2016

The Assessment of Scientific Reasoning Skills of High School Science Students: a Standardized Assessment Instrument

Shane Hanson
Illinois State University, djshaner333@gmail.com

THE ASSESSMENT OF SCIENTIFIC REASONING SKILLS OF HIGH SCHOOL SCIENCE STUDENTS: A STANDARDIZED ASSESSMENT INSTRUMENT

Shane T. Hanson

92 Pages

The main goal of science education has been achieving scientific literacy. However, this has been no easy task considering that scientific literacy has many definitions that involve a plethora of activities. This means that assessing the topic becomes quite challenging, especially if this is done with some sort of overarching instrument. Fortunately, Shamos (1995) has characterized the many dimensions of scientific literacy into three levels. These dimensions can then be assessed individually, making the task of assessment less overwhelming. The highest level, true scientific literacy, contains dimensions discussed in this study that already have individual assessments. Wenning's *Nature of Science Literacy Test* (2006) assesses the dimension of having a proper understanding the nature of science. His *Scientific Inquiry Literacy Test* (2007) assesses the dimension of understanding the scientific processes of knowledge development. The *Lawson Classroom Test of Formal Reasoning* (1978, 2000) and the *Inventory for Scientific Thinking and Reasoning (iSTAR) Assessment* (2013) assess the dimension of using logic for induction and deduction or what can be referred to as scientific reasoning.

The Lawson test and iSTAR assessment were designed to assess six and eight mostly overlapping reasoning dimensions, respectively. When looking at a framework developed by Wenning and Vierya (2015), six to eight reasoning dimensions may not be enough to comprehensively assess scientific reasoning. These authors include 31 scientific reasoning skills in their framework that are organized into six defined categories based on intellectual sophistication. This study was designed to create a test that addresses these 31 skills in order to comprehensively assess high school students in a more systematic fashion.

The final iteration of the test assessed 26 of the 31 skills found in five of the six defined categories of intellectual sophistication. Before the final iteration came to fruition, a bank of test questions and the framework went through a review by five experts. Following the changes made because of this review, a pilot test of 33 questions was administered to high school students in central Illinois. The statistical analysis of this pilot test showed that the test had a mean score percentage well below the ideal 50%, and a KR-20 value considerably lower than the benchmark of .80. In order to increase the performance of the test and move these statistical values to acceptable levels, seven questions were eliminated and 12 questions were replaced or revised. These questions were primarily chosen because of their unacceptable item difficulty indices outside the .40 and .60 range, and point-biserial discrimination indices below the desirable .20 value. A second test of 26 questions reflecting these changes was administered to different high school students in central Illinois. The end result was a test had a mean score percentage relatively close to the ideal 50%, and a KR-20 value higher than the benchmark of .80. By taking these preceding steps of the expert review and administering two rounds of testing to reach the acceptable statistical values, a valid and reliable scientific reasoning

test for high school students that addressed skills above and beyond the dimensions of the Lawson test and iSTAR was created.

KEYWORDS: Assessment, High school science, Scientific literacy, Scientific reasoning, Testing

THE ASSESSMENT OF SCIENTIFIC REASONING SKILLS OF HIGH SCHOOL SCIENCE STUDENTS: A STANDARDIZED ASSESSMENT INSTRUMENT

SHANE T. HANSON

A Thesis Submitted in Partial
Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

School of Teaching and Learning

ILLINOIS STATE UNIVERSITY

2016

THE ASSESSMENT OF SCIENTIFIC REASONING SKILLS OF HIGH SCHOOL SCIENCE STUDENTS: A STANDARDIZED ASSESSMENT INSTRUMENT

SHANE T. HANSON

COMMITTEE MEMBERS:

May Jadallah, Chair

Allison A. Meyer

Carl J. Wenning

ACKNOWLEDGMENTS

The writer wishes to thank several individuals who were an integral part of the development of this thesis and the scientific reasoning test created as part of this study. This needs to begin with the committee members. Carl Wenning was by far the most impactful person from the beginning. He is the one who gave the writer the idea for this thesis. He is the one who gave the writer the most insight on everything surrounding scientific literacy, scientific reasoning, and assessment. He is the one who created the framework that is the basis for the test. Most of all, he deserves at least as much credit as the writer does for the creation of the test questions. Ultimately, the test is as much his as it is the writers. Without May Jadallah's support and guidance throughout the whole process as the committee chair, the thesis would have never reached a published ready form. She kept the writer on track through all of the ins and outs of finishing a thesis. Her commentary from a general thesis writing perspective was extremely helpful. She also added a fresh perspective on various parts of this thesis that would have otherwise gone unnoticed. This resulted in a thesis that had more clarity through format and language. Allison Meyer needs to also be recognized for providing commentary about the thesis that was very beneficial. Like May, Allison added a fresh perspective to parts of the thesis. This made the writer rethink how these parts should be written, which improved this body of work.

The expert reviewers were an integral part of the test development. Lina Avayanti, Rodger Baldwin, Luke Luginbuhl, Rebecca Rosenblatt, and Rebecca Vierya deserve credit for the crucial step of validating the bank of test questions as well as the framework aligned to these questions. Their feedback was comprehensive to say the least and with it the test was fairly sound going into its first round of analysis. The teachers that administered the test to their high school students were also important to the test development. Rodger Baldwin, Kristi Brown, Joe Casey, Jim Covey, Mike Fisher, Kerry Kraus, Luke Luginbuhl, Dave Peeler, Kelli Pochynok, and Ken Reid provided the writer with a sample of over 900 students. This sample size was much needed in order to determine if the test was reliable.

S.T.H.

CONTENTS

TABLES

FIGURES

CHAPTER I

THE PROBLEM AND ITS BACKGROUND

Problem Statement

"It is frequently said that achieving scientific literacy is the main goal of science education" (Wenning, 2006, p. 3). In doing so, educators fulfill their obligation of developing scientifically literate citizens who can better our society. According to the authors of *Science Framework for the 2009 National Assessment of Educational Progress*:

> "Our country has an obligation to provide young people who choose to pursue careers in science and technology with a strong foundation for their post secondary study and work experience. The nation's future depends on scientifically literate citizens who can participate as informed members of society and as a highly skilled scientific workforce, well prepared to address challenging issues at the local, national, and global levels" (National Assessment Governing Board, 2008, p. v).

Unfortunately, achieving scientific literacy has become a daunting task considering that there is a veritable deluge of definitions, and within the science community no consensus exists about the definition (Roberts, 2007). "There are far too many visions at play, ranging from science concepts, processes, the history of science, and the nature of science, to science, society, and technology" (DeBoer, 2000, p. 594).

What makes achieving scientific literacy even more daunting is that it is also referred to as science literacy. According to Roberts (2007), scientific literacy "derives its meaning from the character of situations with a scientific component, situations that students are likely to encounter as citizens," and science literacy looks "inward at the canon of orthodox natural science, that is, the products and processes of science itself" (p. 2). Consequently, assessing scientific literacy, which is the term more consistent with other authors referenced in this study, and measuring progress toward achieving it has become equally daunting.

Currently, no overarching instrument assesses this progress because "a comprehensive assessment instrument would be of unacceptable length" (Wenning, 2006, p. 4). Fortunately, there are dimensions that can be assessed individually. In his publication, *The Myth of Scientific Literacy*, Shamos (1995) characterized the many dimensions of scientific literacy in three levels: cultural scientific literacy, functional scientific literacy, and true scientific literacy. Cultural scientific literacy is the level achieved by most adults who have an understanding of science-based terms through the media. Functional scientific literacy is the level that builds upon cultural scientific literacy, and requires the ability to effectively communicate using the basic terms, concepts, and relationships of science. True scientific literacy is the level that is targeted by this study.

