCI is one who ignores (or considers and rejects) the tempting answers and instead thinks though the problem using sound physics reasoning.

An additional finding is that the new question design method of giving identical questions with and without experimental data seems to work well in probing and distinguishing basic and advanced levels of COV skills in terms of the testability and influence of variables.

6.4.2 Impact of question context on student performance

Our second result compares performance based on the context of the question, which is our third research goal discussed in the introduction. Figure 6.3 shows student performance on real-life context and physics context problems. The real-life context data are from question 1 (fishing); the physics context data are averages of questions 2 (spring) and 3 (pendulum). We continue to see the trend from our first result that students who were not given data (Test A) outperform students who were given data (Test B), but we now turn our focus to the context. On the real-life context problem, those not given data (N=154) had a mean of 58% while those given data (N=149) had a mean of 30%; on
the physics context problems, those not given data (N=149) had a mean of 55% while those given data (N=150) had a mean of 41%.

We can see that students taking Test A (data not given) perform at essentially the same level in both contexts, but students taking Test B (data given) perform significantly better (p<0.05, effect size=0.23) on physics context questions than on real-life context questions. The relative performances on Tests A and B in the two contexts is an important piece of data. In the real-life context, the separation is 28%, while in the physics context, the separation is 14%. The difference between the two separations is statistically significant (p=0.001). Why does student performance vary so much depending on context?

![Figure 6.3. The mean scores on Test A (data not given) and Test B (data given) for each context. The real-life context shows a greater difference between the means of Tests A and B than the physics context. The error bars (±0.04) indicate the standard errors.](image-url)
By carefully studying student written responses, we see that students seem to use different reasoning when looking at a real-life scenario compared to a physics scenario. Two factors appear to impact reasoning. One of these factors is that real-life contexts trigger real-world knowledge, which biases student reasoning. Students may insist on what they believe to be true, regardless of what the question at hand states (Caramazza, McCloskey & Green, 1981). For example, if a student believes firmly that a thin fish hook will work better, he/she will answer the question so that his/her belief remains intact. Boudreaux et al. (p. 166) described this as a “failure to distinguish between expectations and evidence” and noticed that many students had this behavior. In a physics context, prior knowledge is more limited, and the knowledge that students do have is more likely to be something learned formally, not well-established, and less tied with their intuitive beliefs (Boudreaux et al., 2008). Therefore, the physics knowledge is less tempting for students to fall back on, so they may be more likely use their COV reasoning skills to answer the question.

The second factor that appears to influence reasoning when faced with a real-life question is a tendency for students to consider additional variables other than those given in the problem. The written responses to the fishing question make this clear. One student indicated several additional variables he considered in his response, including the types of fish, cleverness of fish, water flow, fish’s fear of hook, difficulty in baiting the hook, and the physical properties of the hook. Note that the only variables provided were hook size, fisherman location, and fishing rod length, and students were instructed to
ignore other variables. It is evident that a familiar real-life context triggers students into considering an extended set of variables that are well associated to the question context through life experience. A total of 17 additional variables (other than those given in the question) were named in the written responses to the fishing question. This can be contrasted with the physics context questions where students indicated far fewer additional variables; only 5 additional variables total were mentioned for the physics context problems (mainly material properties or air resistance).

The consideration of additional variables may be connected to the open-endedness of a question. Compared to a real-life situation, a physics context is often pre-processed to have extraneous information removed, and therefore is more constrained and close-ended. For example, students know through formal education that certain variables, such as color, do not matter in a physics question on mechanics. In this sense, the variables are pre-processed in a way consistent with physics domain knowledge, which filters the variables to present a cleaner, more confined question context. A real-life context problem, on the other hand, is very open-ended. The variables are either not pre-processed or are pre-processed in a wide variety of ways depending on the individuals’ personal experiences. Therefore, there is a richer set of possibilities that can come to the student’s mind for consideration. This makes the task much less confined as there is a more diverse set of variables for the student to manipulate.

Our explanation of open-endedness is similar to what Hammer, Elby, Scherr, and Redish refer to as cognitive framing (Hammer, Elby, Scherr & Redish, 2004). Framing is a pre-established way, developed from life experiences, that an individual interprets a
specific situation. In Hammer et al.’s words, a frame is a “set of expectations” and rules of operation that a student will use when solving a problem that affects how the student handles that problem. This applies to what we see in our study. In a real-life context, the frame a student has is affected by a lifetime of experiences, which explains why so many extra variables make their way into students’ explanations. In a physics context, students are accustomed to using traditionally recognized physics methods, so other thoughts are less likely to be triggered. Typically, physics classes train students to use only the variables given in the problem, which confines the task but also gradually habituates students into a “plug-and-chug” type of problem-solving method limiting their abilities in open-ended explorations.

It can be inferred from the data that context (real-life vs. physics) does have an impact on student reasoning (reflected in the number and types of variables students call on), particularly when the question provides experimental results. This is supported by analysis of student performance and written responses on Test B. However, we do not have enough evidence to clearly determine how context affects reasoning; this is an important topic that warrants future research.

6.4.3 Impact of embedded relationships between variables on student performance

The third result of this study looks at how students handle variables that are testable versus non-testable and influential versus non-influential, which aligns with our first research goal of defining COV reasoning skills and difficulties at smaller grain sizes. The testability and influence of variables are two threads of features and relations in a COV
experiment. The experimental designs provided in the question are enough to determine which variables are testable. In such cases, the students need to identify which experimental trials are controlled or not. If a variable is testable, experimental data provided by the question will help determine if that variable has an influence on the result. As discussed previously, the relationship between these two threads is complicated and needs to be carefully distinguished, particularly when analyzing student reasoning on handling variables. As discussed earlier, we consider that distinguishing between the two threads is an area of difficulty; students may equate non-influential variables with non-testable variables.

Boudreaux et al. studied a related issue, but their design did not distinguish the different types of student reasoning regarding variable testability and influence. In our study, the new features in the question design will allow us to distinguish between testability and influence. For example, the two test forms (giving and not giving data) can help in discriminating advanced COV reasoning from basic COV strategies. In addition, the three questions form a progressive ladder that measures the different levels of related skills. The spring question in Test B (data given) is at the last stage of the measurement scale designed to probe the ability of distinguishing between testability and influence. It uses multivariable conditions and allows multiple responses. Permitting multiple responses also leads to a more open-ended testing approach.

The spring question in Test B probes if students are sensitive to the differences between non-influential and non-testable variables. Among the three variables involved in the spring question, two (un-stretched length and distance pulled from equilibrium) are
testable, and the third (mass) is not. Of the two testable variables, one (un-stretched length) is influential and the other is not. There are three levels of skill tested in this question. The lowest level is to recognize that mass is not a testable variable. Students who succeed on this level have basic COV skills. At the second level, students are able to correctly recognize the influential variable (un-stretched length) as testable. However, these students often miss the non-influential variable that is also testable, indicating a possibility of equating non-testable and non-influential. At the third and highest level, students are able to correctly recognize all testable variables regardless of their influence or non-influence.

Table 6.3 shows that students perform significantly better on the spring question (choice d) when no data are given (p=0.002, effect size=0.37). Among the students taking Test B (data given), 26% chose the influential variable as the only testable variable while only 3% of students taking Test A (data not given) made the same choice. This suggests that students can better pinpoint a testable variable when data are not provided (when there is no interference from considering influence). Therefore, the three questions in Test A constitute the measure for the lower-end basic COV skills.

With Test B, more advanced reasoning can be probed. On the spring question in Test B, a significant portion of students (26%) picked the choice a, which corresponds to the second level of COV skills tested in this question. These students are suspected to have utilized incorrect reasoning in equating non-testable and non-influential variables. The results also show that a little over 1/3 of the students achieved the highest level of COV
skills tested in this question: these students are able to engage correct COV reasoning in complicated multi-variable conditions that include both causal and non-causal relations.

<table>
<thead>
<tr>
<th>COV Skills</th>
<th>Questions (Choice)</th>
<th>Test A (no data)</th>
<th>Test B (with data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deciding if a variable is testable when it is testable</td>
<td>Fishing (a, correct)</td>
<td>58%</td>
<td>30%</td>
</tr>
<tr>
<td>Deciding if any of several variables are testable when all are non-testable</td>
<td>Pendulum (h, correct)</td>
<td>55%</td>
<td>46%</td>
</tr>
<tr>
<td>Being able to decide if any of several variables are testable when some variables are influential (only in Test B)</td>
<td>Spring (a, partially correct)</td>
<td>3%</td>
<td>26%</td>
</tr>
<tr>
<td>Being able to decide if any of several variables are testable when some variables are non-influential (only in Test B)</td>
<td>Spring (b, partially correct)</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>Being able to decide if any of several variables are testable when some variables are influential and some non-influential (only in Test B)</td>
<td>Spring (d, correct)</td>
<td>54%</td>
<td>36%</td>
</tr>
</tbody>
</table>

Table 6.3. Percentage of students responding with selected choices of the three questions on Test A (data not given) and Test B (data given). For the pendulum question, none of the variables were testable, and there is no significant difference between Test A and B. For the spring question, two variables were testable, one of which was influential, and there was a significant difference between Test A and B.

The quantitative data in Table 6.3 points to a group of students who may conflate testability with influence. To find evidence about students’ reasoning, we analyze the written responses to the spring question in Test B, which are truly enlightening. In their explanations, some students explicitly used the fact that a particular variable was not
influential as the reason to not choose the corresponding answer: “The number of oscillations has nothing to do with the distance the bob is pulled from its balanced position at the time of release.” These students often chose the variable that was influential as the only testable variable (choice a). There were also students who clearly stated that non-influential means non-testable: “The mass of the bob and the un-stretched length are kept constant, but the distance the bob is pulled from its equilibrium position at the time of release varies in the second and third trials, and the results in these two experiments do not change. Therefore, we cannot obviously test the influence of the distance the bob is pulled from its equilibrium position on the number of oscillations.”