According to Shamos (1995), the dimensions of the "truly" scientifically literate person are: (1) understanding the scientific processes of knowledge development, (2) understanding the importance of observation and experimentation in science, (3) being capable of questioning, (4) using logic for induction and deduction, (5) relying upon

evidence, (6) having a proper understanding of the nature of science, and (7) having a basic understanding of the history, values, and assumptions of science. Each dimension of true scientific literacy and its assessment instrument could be "one part of a potential battery of tests to assess progress toward the more general goal of scientific literacy" (Wenning, 2006, p. 3). In doing so, the framework for this battery of standardized tests would operationally define scientific literacy. More importantly, these tests would then provide the barometer of progress toward achieving this operationally defined goal instead of merely having ideas and notions of such progress.

Wenning has addressed two dimensions of Shamos' true scientific literacy in order to get quantifiable assessments centered around (6) having a proper understanding of the nature of science and (1) understanding the scientific processes of knowledge development. His *Nature of Science Literacy Test* (2006) addressed dimension (6) while his *Scientific Inquiry Literacy Test* (2007) addressed dimension (1). Both of these tests were developed in conjunction with frameworks that operationally define what each measured. Another dimension of scientific literacy that has been addressed by the *Lawson Classroom Test of Formal Reasoning* (1978, 2000) is (4) using logic for induction and deduction or what will be referred to as scientific reasoning, which is the focus of this study. Recently, the *Inventory for Scientific Thinking and Reasoning (iSTAR) Assessment* (2013) was designed "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 test" (Han, 2013, p. 36). Both of these tests are based on lists of mostly overlapping

reasoning dimensions with the aforementioned expansion by iSTAR. These dimensions can be found in Table 1.

Table 1

Reasoning Dimensions of the Lawson Test (1978, 2000) and iSTAR Assessment (2013)

Lawson Test	iSTAR Assessment
Proportional reasoning	Proportions and ratios
Control of variables	Control of variables
Probability reasoning	Probability
Correlation reasoning	Correlational reasoning
Hypothetical-deductive reasoning	Hypothetical-deductive reasoning
Conservation of matter and volume	Deductive reasoning
	Inductive reasoning
	Causal reasoning

One issue with the list of reasoning dimensions is that some of the dimensions require skills that are more math related than science related. Another issue is that the list of reasoning dimensions may not be inclusive enough. With respect to a framework developed by Wenning and Vierya (2015), this seems to be the case. These authors highlight many more scientific reasoning skills that are slotted into defined categories based on intellectual sophistication. These skills range from those considered the most rudimentary to those of a hypothetical scientist. With all of these additional skills that are mainly scientific, there could be a question of whether or not Lawson and iSTAR comprehensively assesses scientific reasoning.

Purpose Statement

The purpose of this study is to create a test that comprehensively assesses the scientific reasoning skills of high school students in a more systematic fashion. This test is based on a modification of Wenning and Vierya's (2015) framework of intellectual

process skills and scientific practices. The framework, revised and extended working closely with Wenning, contains thirty-one scientific reasoning skills defined by six different levels of increasing intellectual sophistication based on Wenning's latest version of the Levels of Inquiry Model of Science Teaching (Wenning & Vierya, 2015). In doing so, students will not only be assessed more extensively overall, but can be placed at a level of intellectual sophistication based on their results. This allows the teacher to not only find out the baseline level of students before instruction, but also find out the level of students reached following instruction. In addition, the assessment and framework can help these teachers exemplify their goals for student learning. More generally speaking, the test provides "important data required for informed decision making, for holding schools accountable for meeting achievement goals, and for determining program effectiveness" (Wenning, 2006, p. 11).

Research Questions

- Can a valid scientific reasoning test for high school science students be created from the defined scientific reasoning skills in Wenning and Vierya's intellectual process skills and scientific practices framework?
- Can a reliable scientific reasoning test for high school science students be created from the defined scientific reasoning skills in Wenning and Vierya's intellectual process skills and scientific practices framework?
- Can a scientific reasoning test for high school science students address skills that go above and beyond the dimensions addressed by the *Lawson Classroom Test of*

Formal Reasoning and *Inventory for Scientific Thinking and Reasoning Assessment*?

CHAPTER II

REVIEW OF RELATED LITERATURE

Assessing Scientific Literacy

A goal of science education for more than a century has been enhancing scientific literacy. This goal formed Dewey's teaching at the turn-of-the-twentieth century, was given a boost by the 1957 launch of Sputnik I by the U.S.S.R., brought to light in 1983 by The National Commission of Excellence in Education in *A Nation at Risk: The Imperative for Educational Reform*, and achieving it seems to be the main goal of science education today (Wenning, 2006). Unfortunately, achieving scientific literacy has become a daunting task considering what is involved. According to the *National Science Education Standards*:

> Scientific literacy means that a person can ask, find, or determine answers to questions derived from curiosity about everyday experiences. It means that a person has the ability to describe, explain, and predict natural phenomena. Scientific literacy entails being able to read with understanding articles about science in the popular press and to engage in social conversation about the validity of the conclusions. Scientific literacy implies that a person can identify scientific issues underlying national and local decisions and express positions that are scientifically and technologically informed. A literate citizen should be able to evaluate the quality of scientific information on the basis of its source and the

> methods used to generate it. Scientific literacy also implies the capacity to pose and evaluate arguments based on evidence and to apply conclusions from such arguments appropriately. (National Research Council, 1996, p. 22)

The authors of *Science for All Americans* also have their definition:

> Science literacy—which encompasses mathematics and technology as well as the natural and social sciences—has many facets. These include being familiar with the natural world and respecting its unity; being aware of some of the important ways in which mathematics, technology, and the sciences depend upon one another; understanding some of the key concepts and principles of science; having a capacity for scientific ways of thinking; knowing that science, mathematics, and technology are human enterprises, and knowing what that implies about their strengths and limitations; and being able to use scientific knowledge and ways of thinking for personal and social purposes. (American Association for the Advancement of Science, 1993, p. 20)

These definitions exemplify the "loaded" (i.e. containing many dimensions) nature of science literacy. To make things more overwhelming, each definition can be considered distinct in the term it is defining. The former uses the term scientific literacy while the latter uses the term science literacy. As was mentioned in the previous chapter, Roberts (2007) defined scientific literacy as more about science related situations while defining science literacy as more about the process of science. Whether this distinction is significant or not, the point that needs to be highlighted once again is that science-related literacy is loaded. This point can be strengthened even further with the fact that there is a veritable deluge of definitions, and within the science education community no consensus

exists about the definition (Roberts, 2007). With that said, scientific literacy will be the term of choice moving forward, because the term is more consistent with other authors referenced in this study.

A deluge of loaded scientific literacy definitions that have no existing consensus should rightfully make one wonder if there is a way to assess the subject. This is an important thought considering that achieving scientific literacy is the main goal of science education, and assessment is the barometer of progress toward this vast and complex goal. Presently, no overarching instrument assesses this progress. Even if such an instrument were to exist, the sheer volume of items necessary to show this comprehensive progress would make the assessment too lengthy and cumbersome for practical use. Fortunately, there are dimensions that can be assessed individually.

In his publication, *The Myth of Scientific Literacy*, Shamos (1995) characterized the many dimensions of scientific literacy in three levels: cultural scientific literacy, functional scientific literacy, and true scientific literacy. The dimensions of cultural scientific literacy are those achieved by most who believe they are reasonably literate in science. The dimensions of functional scientific literacy, reached by 40% of the population, build upon cultural scientific literacy. The dimensions of true scientific literacy contain "the same mental qualities that John Dewey called 'scientific habits of the mind' nearly a century ago and which he proposed to be the main rationale for compulsory science education, a rationale that today is often called critical thinking" (Shamos, 1995, pp. 89-90). This is the level of the "truly" scientifically literate person, which makes up 4% or 5% of the U.S. population. The dimensions for each level can be found in Table 2.