On the spring question in Test B, 21% of the students showed reasoning similar to the two examples discussed above. Nearly all the students who picked the correct answer (choice d) used the correct reasoning to choose variables that were both influential and non-influential as testable. The 21% using improper reasoning is a large number, but it can go undetected if only one of the two versions of questions were used, as it can appear to be a mistake in COV reasoning rather than an important misconception. By designing two versions of the test, this result can be more clearly revealed by contrasting results between Tests A and B. For example in Test B (data given), over a quarter of students missed the non-influential testable variable, while only 3% of the students taking Test A (data not given) missed the same variable. This result suggests that the new format of giving and not giving data in questions can be a useful assessment design in measuring student reasoning at a finer grain size.
In Table 6.3, the results from the fishing and pendulum questions are also included. The results on the fishing question reflect largely the impact of context features, which have been discussed in the previous section. With the pendulum question, since none of the variables are testable, the interference from the influence relations is less, which resulted in a non-significant difference between Tests A and B (p=0.119, effect size=0.18). It appears that students can identify a non-COV experiment with less distraction from experimental data, which is confirmed by students’ written responses: “There is more than one variable changing in each trial. It can be tested only if one variable varies and other variables keep constant” and “Two variables are different in each pair of experiments, so all of the variables are non-testable.” From the students tested, 41% in Test A and 36% in Test B showed similar explanations.

6.6 Conclusions and Discussions

COV is a very important aspect of scientific reasoning. Current research tends to address broad definitions within COV, but it is also necessary to identify COV skills at a smaller grain size. Based on the literature and our results, we begin to see a progression of different levels of COV skills and their assessment settings, which are listed in Table 6.5. We have made a rough estimate of difficulty and ordered the skills accordingly. The actual order remains up for debate and calls for further research to make mapping the progress from naïve to expert learners possible.

From this study, we found that students perform better on COV tasks when experimental data are not provided. Providing data seems to trigger students into
thinking beyond the testability of the variables and attempting to determine if variables are influential. The reason for this behavior could be that students may mix the concepts of testability and influence; in particular, students seem to have a tendency to equate non-influential variables to non-testable variables.

<table>
<thead>
<tr>
<th>COV Skill</th>
<th>Possible Level</th>
<th>Test Versions and Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deciding if a variable is testable when it is testable (without experimental data)</td>
<td>Low-end</td>
<td>Test A Fishing</td>
</tr>
<tr>
<td>Deciding if a set of variables are testable when none is testable (without experimental data)</td>
<td>Low-end</td>
<td>Test A Pendulum</td>
</tr>
<tr>
<td>Deciding if an experimental design involving multiple testable and non-testable variables (&gt;2) is a valid COV experiment (without experimental data)</td>
<td>Low-end to Intermediate</td>
<td>Test A Spring</td>
</tr>
<tr>
<td>Deciding if a variable is testable when it is not testable (with experimental data)</td>
<td>Low-end to Intermediate</td>
<td>Test B Pendulum</td>
</tr>
<tr>
<td>Deciding if a variable is testable when it is testable and influential in a real-life context (with experimental data)</td>
<td>Intermediate</td>
<td>Test B Fishing</td>
</tr>
<tr>
<td>Deciding if any of several variables are testable when some variables are influential (with experimental data given)</td>
<td>Intermediate</td>
<td>Test B Spring</td>
</tr>
<tr>
<td>Deciding if any of several variables are testable when some variables are non-influential (with experimental)</td>
<td>High-end</td>
<td>Test B Spring</td>
</tr>
</tbody>
</table>

Table 6.5. A progression of COV skills tested in this study.

When a task is perceived as asking for influence (rather than testability), it becomes more demanding as it requires coordinating evidence and hypothesis, which is a high-end
skill that students may not have mastered, despite having basic understanding of the COV method. For example, multivariable causal reasoning can be tested in a COV context, but in this case, COV is not the thought process being tested. Rather, COV needs to be understood first and can then be used as the context for testing another skill. This is supported by Kuhn’s argument that COV is not the only challenge students face during scientific reasoning (Kuhn, 2007). COV does, however, deserve much of the focus Klahr and others (Chen & Klahr, 1999 for example) give it because it is the foundation that supports higher level skills.

Furthermore, it is found that students perform better on physics context questions than on real-life context questions. When faced with a real-life question, students tend to call on additional variables (other than those given) more so than in a physics context. We believe this is influenced by the open-endedness of the questions and how students’ knowledge is pre-processed. Students are trained in answering physics context questions in school, so their knowledge is pre-processed in a way that limits the variables they will consider. Real-life knowledge, on the other hand, is pre-processed by a lifetime of experiences, so there are many more variables for a student to consider.

The instrument developed in this study has helped identify a unique assessment structure – providing experimental data versus not providing data. Our results show that giving data triggers a different thought process, one that can exclude COV knowledge. Thus, a test that provides experimental data will separate out students more as it is fundamentally more difficult. This is important for teachers who want to assess their students’ knowledge. Designing and evaluating COV experiments are low-end skills, but
doing the same tasks when considering experimental data is a high-end skill. Teachers need to be conscious of this and make sure they are writing tests aimed at the proper level. Young students should be tested on low-end skills, so data should not be provided. Older students need to be tested for high-end skills, so data can be provided. The ages at which these skills should be developed have not yet been determined, but doing so is the goal of future research.
As part of the ongoing education research on scientific reasoning, this dissertation project conducted a series of studies on the assessment of scientific reasoning. In current literature, the Lawson’s test of scientific reasoning is the most widely used quantitative tool in assessing scientific reasoning. However, the test’s validity has not been thoroughly studied.

This dissertation project started with an in-depth study to evaluate the validity of the Lawson’s test. The research has shown multiple test design issues with the current version of the Lawson’s test and also suggested ways to improve the instruments. For example, the choices of question 8 seem to have two correct answers, which have resulted in many students scoring low on this question although they have shown correct understanding and reasoning from interviews. Similar issues also exist in several other items. As a result, the Lawson’s test has a low ceiling of approximately 80% even for well-developed learners.

Although there are validity concerns with the Lawson’s test, since it is widely used in almost all of the fields in STEM education, the existing data provides a rich resource for educators and researchers to compare and evaluate their own students and research outcomes. Therefore, a collection of well-established baseline results would greatly add
educators and researchers in using Lawson’s test in teaching and research. To this end, the work discussed in Chapter 4 provides a detailed analysis of the developmental trends of scientific reasoning abilities for US and Chinese students measured with the Lawson’s test.

A logistic model is developed and fitted to the data. The model fits the data reasonably well with a mean score RMSD at the 10% level of the average standard deviations. The parameters obtained from the model fitting provide quantitative measures to compare the overall scientific reasoning ability and the individual skill dimensions. The results show that the Chinese and US students are similar in their total scores on the Lawson’s test but diverge in five out of the six skill dimensions. The actual causes for the differences are still under investigation; however, initial clues suggest that cultural and educational settings of the two countries have a lot to contribute to the differences.

The analysis also provides quantitative metrics for the difficulty levels of the Lawson’s test and the individual skill dimensions. The results show that this test is optimized for assessing high school students.

In most existing research, the Lawson’s test data has been analyzed using the total scores, which significantly limits the capacity of information that can be extracted. The Lawson’s test has a unique design using a two-tiered structure, which can produce a wealth of information regarding students’ abilities in coordinating conclusions and explanations. Such information is not well utilized with the existing score based analysis. The work in Chapter 5 aims at developing a data mining method using combinational
patterns of responses study the Lawson’s test data. The method can help identify fine grained learning progression behaviors of selected scientific reasoning skills.

Four items on the Lawson’s test have been studied using this method, which revealed three results. First, we acquired some theoretical insight into student responses; students will be able to answer a question correctly before they can provide the correct reasoning. Second, we established six performance levels based on student responses to the four items. These levels revealed a learning progression with distinct patterns being seen at all levels. Third, we proposed a new scoring method for the Lawson Test. In line with our first result, we believe a three-level scoring system (where students get credit for providing correct answers with incorrect reasoning) better reflects student understanding and is therefore more accurate in assessment. This pattern analysis method seems to have great potential as vast quantities of previously collected Lawson’s test data can be mined to identify new information about student learning.

As another effort to identify fine grained levels for more precise assessment of student scientific reasoning skills, Chapter 6 presents a case study on the learning progressions of the ability of control of variables. Different from the data mining method, this study uses a randomized testing approach to obtain and validate fine grained level of students’ ability in using control of variables. It found that students perform better on COV tasks when experimental data are not provided. Providing data seems to trigger students into thinking beyond the testability of the variables and attempting to determine if variables are influential. The reason for this behavior could be that students may mix the concepts of testability and influence; in particular, students seem to have a tendency
to equate non-influential variables to non-testable variables. The new form of assessment design developed in this study provides a practical means for researchers and teachers to evaluate student learning progression on control of variables.

Summarizing the overall scope of this dissertation work, the project started with a detailed review of the research in the area of scientific reasoning with an emphasis on its assessment aspect. A significant amount of effort has been devoted to evaluate the validity of the Lawson’s test and to establish a solid baseline of students’ performances on this test. Building on the work with the Lawson’s test, two further studies have been conducted, one on a data mining method and the other on a question design method, pave the ways for the on-going research that moves beyond the Lawson’s test to develop a new generation of assessment instruments on scientific reasoning.
References


Honolulu: Kamehameha Schools, Research & Evaluation.