Table 2

Dimensions of Scientific Literacy According to Shamos (1995)

Cultural	Functional	True
- Understand basic background information and vocabulary. - Recognize many of the science-based terms used by popular media.	- Ability to effectively communicate basic terms, concepts, and relationships of science. - Be familiar with simple everyday facts of nature such as the concepts of Earth's orbital and diurnal motion, eclipses of the sun and moon, the sun as a source of energy, the greenhouse effect, the origin of the oxygen we breath, and the effects of pollution.	- Understand the scientific process of knowledge development. - Understand the importance of observation and experimentation in science. - Be capable of questioning. - Use logic for induction and deduction. - Rely upon evidence. - Have a proper understanding of the nature of science. - Have a basic understanding of the history, values, and assumptions of science.

Each dimension, especially those found in true scientific literacy, and the assessment instrument associated with the dimension could be an individual part of a group of tests which assesses progress toward the more general goal of scientific literacy (Wenning, 2006). In doing so, the framework for this battery of tests would operationally define scientific literacy. More importantly, these tests would then provide the barometer of progress toward achieving this operationally defined goal. Recently, there has been some work done to provide a barometer of individual dimensions. Two dimensions that

have been addressed are nature of science and science process with two assessments developed by Wenning (2006, 2007).

The Nature of Science Literacy Test (NOSLiT) is a 35-item assessment instrument "that can be used in part, to measure student understanding of the nature of science and thereby track progress toward the more elusive goal of achieving scientific literacy" (Wenning, 2006, p. 11). The *NOSLiT* was developed using an eight-step process outlined by DeVellis (1991) with the first step being the development of the author's framework that operationally defines what is being measured. The framework addresses the essential understandings about nature of science. These understandings can be found in Table 3.

Table 3

Understandings about Nature of Science According to Wenning (2006)

- Knowledge of the content and history of at least one science discipline
- Knowledge of associated scientific nomenclature
- Intellectual process skills
- Rules of scientific evidence
- Postulates of science
- Scientific dispositions
- Major misconceptions about NOS

The framework was reviewed by several physics teaching majors, scientists, educators, and philosophers of science for completeness, clarity, and to provide a reasonable certainty of validity. An item pool consisting of one or more multiple-choice questions for each of the understandings in the framework was generated for possible inclusion in the assessment instrument. A team of physics teacher education majors reviewed these items for clarity, accuracy, reading difficulty, and redundancy, and aligned each of the items with the framework to ensure coverage and agreement (Wenning, 2006).

Following this review, a pilot test was administered to 386 high school physical science students from central Illinois high schools. The analysis of the test overall included range, mean, standard deviation, KR20 reliability, and mean item difficulty. An analysis was also conducted on each item looking at such things as difficulty index, discrimination index, and suitability of foils. An unacceptably low KR20 reliability coefficient, as well as mean item difficulty that was a bit low was cause for the review and revision of poor performing test items. The pilot test was then administered a second time to 354 of the same high school students who took the initial assessment, and went through the same analysis as previously. This revised version increased the mean item difficulty and KR20 reliability coefficient to acceptable levels. However, some low-performing items were still found during this analysis, and revisions were made after a discussion with experts concerning the issues (Wenning, 2006).

This second revision of the test, now the final version, was administered to 36 in-service high school physics teachers, nearly all from the Chicago metropolitan area. The teachers had an overall mean score of 84.8%, which was considerably higher than the 59.6% score of the high school science students. Also, the teachers had a substantially lower standard error of the mean of 1.91 than the students' standard error of the mean of 2.59. "The fact that experienced teachers have a significantly higher mean score than high school students and a smaller standard error is evidence of construct validity for the test" (Wenning, 2006, p. 12). This meant that the test was measuring the construct it claimed to be measuring (Brown, 2000).

The *Scientific Inquiry Literacy Test (ScInqLiT)*, which addresses science process, is another 35-item assessment instrument that was developed using the same eight-step

process. The author's framework that operationally defines what this test measures is based on his levels-of-inquiry spectrum, which is a systematic approach of students developing "increased understanding by moving through progressively more sophisticated levels of inquiry and carrying out various stages of inquiry repeatedly. As the level of intellectual sophistication required to conduct the various levels of inquiry grows, the locus of control shifts from teacher to student" (Wenning, 2007, p. 22). At the time of the assessment's creation, the levels-of-inquiry spectrum (Figure 1) included, in progressively sophisticated order, discovery learning, interactive demonstrations, inquiry lessons, inquiry labs, and hypothetical inquiry.

Discovery Learning	**Interactive Demonstrations**	**Inquiry Lessons**	**Inquiry Labs**	**Hypothetical Inquiry**
Low	**<=== Intellectual Sophistication ===>**			**High**
Teacher	**<=== Locus of Control ===>**			**Student**

Figure 1. The Levels-of-Inquiry Spectrum. As students become more intellectually sophisticated, the level of inquiry utilized by teachers correspondingly can become more sophisticated. At the same time, the locus of control shifts gradually from the teacher to the student (Wenning, 2007).

Discovery learning has the teacher directing students to make specific observations and guiding them to draw specific conclusions using "funneling" questions (Wood, 1998). Interactive demonstrations still have the teacher directing, but the control shifts slightly to the students, as they are required to make explanations of their observations. Inquiry lessons have the teacher use think aloud protocol to guide the students through various scientific practices. Although the teacher maintains control of equipment and the experiment, students are encouraged through "focusing" questions (Wood, 1998). Inquiry labs have students take greater control of the entire learning process, from answering a series of questions and developing problems, to designing

experimental procedures and drawing conclusions on their own. Hypothetical inquiry has the students in full control as they identify their own problems, develop hypotheses or models, make predictions, conduct experiments or observations, and draw conclusions on the basis of logic using empirical evidence (Wenning, 2007).

The spectrum was tied to increasingly sophisticated scientific inquiry skills that can be found in Table 4. This spectrum and associated inquiry skills have since been expanded to include a real-world applications level, and will be discussed later as it relates to the framework for a new scientific reasoning assessment. Also, this framework will show where each of the skills fall into the spectrum.

Table 4

Scientific Inquiry Skills According to Wenning (2007)

- Identifying a problem to be investigated
- Using induction to formulate a hypothesis or model
- Using deduction to generate a prediction
- Designing experimental procedures
- Conducting a scientific experiment
- Observation, or simulation
- Collecting, organizing, and analyzing data
- Applying numerical and statistical methods
- Explaining any unexpected results
- Using available technology to report, display, and defend results

Following the same expert review of the framework and item pool as the *NOSLiT*, a pilot test was administered to 425 high school physical science students from central Illinois high schools. The analysis of the test included range, mean, standard deviation, KR20 reliability, and mean item difficulty. An analysis was also conducted on each item looking at such things as difficulty index, discrimination index, and suitability of foils. Like the *NOSLiT* pilot, the mean item difficulty for the *ScInqLiT* was a bit low. Some

poor performing test items that had very high or low difficulty and/or small to negative discrimination were either removed or revised. The revised pilot test was then administered a second time to 61 entirely different high school students that were highly motivated and relatively homogeneous, and went through the same analysis as previously. The mean item difficulty actually exceeded the acceptable value, which was expected from a motivated and homogenous group. Following this second pilot study, one item was replaced and several others were revised as part of a final review process. "It is expected that that the finalized version of *ScInqLiT* has increased validity and reliability as a result of these changes" (Wenning, 2007, p. 23).