Appendix A: Group Assessment of Logical Thinking (GALT) (An online version)

Item 1 - Piece of Clay

Tom has two balls of clay. They are the same size and shape. When he places them on the balance, they weigh the same.

The balls of clay are removed from the balance pans. Clay 2 is flattened like a pancake.

Which of these statements is true?

- The pancake-shaped clay weighs more.
- The two pieces weigh the same.
- The ball weighs more.

What is your reason for this Answer?

1. You did not add or take away any clay.
2. When clay 2 was flattened like a pancake, it had a greater area.
3. When something is flattened, it loses weight.
4. Because of its density, the round ball had more clay in it.
**Item 3 - Glass Size**

The drawing shows two glasses, a small one and a large one. It also shows two jars, a small one and a large one.

It takes 15 small glasses of water or 9 large glasses of water to fill the large jar. It takes 10 small glasses of water to fill the small jar.

**How many large glasses of water does it take to fill the same small jar?**

- 4
- 5
- 6
- Other

**How many large glasses of water does it take to fill the same small jar?**

- It takes five less small glasses of water to fill the small jar. So it will take five less large glasses of water to fill the same jar.
- The ratio of small to large glasses will always be 5 to 3.
- The small glass is half the size of the large glass. So it will take about half the number of small glasses of water to fill up the same small jar.
- There is no way of predicting
**Item 5 - Pendulum Length**

Three strings are hung from a bar. String #1 and #3 are of equal length. String #2 is longer. Charlie attaches a 5-unit weight at the end of string #2 and at the end of #3. A 10 unit weight is attached at the end of string #1. Each string with a weight can be swung.

Charlie wants to find out if the length of the string has an effect on the amount of time it takes the string to swing back and forth.

**Where would he hang a 5-unit weight to make the scale balance again?**

- strings #1 and #2
- strings #1 and #3
- strings #2 and #3
- strings #1, #2, and #3
- string #2 only

**What is your reason for this Answer?**

- The length of the strings should be the same. The weights should be different
- Different lengths with different weights should be tested
- All strings and their weights should be tested against all others.
- Only the longest string should be tested. The experiment is concerned with length not weight.
- Everything needs to be the same except the length so you can tell if length makes a difference.
**Item 7 - Squares and Diamonds**

In a cloth sack there are

All of the square pieces are the same size and shape. The diamond pieces are also the same size and shape. One piece is pulled out of the sack

**What are the chances that it is a spotted piece?**

- 1 Out of 3
- 1 out of 4
- 1 out of 7
- 1 out of 21
- other

**What is your reason for this Answer?**

- There are twenty-one pieces in the cloth sack One spotted piece must be chosen from these.
- One spotted piece needs to be selected from a total of seven spotted pieces.
- Seven of the twenty-one pieces are spotted pieces
- There are three sets in the cloth sack One- of them is spotted
- One fourth of the square pieces and 4/9 of the diamond pieces are spotted.
**Item 9 - The Mice**

A farmer observed the mice that lived in his field. He found that the mice were either fat or thin. Also, the mice had either black or white tails.

This made him wonder if there might be a relation between the size mouse and the color of its tail. So he decided to capture all of the mice in one part of his field and observe them. The mice that he captured are shown below.

---

**Do you think there is a relation between the size of the mice and the color of their tails (That is, is one size of mouse more likely to have a certain color tail and vice versa)?**

- [ ] yes
- [ ] no

**What is your reason for this Answer?**

- [ ] 8/11 of the fat mice have black tails and 3/4 of the thin mice have white tails.
- [ ] Fat and thin mice can have either a black or a white tail.
- [ ] Not all fat mice have black tails. Not all thin mice have white tails.
- [ ] 18 mice have black tails and 12 have white tails.
- [ ] 22 mice are fat and 8 mice are thin.
**Item 11 - The Dance**

After supper, some students decide to go dancing. There are three young men: ALBERT (A), BOB (B), and CHARLES (C), and three young women: LOUISE (L), MARY (M), AND NANCY (N).

One possible pair of dance partners is AL which means ALBERT and LOUISE.

**In the box below, list all of the possible man-woman couples of dancers. Only man-woman dance couples are allowed. The first possible couple is done for you.**

Well, that’s it! You may want to carefully review your answers and make sure that you answered ALL the questions. Remember, in order for an item to be scored as correct, both the answer and the reason must be correct.
Appendix B: The Test of Logical Thinking (TOLT)

Questions and Reasoning

A series of eight problems is presented. Each problem will lead to a question. Record the answer you have chosen and reason for selecting that answer.

1. Orange Juice

Four large oranges are squeezed to make six glasses of juice. How much juice can be made from six oranges?

a. 7 glasses  b. 8 glasses  c. 9 glasses  d. 10  e. other

Reason:
1. The number of glasses compared to the number of oranges will always be in the ratio 3 to 2.
2. With more oranges, the difference will be less.
3. The difference in the numbers will always be two.
4. With four oranges the difference was 2. With six oranges the difference would be two more.
5. There is no way of predicting.

2. Orange Juice

How many oranges are needed to make 13 glasses of juice?

a. 6 1/2 oranges  
b. 8 2/3 oranges  
c. 9 oranges  
d. 11 oranges  
e. other

Reason:
1. The number of oranges compared to the number of glasses will always be in the ratio of 2 to 3.
2. If there are seven more glasses, then five more oranges are needed.
3. The difference in the numbers will always be two.
4. The number of oranges will always be half the number of glasses.
5. There is no way of predicting the number of oranges.
3. The Pendulum's Length

Suppose you wanted to do an experiment to find out if changing the length of a pendulum changed the amount of time it takes to swing back and forth. Which pendulums would you use for the experiment?

a. 1 and 4  b. 2 and 4  c. 1 and 3  d. 2 and 5  e. all

Reason
1. The longest pendulum should be tested against the shortest pendulum.
2. All pendulums need to be tested against one another.
3. As the length is increased the number of washers should be decreased.
4. The pendulums should be the same length but the number of washers should be different.
5. The pendulums should be different lengths but the numbers of washers should be the same.

4. The Pendulum's Weight

Suppose you wanted to do an experiment to find out if changing the weight on the end of the string changed the amount of time the pendulum takes to swing back and forth. Which pendulums would you use for the experiment?

a. 1 and 4  b. 2 and 4  c. 1 and 3  d. 2 and 5  e. all
**Reason:**
1. The heaviest weight should be compared to the lightest weight.
2. All pendulums need to be tested against one another.
3. As the number of washers is increased the pendulum should be shortened.
4. The number of washers should be different but the pendulums should be the same length.
5. The number of washers should be the same but the pendulums should be different lengths.

**5. The Vegetable Seeds**
A gardener bought a package containing 3 squash seeds and 3 bean seeds. If just one seed is selected from the package, what are the chances that it is a bean seed?

a. 1 out of 2  
b. 1 out of 3  
c. 1 out of 4  
d. 1 out of 6  
e. 4 out of 6

**Reason:**
1. Four selections are needed because the three squash seeds could have been chosen in a row.
2. There are six seeds from which one bean seed must be chosen.
3. One bean seed needs to be selected from a total of three.
4. One half of the seeds are bean seeds.
5. In addition to a bean seed, three squash seeds could be selected from a total of six.

**6. The Flower Seeds**
A gardener bought a package of 21 mixed seeds. The package contents listed:
3 short red flowers
4 short yellow flowers
5 short orange flowers
4 tall red flowers
2 tall yellow flowers
3 tall orange flowers

If just one seed is planted, what are the chances that the plant that grows will have red flowers?

a. 1 out of 2  
b. 1 out of 3  
c. 1 out of 7  
d. 1 out of 21  
e. other
Reason:
1. One seed has to be chosen from among those that grow red, yellow or orange flowers.
2. 1/4 of the short and 4/9 of the tall are red.
3. It does not matter whether a tall or a short is picked. One red seed needs to be picked from a total of seven red seeds.
4. One red seed must be selected from a total of 21 seeds.
5. Seven of the twenty one seeds will produce red flowers.

7. The Mice
The mice shown represent a sample of mice captured from a part of a field. Are fat mice more likely to have black tails and thin mice more likely to have white tails?

a. Yes
b. No

Reason:
1. 8/11 of the fat mice have black tails and 3/4 of the thin mice have white tails.
2. Some of the fat mice have white tails and some of the thin mice have white tails.
3. 18 mice out of thirty have black tails and 12 have white tails.
4. Not all of the fat mice have black tails and not all of the thin mice have white tails.
5. 6/12 of the white tailed mice are fat.
8. The Fish
Are fat fish more likely to have broad stripes than thin fish?

a. Yes

b. No

Reason:
1. Some fat fish have broad stripes and some have narrow stripes.
2. 3/7 of the fat fish have broad stripes.
3. 12/28 are broad striped and 16/28 are narrow striped.
4. 3/7 of the fat fish have broad stripes and 9/21 of the thin fish have broad stripes.
5. Some fish with broad stripes are thin and some are fat.

9. The Student Council

Three students from grades 10, 11, 12 were elected to the student council. A three member committee is to be formed with one person from each grade. All possible combinations must be considered before a decision can be made. Two possible combinations are Tom, Jerry and Dan (TJD) and Sally, Anne and Martha (SAM). List all other possible combinations in the spaces provided.

More spaces are provided on the answer sheet than you will need.

STUDENT COUNCIL

216
10. The Shopping Center
In a new shopping center, 4 store locations are going to be opened on the ground level.

A BARBER SHOP (B), a DISCOUNT STORE (D), a GROCERY STORE (G), and a COFFEE SHOP (C) want to move in there. Each one of the stores can choose any one of four locations.