Assessing Scientific Reasoning

Another dimension that has been addressed Lawson (1978, 2000) and Han (2013) is the use of logic for induction and deduction or what will be referred to as scientific reasoning. "Scientific reasoning is the process by which the principles of logic are applied to scientific processes – the pursuit of explanations, the formulation of hypotheses, the making of predictions, the solutions of problems, the creation of experiments, the control of variables, the analysis of data, the development of empirical laws – all in a logical manner – with the intent of developing meaning" (Wenning and Vieyra, 2015). Like nature of science and scientific inquiry, there has been work done to assess scientific reasoning.

The most used assessment is the *Lawson Classroom Test of Formal Reasoning*. The test, originally developed and validated in 1978, was designed to measure concrete and formal reasoning, be administered to high school and college age students, be easily

scored, use physical materials, require as little reading and writing as possible, and include a large enough number and variety of problems to assure reliability. Fifteen items were selected for the original that required "the isolation and control of variables, combinatorial reasoning, probabilistic reasoning, and proportional reasoning…In addition, one item involving conservation of weight (Piaget & Inhelder 1962) and one item involving displaced volume (Karplus & Lavatelli 1969)" (Lawson, 1978, p. 12). Each item involved a demonstration with physical materials that was used to either pose a question or call for a prediction. Students responded in booklets that contained the questions and possible answers, and were instructed to choose the best answer and then explain why they chose the answer (Lawson, 1978).

Lawson administered the test to 513 students in eighth, ninth, and tenth grades. A subgroup of 72 students were randomly selected and individually administered a battery of Piagetian tasks in individual interviews in order to determine if the group test results correlated with interview data. Three types of evidence were sought to assess validity of the test. The first type of evidence involved a panel of six judges with Piagetian research expertise responding with 100% agreement that the test items appear to require concrete and/or formal reasoning. The second type of evidence involved a parametric statistics and principal components analysis of the relationship between the test total scores and the level of subject response on two of the interview tasks, which showed a correlation of 0.76 that was statistically significant ($p < 0.001$). The third type of evidence involved a principal components analysis of the relationship between the test and all four interview tasks, which, instead of yielding the expected two principal factors of concrete reasoning and formal reasoning, showed that three principal factors accounted for 66% of the total

variance. The third factor was identified as early formal reasoning and could be considered intermediate. Overall, the results supported the hypothesis that the test measures aspects of formal and concrete reasoning as well as the intermediate early formal reasoning (Lawson, 1978).

Lawson designed the test so that teachers and/or researchers could classify student performance into development levels. He used four types of information to create this classification scheme. The first type was a comparison of test scores with concrete, transitional, formal responses on two of the interview tasks. The second type was knowledge gained through previous investigations that gave insight into what the test items were measuring. The third type was an item analysis of the test items. The fourth type was the principal components analysis. Of these types of information, the first brought forth evidence of note with a significant relationship between the test scores and summed interview task scores. Because the scores on the interview tasks reflected concrete, transitional, and formal reasoning, it could be seen from the analysis that the majority of the 72 subgroup students who scored 0-5 were those classified as concrete, 6-10 were those classified as transitional, and 11-15 were those classified as formal. The other three types of information all suggested that this scoring scheme was reasonable. For those that did not fit the scheme, the data showed that the test underestimated more than overestimated the abilities of more students. Using this scheme for the 513 students, it was found that 35.3% responded at the concrete level, 49.5% responded at the transitional level, and 15.2% responded at the formal level. Overall, Lawson concluded that the parameters measured by the Piagetian interview tasks were also measured by the test items with a fairly high degree of reliability.

There has been research, although minimal, to assess the performance of Lawson's test since it was published. One such study by Pratt and Hacker (1984), who administered the test to 150 students to examine its construct validity within the context of a unidimensional trait model. Critical of Lawson's use of factor analysis, the researchers chose the unidimensional trait model because it was better suited to testing a single factor hypothesis. In doing so, they concluded that the test failed to reflect the unitary nature of formal reasoning. In other words, the test measured several factors instead of just formal reasoning.

Lawson's test was updated in 2000 changing to a completely multiple-choice format with no demonstrations. Instead of fifteen items, the 2000 test contained twelve items in question pairs, totaling twenty-four questions. The score for this version was a count of the number of questions answered correctly as opposed to the number of items where the question and follow-up explanation of the original version had to be answered correctly. Seven items were carried over from the original with the only difference being the follow-up explanation in the 2000 version was multiple-choice. Three other items followed this format, but involved correlational instead of combinatorial reasoning found in the original. The last two items introduced hypothetical-deductive questions. One of the items contained a first question about experimental design and a second question about what outcome would refute a stated hypothesis. The other item had both questions concerning what experimental results would refute a stated hypothesis (Lawson, 2000).

Unlike the 1978 Lawson test, the 2000 version "was not presented as part of a formal study proving its efficacy, instead resting on its laurels of its earlier incarnation" (Han, 2013). In response to this issue, Han (2013) performed a data-driven study on the

validity of the test. The data was collected in three forms. The first form provided quantitative data of 3rd grade to graduate level students, which indicated issues with questions that would be investigated further. Three item pairs, including two that were not part of the 1978 version, from this analysis showed abnormal results. A high percentage of students scored correctly on the first question of the pair, but a low percentage scored correctly on the second question that required students to explain their reasoning for the first question.

The second form of data collection provided quantitative data of college freshman students that indicated inconsistencies with item pairs. This involved an analysis of two-tiered response patterns of the item pairs. This analysis showed the percentage of students who either (1) responded incorrectly on both questions; (2) responded correctly on both questions; (3) responded incorrectly on the content question, but with correct reasoning; or (4) responded correctly on the content question, but with incorrect reasoning. Patterns (3) and (4) were relatively low for most questions, which means both content and reasoning parts were consistent. However, these patterns were much more prevalent in the three problematic item pairs found in the first form of data collection, implying that there may be a problem in the question design (Han, 2013).

The third form of data collection provided qualitative data of college freshman science and engineering majors from the same pool who were asked to provide open-ended reports on their reasoning to each of the test questions. Also, a subgroup of these students were asked in a follow-up interview to go over the test after completing it and explaining their reasoning on how they solved each of the questions. This was done in order to further validate that the high percentage of patterns (3) and (4) in the three

problematic item pairs were caused by question design. One of the items pairs was on proportional reasoning while the other two pairs were on correlational reasoning. Through the analysis of the proportional reasoning item pair, it was concluded that the question wording was problematic which had an adverse impact on the validity of the assessment. Through the analysis of the correlational reasoning item pairs, the conclusion was that the choices of the reasoning portion of the questions could cause significant uncertainties among students and needed to be reworked. Also, the graphical representations of the questions needed to be improved. The qualitative data collection was also done out of concern that the two-tier format of the test would allow students to answer a question correctly without real understanding of what is being measured. Because about 10% of the students changed their answer to one item pair question after reading the other or their answer to both item pair questions after reading other item pair questions, it was concluded that something in the questions cued the student into finding the answer to other questions. As a result, these questions may be measuring simple logic instead of scientific reasoning (Han, 2013).

In response to the issues with the Lawson test found in Han's data-driven study as well as addressing the need to fully assess students' scientific reasoning ability and provide fine-tuned guidance for teachers, Han and his research team created the *Inventory for Scientific Thinking and Reasoning (iSTAR) Assessment*. The research team identified eight dimensions for reasoning, which was an expansion of the six dimensions assessed by Lawson. To see how these dimensions are defined and compared using examples from each test, refer to Table 5.

Table 5

Reasoning Dimensions of the Lawson Test (1978, 2000) and iSTAR Assessment (2013) with Definitions and Examples

Dimension	Definition	Lawson Example	iSTAR Example
Control of variables	Determining which variables influence the outcome by changing the variable of interest while controlling all other variables (Han, 2013).	Three strings with weights at the end are hung from a bar. Two of the three strings are the same length, and two of the three weights are the same. The strings and weights are chosen to find out if the length of the string affects the time of a swing.	A student wants to know if coffee grounds are good for plants. An experiment is done on two similar plants. One plant is put in sunlight, soil, water, and coffee grounds. The set up for the other plant is chosen.
Proportional reasoning	Using the equality of two ratios (a/b = c/d) to solve for a term when given the other three terms (Han, 2013).	Two cylinders of different diameter are filled with the same amount of water. The narrow cylinder is 2/3 the diameter of the wide cylinder. Knowing how much the water rises in one cylinder, the level to where the water rises in the other cylinder is chosen.	A certain number of bottles of orange juice fill a certain number of glasses. The number of glasses filled by a different number of bottles is chosen.
Probability reasoning	Determining the fraction of the times an event will occur as the outcome of some repeatable process when that process is repeated (Han, 2013).	Pieces of wood of various shape and color are put into a bag. The chance of particular shape and color piece being pulled out of the bag is chosen.	Nine students from three different grade levels are randomly picked. The chance that two specific students will be two of three members picked for the committee is chosen.

Table Continues

Dimension	Definition	Lawson Example	iSTAR Example
Correlational Reasoning	Determining the strength of mutual or reciprocal relationships between variables (Lawson, Adi, and Karplus 1979).	A collection of mice that are either big or small and have either white or black tails. The link between mouse size and tail color is chosen.	A collection of apples that are either big or small and either red or yellow. The link between apple size and color is chosen.
Deductive Reasoning	Drawing a conclusion from premises (Han, 2013).		A pattern is noticed in a card game where any card with an even number is gray on the other side and any card with an odd number is white on the other side. Four cards are shown that have each trait, and which cards that need to be turned over to see if the pattern is true is chosen.
Inductive Reasoning	Drawing a conclusion from particular cases (Han, 2013).		Various combinations of three ants that are either from the same colony (get along well) or from different colonies (fight each other) are shown. The combinations that have ants from all different colonies are chosen.
Causal Reasoning	Establishing the presence of causal relationships among events, which leads to the belief that events of one sort (the causes) are systematically related to events of some other sort (the effects) (Han, 2013).		A possible link between forest fire recovery and wild wolf population was noticed due to an increase of wolves being spotted. After tourists were encouraged to report their spotting of wild wolves, incidents went up four times. The reason for wolf population increase is chosen.

Table Continues

Dimension	Definition	Lawson Example	iSTAR Example
Conservation of weight and volume	The ability to retain the knowledge that although the appearance of an object is changed, certain properties of an object remains the same (Siegal, 2003).	Two clay balls begin with equal size and shape, and then one ball is flattened into a pancake shape. The relative weight of the two balls is chosen.	

Note. Examples were left blank for dimensions not assessed.

According to Han (2013), the assessment is designed "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 test." This includes "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" (p. 36). The result is an assessment that contains 21 items. Like Lawson's 2000 version, the iSTAR is a completely multiple-choice format in which the score is count of the number of questions answered correctly. However, gone is the two-tier format where the question and answer to one part of the pair is so reliant on the question and answer to the other part of the pair. The two-tier items are now replaced with items that contain anywhere from one to three questions with all questions having no bearing on another question in each item. Although this would appear to be an improvement on the Lawson test, there has yet to be any published research about the iSTAR questions, so it remains to be seen if these questions perform any better.

With the introduction of an expanded dimension set, a logical step toward fully assessing scientific reasoning has been taken. However, it appears that there are many scientific reasoning skills beyond these dimensions that can be assessed. A framework developed by Wenning and Vierya (2015) contains these skills. These authors highlight many skills and practices slotted into defined categories based on intellectual sophistication. These skills and practices range from those considered the most rudimentary to those of a hypothetical scientist.

Framework for Scientific Reasoning

Wenning and Vierya's framework includes intellectual process skills and scientific practices that are categorized into increasing levels of intellectual sophistication and tied to the levels of inquiry found in Figure 1 (Wenning, 2007). Furthermore, each level is loosely connected to Bloom's Taxonomy of Educational Objectives to help substantiate why each skill falls into the level. Like the levels of inquiry, Bloom's Taxonomy contains levels of objectives that move from lower to higher intellectual sophistication: remembering, understanding, applying, analyzing, evaluating, and synthesizing.

The first level includes rudimentary skills and practices that are most closely tied to discovery learning, and loosely connected with remembering in Bloom's Taxonomy. These skills and practices are promoted and developed as students generate concepts on the basis of first-hand experiences (a focus on active engagement to construct knowledge). The second level includes basic skills and practices that are most closely tied to interactive demonstrations, and loosely connected to understanding in Bloom's

Taxonomy. These skills and practices are promoted and developed as students engage in explanation and prediction-making that allows teachers to elicit, identify, confront, and resolve alternative conceptions (addressing prior knowledge). The third level includes intermediate skills and practices that are most closely tied to inquiry lessons, and loosely connected to applying in Bloom's Taxonomy. These skills and practices are promoted and developed as students identify scientific principles and/or relationships (cooperative work used to construct more detailed knowledge). The fourth level includes integrated skills and practices that are most closely tied to inquiry labs, and loosely connected to analyzing in Bloom's Taxonomy. These skills and practices are promoted and developed as students establish empirical laws based on measurement of variables (cooperative or collaborative work is used to construct more detailed knowledge). The fifth level includes culminating skills and practices that are most closely tied to real-world applications, the level of inquiry that has been added in the expansion, and loosely connected to evaluating in Bloom's Taxonomy. These skills and practices are promoted and developed as students solve problems related to authentic situations while working individually or in cooperative and collaborative groups using problem-based and project-based approaches. The sixth level includes advanced skills and practices that are most closely tied to hypothetical inquiry, and loosely connected to synthesizing in Bloom's Taxonomy. These skills and practices are promoted and developed as students generate explanations for observed phenomena (experience a more realistic form of science). The categorized skills and practices for each level can be found in Table 6.

Table 6

Wenning and Vierya's Intellectual Process Skills and Scientific Practices Framework (2015)

Practice/Skill Category	Level of Inquiry	Bloom's Taxonomy	Intellectual Process Skill/ Scientific Practice	Classification
Rudimentary	Discovery Learning	Remembering	Acquiring qualitative data	Scientific practice
			Classifying	Scientific reasoning
			Conceptualizing	Scientific reasoning
			Concluding	Scientific reasoning
			Contextualizing	Scientific reasoning
			Generalizing	Scientific reasoning
			Observing	Scientific practice
			Ordering	Scientific reasoning
			Problematizing	Scientific reasoning
Basic	Interactive demonstrations	Understanding	Estimating	Scientific reasoning
			Explaining	Scientific reasoning
			Formulating and revising scientific explanations using logic and evidence	Critical thinking
			Predicting	Scientific reasoning
			Recognizing and analyzing alternative explanations and models	Critical thinking
			Using conditional thinking	Scientific reasoning
Intermediate	Inquiry lessons	Applying	Applying information	Scientific reasoning
			Assisting with the design and execution of controlled scientific investigations	Scientific practice
			Collecting and recording quantitative data	Scientific practice
			Describing relationships	Scientific reasoning
			Making simple sense of quantitative data	Scientific reasoning
			Measuring	Scientific practice
			Using combinatorial thinking	Scientific reasoning
			Using correlational thinking	Scientific reasoning