One way that the stores could occupy the 4 locations is BDGC. List all other possible ways that the stores can occupy the 4 locations.

More spaces are provided on the answer sheet than you will need.
Appendix C: Lawson’s Class Room Test of Scientific Reasoning
1. Suppose you are given two clay balls of equal size and shape. The two clay balls also weigh the same. One ball is flattened into a pancake-shaped piece. *Which of these statements is correct?*

   a. The pancake-shaped piece weighs more than the ball
   b. The two pieces still weigh the same
   c. The ball weighs more than the pancake-shaped piece

2. *because*

   a. the flattened piece covers a larger area.
   b. the ball pushes down more on one spot.
   c. when something is flattened it loses weight.
   d. clay has not been added or taken away.
   e. when something is flattened it gains weight.

3. To the right are drawings of two cylinders filled to the same level with water. The cylinders are identical in size and shape.

   Also shown at the right are two marbles, one glass and one steel. The marbles are the same size but the steel one is much heavier than the glass one.

   When the glass marble is put into Cylinder 1 it sinks to the bottom and the water level rises to the 6th mark. *If we put the steel marble into Cylinder 2, the water will rise*

   a. to the same level as it did in Cylinder 1
   b. to a higher level than it did in Cylinder 1
   c. to a lower level than it did in Cylinder 1

4. *because*

   a. the steel marble will sink faster.
   b. the marbles are made of different materials.
   c. the steel marble is heavier than the glass marble.
   d. the glass marble creates less pressure.
   e. the marbles are the same size.
5. To the right are drawings of a wide and a narrow cylinder. The cylinders have equally spaced marks on them. Water is poured into the wide cylinder up to the 4th mark (see A). This water rises to the 6th mark when poured into the narrow cylinder (see B).

Both cylinders are emptied (not shown) and water is poured into the wide cylinder up to the 6th mark. *How high would this water rise if it were poured into the empty narrow cylinder?*

- a. to about 8
- b. to about 9
- c. to about 10
- d. to about 12
- e. none of these answers is correct

6. *because*

- a. the answer can not be determined with the information given.
- b. it went up 2 more before, so it will go up 2 more again.
- c. it goes up 3 in the narrow for every 2 in the wide.
- d. the second cylinder is narrower.
- e. one must actually pour the water and observe to find out.

7. Water is now poured into the narrow cylinder (described in Item 5 above) up to the 11th mark. *How high would this water rise if it were poured into the empty wide cylinder?*

- a. to about 7 1/2
- b. to about 9
- c. to about 8
- d. to about 7 1/3
- e. none of these answers is correct

8. *because*

- a. the ratios must stay the same.
- b. one must actually pour the water and observe to find out.
- c. the answer can not be determined with the information given.
- d. it was 2 less before so it will be 2 less again.
- e. you subtract 2 from the wide for every 3 from the narrow.
9. At the right are drawings of three strings hanging from a bar. The three strings have metal weights attached to their ends. String 1 and String 3 are the same length. String 2 is shorter. A 10 unit weight is attached to the end of String 1. A 10 unit weight is also attached to the end of String 2. A 5 unit weight is attached to the end of String 3. The strings (and attached weights) can be swung back and forth and the time it takes to make a swing can be timed.

Suppose you want to find out whether the length of the string has an effect on the time it takes to swing back and forth. Which strings would you use to find out?

a. only one string
b. all three strings
c. 2 and 3
d. 1 and 3
e. 1 and 2

10. because

a. you must use the longest strings.
b. you must compare strings with both light and heavy weights.
c. only the lengths differ.
d. to make all possible comparisons.
e. the weights differ.
11. Twenty fruit flies are placed in each of four glass tubes. The tubes are sealed. Tubes I and II are partially covered with black paper; Tubes III and IV are not covered. The tubes are placed as shown. Then they are exposed to red light for five minutes. The number of flies in the uncovered part of each tube is shown in the drawing.

This experiment shows that flies respond to (respond means move to or away from):

a. red light but not gravity
b. gravity but not red light
c. both red light and gravity
d. neither red light nor gravity

12. because

a. most flies are in the upper end of Tube III but spread about evenly in Tube II.
b. most flies did not go to the bottom of Tubes I and III.
c. the flies need light to see and must fly against gravity.
d. the majority of flies are in the upper ends and in the lighted ends of the tubes.
e. some flies are in both ends of each tube.
13. In a second experiment, a different kind of fly and blue light was used. The results are shown in the drawing.

These data show that these flies respond to (respond means move to or away from):

a. blue light but not gravity  
b. gravity but not blue light  
c. both blue light and gravity  
d. neither blue light nor gravity

14. because

a. some flies are in both ends of each tube.  
b. the flies need light to see and must fly against gravity.  
c. the flies are spread about evenly in Tube IV and in the upper end of Tube III.  
d. most flies are in the lighted end of Tube II but do not go down in Tubes I and III.  
e. most flies are in the upper end of Tube I and the lighted end of Tube II.