Table Continues

Practice/Skill Category	Level of Inquiry	Bloom's Taxonomy	Intellectual Process Skill/ Scientific Practice	Classification
Integrated	Inquiry labs	Analyzing	Defining precisely a problem to be studied	Scientific reasoning
			Defining precisely the system to be studied	Scientific reasoning
			Designing and conducting controlled scientific investigations	Scientific reasoning
			Distinguishing independent and dependent variables	Scientific practice
			Interpreting quantifiable data to establish laws using logic	Scientific reasoning
			Using technology and math during investigations	Scientific practice
Culminating	Real-world applications	Evaluating	Collecting and evaluating data from various sources	Critical thinking
			Determining if an answer to a problem or question is reasonable including size and/or units	Scientific reasoning
			Making and defending evidence-based conclusions and judgments of arguments based on the logical interpretation of scientific evidence and other criteria	Critical thinking
			Solving complex real-world problems	Critical thinking
			Summarizing for the purpose of logically justifying a conclusion on the basis of empirical evidence	Scientific reasoning
			Using causal reasoning to distinguish co-incidence from cause and effect	Scientific reasoning
			Using causal reasoning to distinguish correlation from cause and effect	Scientific reasoning
			Using data and math in the solution of real-world problems	Scientific reasoning
			Using proportional reasoning to make predictions	Scientific reasoning

Table Continues

Practice/Skill Category	Level of Inquiry	Bloom's Taxonomy	Intellectual Process Skill/ Scientific Practice	Classification
Advanced	Hypothetical inquiry	Synthesizing	Analyzing and evaluating scientific arguments	Critical thinking
			Creating abstract hypothetical explanations	Critical thinking
			Creating a unique communication	Scientific practice
			Evaluating and revising hypotheses in light of new evidence	Critical thinking
			Generating and evaluating analogies	Scientific reasoning
			Generating predictions through the process of deduction	Scientific reasoning
			Thinking analogically	Scientific reasoning
			Thinking to assimilate concepts	Scientific reasoning
			Thinking deliberately	Scientific reasoning
			Using probabilistic thinking	Critical thinking

Because the framework is an exhaustive collection of intellectual process skills and scientific practices, not all can be categorized as scientific reasoning. Some skills are classified as critical thinking which is the process of evaluating statements, opinions, and hypotheses by collecting, analyzing, and evaluating data, issues, and arguments from different sources and perspectives (Herr, 2008). Others are classified as nothing more than scientific practice, which more closely resembles the actions of a scientist. These can be found among the scientific inquiry skills mentioned previously with the *ScInqLiT* or within the eight practices that the *Next Generation Science Standards* (NGSS) *Framework* (National Research Council, 2013) identifies as essential. These practices can be found in Table 7.

Table 7

Scientific Practices of the Next Generation Science Standards Framework (2013)

- Asking questions and defining problems - Developing and using models - Planning and carrying out investigations - Using mathematics and computational thinking - Constructing explanations and designing solutions - Engaging in argument from evidence - Obtaining, evaluating, and communicating information

After removing critical thinking skills and scientific practices, the remaining are scientific reasoning skills. What separates these skills from the rest is that they either involve inductive reasoning or deductive reasoning. Inductive reasoning is the process of making generalizations from specific information. Deductive reasoning is the process of drawing specific conclusions from general principles or premises (Herr, 2008). This may seem to be too simplified considering Han and Lawson also name hypothetical-deductive, causal, proportional, probability, and correlational as reasoning types. However, inductive or deductive appears to be found within each of these types, and essentially covers all the bases of reasoning in the sciences. For example, correlational reasoning can be considered inductive because a generalization of a correlation is drawn from specific concurrent events. If a skill displays either process, then it can be classified as scientific reasoning. In doing so, the quantity of skills and practices in Wenning and Vierya's framework can be reduced from a total of forty-eight to thirty-one. To make it clearer as to why a skill is classified as scientific reasoning, Wenning and Vierya's (2015) operational definition for each skill can now be displayed in a framework that is solely scientific reasoning focused. These defined skills along with examples are found in Table

8. Table 8 also shows how much more comprehensive the framework is by noting which skills Lawson and iSTAR address.

Table 8

Framework of Scientific Reasoning Skills Operationally Defined by Wenning and Vierya (2015)

Category	Scientific Reasoning Skill	Definition	Example	Lawson	iSTAR
Rudimentary	Classifying	Categorizing phenomena on the basis of commonalities, dissimilar attributes, or other criteria.	Grouping objects based on observable traits, such as asking students to classify different types of lenses or mirrors based upon their shape.		
	Conceptualizing	Generalizing critical observations of specific instances of a phenomenon to create an abstraction.	Dropping balls of different mass from different heights into one another's hands, students come to understand the concept of kinetic energy.		
	Concluding	Processing data using scientific reasoning to establish if-then statements or similar relationships based on commonalities.	As one trait of an example increases, so does another, such as the formulation of the statement, "If the two surfaces in contact with one another become more slippery, then there will be less friction between the two surfaces."		
	Contextualizing	After being introduced to a topic, students are asked to brainstorm particular instances of the phenomenon.	When being introduced to electricity, students are asked to provide a number of examples where they encounter this phenomenon in their daily lives.		

Table Continues

Category	Scientific Reasoning Skill	Definition	Example	Lawson iSTAR
Rudimentary	Generalizing	Making general or broad statements by inferring from specific cases. Using critical observations of specific instances of a phenomenon to generate a qualitative principle that describes a relationship among variables.	Recognizing that all objects moving away from a motion detector create a position-time graph with a positive slope.	
	Ordering	Arranging sets of objects in sequence using a common characteristic.	Arranging objects on the basis of some progressively changing observable trait, such as ranking the mass or volume of objects.	
	Problematizing	Having reviewed the physical examples of the topic being introduced, the students identify a number of problems in need of solution.	With the concept of momentum, what happens when a car and a truck of different masses and speeds collide head on?	
Basic	Estimating	Determining roughly through calculation or other reasoning processes the approximate value of a quantity or extent of a phenomenon under consideration.	How thick is a sheet of paper or what is the mass of the moon in kilograms or how many times does a person's heart beat during an average human lifetime?	
	Explaining	Simple hypothesizing, translating, interpreting, or otherwise making clear by providing additional details, information, or ideas.	Following the making of a prediction, students explain their reasoning, as in "A red dot viewed through a blue filter will appear black, because all colors except blue get filtered out (absorbed), and don't make it through the filter. As a result, the light from a red dot won't make it through to the viewer's eye."	

Table Continues

Category	Scientific Reasoning Skill	Definition	Example	Lawson iSTAR
Basic	Predicting	Foretelling what will happen or will be the consequence an event under a given set of circumstances or conditions using the process of extrapolation.	Given a sequence of events set into motion, students will state a probable outcome assuming some form of causality such as stating "Changing the mass of a pendulum bob will not have any effect on the period of the swing so long as the length of the pendulum remains unchanged in doing so."	x
	Using conditional thinking	Drawing conclusions from if-then statements.	"If I drop an object, it will increase in both kinetic energy and momentum" and "If the size of a sample of given material is larger, it is heavier. Sample A is larger than sample B. Therefore sample A is the heavier than sample B."	x
Intermediate	Applying information	Solving problems in new situations by applying previously acquired knowledge, facts, techniques and rules in a different way.	Information from prior experiences with a phenomenon is used to develop an experiment. Given students' understandings about how to measure the final velocity of a ball rolling down a tube with a known length and a stopwatch, students can predict where a marble projectile might land on the ground as it slides horizontally off of the table edge.	
	Describing relationships	Identifying and summarizing if-then relationships in quantifiable physical form including relevant characteristics or qualities.	If the average speed of an object is increased over a given interval, the time required for that object to travel the interval will decrease. Similarly, if the voltage applied to a given electrical circuit increases, the current passing through that circuit will likewise increase.	

Table Continues

Category	Scientific Reasoning Skill	Definition	Example	Lawson	iSTAR
Intermediate	Making simple sense of quantitative data	Examining data to look for and identify trends and possible physical or mathematical relationships using approaches such as graphing or correlation.	Students might count the number of repeating images in a kaleidoscope while varying the angle between the two mirrors. Students can easily find that the number of images is equivalent to 360 degrees divided by the angle between the two mirrors. Likewise, students should be able to identify outlying data points that might not fall within a sensible relationship for the entire group of data.		
	Using combinatorial thinking	Reasoning about all possible combinations, identifying all possible ways in which a number of variables in a given system can interact.	Students explain that cause-and-effect relationships involving more than two variables (e.g., $\Sigma F = ma$ or $\Delta V = IR$) exist and note their interconnection.		x
	Using correlational thinking	Recognizing or rejecting the presence of cause-and-effect relationships despite the presence of concurrent (literally co-incident) events.	Students can explain that while correlation does not imply causation, the lack of correlation does imply the lack of causation. Although frequency and wavelength share a proportional relationship with wave speed, wave speed is dependent only upon the medium type, and not either wavelength or frequency.	x	x
Integrated	Defining precisely a problem to be studied	Clearly stating, following a review of empirical evidence, a problem in need of a solution.	A dynamics cart rolls down an incline plane and its distance is observed to increase disproportionally as a function of time. The student states, "What is the relationship between distance and time for a cart whose acceleration is constant?"		

Table Continues

Category Scientific Reasoning Skill	Definition	Example	Lawson	iSTAR
Defining precisely the system to be studied	Analyzing and identifying all interacting parts of a physical phenomenon including those parts of the natural environment that relate to the question to be answered by an experiment.	Realizing that the amplitude of pendulum will decrease with time as a result of wind resistance.		
Integrated Designing and conducting controlled scientific investigations	Allowing for only one independent variable and one dependent variable at a time, holding all other pertinent variables constant during the experiment.	Holding mass of a ball constant but varying the height of release, determine the relative amount of kinetic energy upon impact by measuring the volume of a depression it makes in clay. Similarly, holding the height of balls constant but varying the mass, determine the relative amount of kinetic energy by measuring the volume of the depressions they make in clay upon impact. Combining the results leads to the final relationship between all three variables.	x	x
Interpreting quantifiable data to establish laws using logic	Using graphs or other representations or depictions to analyze the consequences of the change of independent variables on the dependent variable thereby identifying organizational principles.	In a determination of Hooke's Law for springs, students might realize that each spring has its own unique constant (ratio of $F/\Delta x$), but that every spring's applied force can be represented by the same general equation ($F = k\Delta x$).		

Table Continues

Category	Scientific Reasoning Skill	Definition	Example	Lawson	iSTAR
Culminating	Determining if an answer to a problem or question is reasonable including size and/or units	Calculated answers in science typically are derived from measured values that include magnitude and unit of measurement. It is important to be able to determine if the magnitude and units are reasonable so that answers can be self-checked.	The mass of the Earth is calculated to be $3.87x10^9$ kilograms or the number of kilometers in a light year is $3x10^7$ seconds or the momentum of a 1-kg dynamics cart moving at 3 m/s is 3 Newtons. Are these correct?		
	Summarizing for the purpose of logically justifying a conclusion on the basis of empirical evidence	Explaining in a comprehensible form decisions developed through the analysis of specific instances of a phenomenon.	Using symbols or words to provide in written and/or oral form the meaning of a set of observations, such as the verbal formulation of relationships ("If…then…" or "As _____ increases, then _____ decreases."), or the simple representations of equations, ratios, graphs, charts, images, or drawings.	x	x
	Using causal reasoning to distinguish co-incidence from cause and effect	Just because two things are temporally related, it does not mean that there is a causal mechanism.	Bears hibernate in the autumn but that doesn't bring about winter; birds migrate north in the spring but that doesn't bring about summer.		
	Using causal reasoning to distinguish correlation from cause and effect	Just as one thing increases as another increases or decreases and vice versa there is not necessarily a cause-and-effect relationship at work here. Only when a controlled experiment is conducted might one say that such a relationship is supported by evidence.	Ice cream sales and shark attacks on swimmers both increase during the summer, but that doesn't mean that the increase in ice cream sales is the cause for the increase in shark attacks even though there is a correlation between the two.		x

Table Continues

Category	Scientific Reasoning Skill	Definition	Example	Lawson	iSTAR
Culminating	Using data and math in the solution of real-world problems	It is important not only to know math, but know how and when to apply it to real-world problems. This can range from correctly interpreting graphs to making simple calculations to draw independent conclusions from data.	Drawing a conclusion from a data table that shows how many days a patient is cured after taking a certain dosage of new medication.		
	Using proportional reasoning to make decisions	Given a mathematical law and a change of variables, correctly forecast the consequences from those changes.	Given the relationship $F_1 = kQq/r^2$ and the facts that Q is doubled, q is halved, and r is doubled, indicate that the new force in comparison with the initial force ($F_2/F_1 = ¼$).	x	x
Advanced	Generating predictions through the process of deduction	Using the supposed correctness of a law, principle, or hypothetical explanation to forecast the outcome of a specific situation.	Given the thin lens formula, predict the object distance given the image distance for a lens with a known focal length.		
	Generating and evaluating analogies	Defining an analog to some system and then determining the appropriateness of various comparative features in supposedly analogous systems.	How are electrical force, field strength, and potential analogous to gravitational force, field strength, and potential? Is pressure in a water paper system analogous to voltage in an electrical circuit?		

Table Continues

Category	Scientific Reasoning Skill	Definition	Example	Lawson	iSTAR
Advanced	Thinking analogically	Using reasoning based on the idea that two things are similar in many if not all ways allowing inferences generated in one domain to be applied to another domain.	Explaining how energy is transported in an electrical circuit using hot water flowing in pipes with insulation and radiator fins as an analogy. Additionally, determining the aptness of such an analogy by comparing corresponding parts between two models.		
	Thinking deliberatively				
	Thinking to assimilate concepts	This is the type of thinking that occupies students when they seek to understand observations or ideas, and relate them to or reconcile them with knowledge they already possess.			

Note. Skills found on either the Lawson Test or iSTAR Assessment are marked with "x".

With these 31 defined scientific reasoning skills, a test can be created to address another dimension of science literacy in the same manner as the *NOSLiT* and *ScInqLiT* addressed the nature of science and scientific inquiry dimensions. This test should be a valid and reliable instrument tailored for high school science students, and go above and beyond the scientific reasoning dimensions addressed by Lawson and iSTAR. Furthermore, this test should be aligned with a defined framework such as Wenning and Vierya's. This framework contains skills that are mainly science related, and as displayed in Table 8, is a vast expansion of the dimensions included in Lawson's test and iSTAR.

Additionally, the skills in this framework are defined as more basic or advanced. As a result, students will be assessed comprehensively and in a systematic fashion.

CHAPTER III

RESEARCH DESIGN

Overview of the Assessment Instrument

The test created for and used in this study is designed to answer the following questions:

- Can a valid scientific reasoning test for high school science students be created from the defined scientific reasoning skills in Wenning and Vierya's intellectual process skills and scientific practices framework?
- Can a reliable scientific reasoning test for high school science students be created from the defined scientific reasoning skills in Wenning and Vierya's intellectual process skills and scientific practices framework?
- Can a scientific reasoning test for high school science students address skills that go above and beyond the dimensions addressed by the *Lawson Classroom Test of Formal Reasoning* and *Inventory for Scientific Thinking and Reasoning Assessment*?

This test is intended to assess scientific reasoning skills of high school students in a comprehensive and systematic manner, and is based on a modified version of Wenning and Vierya's (2015) theoretical framework of intellectual process skills and scientific practices. The framework contains 31 scientific reasoning skills defined by six different levels of increasing intellectual sophistication. However, the test only addresses the 26

skills found in the lower five levels. The advanced level skills are not included as part of this test for two main reasons. First, these skills are considered too challenging to assess with multiple-choice questions, and would most likely need some sort of questions requiring constructed responses that take much more time for a teacher to evaluate. Second, these skills are associated with activities rarely taught in high school. This test (Appendix A) contains 26 multiple-choice questions, one question for each skill. As a result, the test is not so lengthy that teachers will not be hindered from administering it. On the surface, it would appear that dedicating only one question to each of the 26 different skills works against content validity and possibly creates an internal consistency problem. However, with each skill connected to the defined categories, four to seven questions are utilized to assess the grouping of comparable skills found in each level of intellectual sophistication. All questions are multiple-choice with five possible answers.