15. Six square pieces of wood are put into a cloth bag and mixed about. The six pieces are identical in size and shape, however, three pieces are red and three are yellow. Suppose someone reaches into the bag (without looking) and pulls out one piece. What are the chances that the piece is red?

a. 1 chance out of 6  
b. 1 chance out of 3  
c. 1 chance out of 2  
d. 1 chance out of 1  
e. can not be determined
16. *because*
   
   a. 3 out of 6 pieces are red.
   b. there is no way to tell which piece will be picked.
   c. only 1 piece of the 6 in the bag is picked.
   d. all 6 pieces are identical in size and shape.
   e. only 1 red piece can be picked out of the 3 red pieces.

17. Three red square pieces of wood, four yellow square pieces, and five blue square pieces are put into a cloth bag. Four red round pieces, two yellow round pieces, and three blue round pieces are also put into the bag. All the pieces are then mixed about. Suppose someone reaches into the bag (without looking and without feeling for a particular shape piece) and pulls out one piece.

   ![Diagram of pieces]

   *What are the chances that the piece is a red round or blue round piece?*
   
   a. can not be determined
   b. 1 chance out of 3
   c. 1 chance out of 21
   d. 15 chances out of 21
   e. 1 chance out of 2

18. *because*

   a. 1 of the 2 shapes is round.
   b. 15 of the 21 pieces are red or blue.
   c. there is no way to tell which piece will be picked.
   d. only 1 of the 21 pieces is picked out of the bag.
   e. 1 of every 3 pieces is a red or blue round piece.
19. Farmer Brown was observing the mice that live in his field. He discovered that all of them were either fat or thin. Also, all of them had either black tails or white tails. This made him wonder if there might be a link between the size of the mice and the color of their tails. So he captured all of the mice in one part of his field and observed them. Below are the mice that he captured.

![Mice](image)

*Do you think there is a link between the size of the mice and the color of their tails?*

a. appears to be a link  

b. appears not to be a link  

c. can not make a reasonable guess

20. *because*

a. there are some of each kind of mouse.  

b. there may be a genetic link between mouse size and tail color.  

   c. there were not enough mice captured.  

   d. most of the fat mice have black tails while most of the thin mice have white tails.  

   e. as the mice grew fatter, their tails became darker.
21. The figure below at the left shows a drinking glass and a burning birthday candle stuck in a small piece of clay standing in a pan of water. When the glass is turned upside down, put over the candle, and placed in the water, the candle quickly goes out and water rushes up into the glass (as shown at the right).

![Diagram of experiment](image)

This observation raises an interesting question: Why does the water rush up into the glass?

Here is a possible explanation. The flame converts oxygen into carbon dioxide. Because oxygen does not dissolve rapidly into water but carbon dioxide does, the newly-formed carbon dioxide dissolves rapidly into the water, lowering the air pressure inside the glass.

Suppose you have the materials mentioned above plus some matches and some dry ice (dry ice is frozen carbon dioxide). Using some or all of the materials, how could you test this possible explanation?

a. Saturate the water with carbon dioxide and redo the experiment noting the amount of water rise.
b. The water rises because oxygen is consumed, so redo the experiment in exactly the same way to show water rise due to oxygen loss.
c. Conduct a controlled experiment varying only the number of candles to see if that makes a difference.
d. Suction is responsible for the water rise, so put a balloon over the top of an open-ended cylinder and place the cylinder over the burning candle.
e. Redo the experiment, but make sure it is controlled by holding all independent variables constant; then measure the amount of water rise.

22. What result of your test (mentioned in #21 above) would show that your explanation is probably wrong?

a. The water rises the same as it did before.
b. The water rises less than it did before.
c. The balloon expands out.
d. The balloon is sucked in.
23. A student put a drop of blood on a microscope slide and then looked at the blood under a microscope. As you can see in the diagram below, the magnified red blood cells look like little round balls. After adding a few drops of salt water to the drop of blood, the student noticed that the cells appeared to become smaller.

![Magnified Red Blood Cells](Image)

This observation raises an interesting question: Why do the red blood cells appear smaller?

Here are two possible explanations: I. Salt ions (Na+ and Cl-) push on the cell membranes and make the cells appear smaller. II. Water molecules are attracted to the salt ions so the water molecules move out of the cells and leave the cells smaller.

To test these explanations, the student used some salt water, a very accurate weighing device, and some water-filled plastic bags, and assumed the plastic behaves just like red-blood-cell membranes. The experiment involved carefully weighing a water-filled bag in a salt solution for ten minutes and then reweighing the bag.

*What result of the experiment would best show that explanation I is probably wrong?*

a. the bag loses weight  
b. the bag weighs the same  
c. the bag appears smaller

24. *What result of the experiment would best show that explanation II is probably wrong?*

a. the bag loses weight  
b. the bag weighs the same  
c. the bag appears smaller