Expert Review

To begin the process of developing a valid and reliable test, a team of reviewers consisting of three high school science teachers and two university physics professors reviewed a pool of 38 questions for clarity, accuracy, reading difficulty, and redundancy. They were instructed to first determine if the test had construct validity, which "refers to the extent to which a test reflects constructs presumed to underlie the test performance and also the extent to which it is based on theories regarding those constructs" (Ary, Jacobs, & Razavieh, 1972, p. 197). This entailed reviewing the framework defining scientific reasoning skills to make sure that it is comprehensive and properly defined. At the time, the framework included 31 skills that covered all six levels of intellectual

sophistication. They were instructed to determine if the test had content validity, which "refers to the degree to which a test samples the content area to be measured" (Ary, Jacobs, & Razavieh, 1972, p. 191). This entailed checking to see if the questions of the test were aligned to the skills found in the framework. Beyond validity, they were asked for any feedback that they would share concerning questions that were inaccurate, incomplete, confusing, poorly worded, illogical, and/or had multiple or no correct answers. The review occurred in June of 2015.

The review process brought forth substantial changes to the pool of questions. The first change was that the number of questions was decreased from 36 to 33. This involved eliminating three questions originally aligned to the scientific reasoning skills found in the advanced level. One such question associated with thinking analogically asked in what way a battery in an electric circuit follows the analogy that electric charges flow through a circuit like water flows through a piping system. A comment regarding this question was that it was too content based and more about knowing instead of reasoning. This comment as well as the commentary of the other two questions coupled with the extreme difficulty of developing questions for these skills (only two of the five advanced skills had questions) made it necessary to drop this level from the test, and merely focus only on the other five levels. Wenning's stance that high school students rarely encounter activities involving these advanced skills strengthened the choice of eliminating them.

The second change was that three questions were completely replaced by new ones. These questions addressed contextualizing, defining precisely a system to be studied, and summarizing for the purpose of logically justifying a conclusion on the basis

of empirical evidence. The contextualizing question, which proved to be one of the most challenging questions outside of the advanced level to create, involved students engaged in discovering properties of bar magnets, such as how they are oriented in order to repel and attract each other. The best response to the teacher's question of where they might have seen this effect in their daily lives needed to be chosen. There were several comments regarding issues with this question. One comment stated that the answer, "I once was in a junkyard, and a big crane picked up metal like this" could be correct because magnets pick up cars and metals in junkyards. Although this is true, the big crane referenced in the answer was one that uses a scooping claw-like mechanism to pick up metal. Another comment inquired about kids without cell phones answering the question. This was problematic considering that the answer, "I have cell phone and tablet covers that have latches that work like this" was supposed to be the correct choice. One final comment stated that the question tests familiarity with forces and objects in the answers. This comment actually sums up the previous two comments, and truly highlights the main issue with creating a contextualizing question. This type of question requires that all students make the necessary contextual connections within the topic and answer choices. When students lack prior experiences related to the question, it becomes difficult if not impossible to answer correctly, and consequently is biased toward those students who have had the experiences. The replacement question was changed to the contextual topic of static electricity in hopes that it was common enough to remove issues related to prior experience deficits.

The other two questions had their own set of issues although not as challenging to rectify. The defining precisely a system to be studied question was considered far too

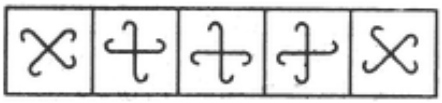
easy and therefore it was difficult to differentiate from the skill problematizing. The question involved different sized materials being packed into a container and what problem needed to be solved. The idea that the question was problematizing had more to do with how the skill it was addressing was defined at the time of the review. It was actually targeting the skill defining the problem and system to be studied, so ultimately it was a question related defining the problem. Once it was determined that the explanation of the question's situation made the answer choice fairly obvious, the question was scrapped and the defining the problem and system skill was split. Because there was another defining the problem question that was accepted by the reviewers, a defining the system question needed to be created. The summarizing for the purpose of logically justifying a conclusion on the basis of empirical evidence question seemed to be too much of an interpreting quantifiable data to establish laws using logic question. The question included a data table of flux and distance values; and the general equation relating the two variables needed to be chosen. The issue here is that making the choice of a general equation better reflects choosing a law than choosing a summarization. Consequently, a question was created that required a summary statement about a graph be chosen.

The third change was that two questions addressing skills outside of the advanced level were eliminated. One question associated with estimating asked for the best estimate for the time it would take to continuously count to a million by 10's. There was issue with the cadence of counting smaller numbers (i.e. 10, 20, 30…) being different than the cadence of larger numbers (i.e. 110250, 110260, 110270…). As a result, it was determined that students would have difficulty selecting a set of counts that best

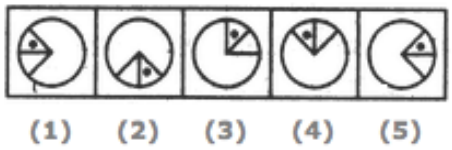
represented the average cadence of all of the numbers. In the end, eliminating the question due to this issue was easily justified, because there was another similar estimating question.

The other question was associated with the skill using probabilistic thinking, which is no longer included in the scientific reasoning skills framework. Wenning and Vierya (2014) define using probabilistic thinking as "recognizing the fact that observations are probabilistic in nature (e.g., all observations are subject to random errors) and require that conclusions must include considerations for such probabilities." The issue with this question was two-fold. First, using probabilistic thinking was determined by Wenning to be more of a critical thinking skill as opposed to a scientific reasoning skill. Second, the question originally created for this skill was very similar to the probability reasoning questions of Lawson and iSTAR. However, this type of question does not properly follow the definition of using probabilistic thinking. Furthermore, it was determined that probability reasoning as defined by Lawson and iSTAR is for the most part a mathematical reasoning skill that is not science specific enough for a scientific reasoning skills framework.

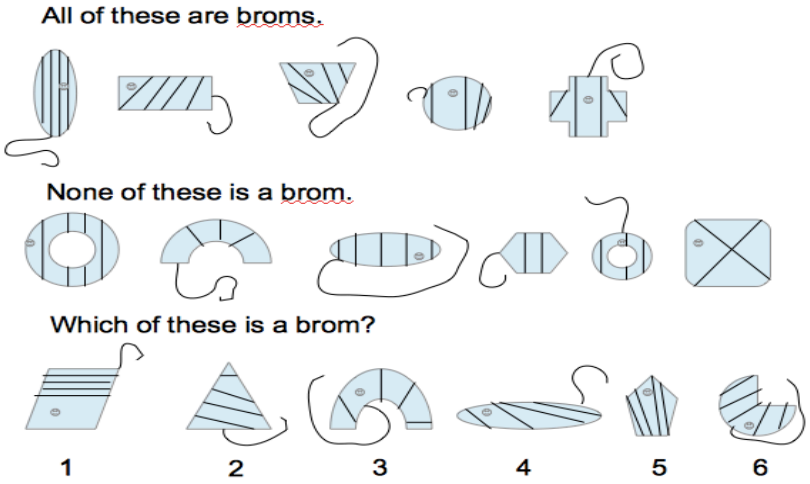
Beyond the substantial changes to the pool of questions, six questions required revisions that for the most part kept the majority of the question in tact. These questions were aligned with the skills classifying, conceptualizing, concluding, problematizing, and correlational thinking. The classifying question initially contained the following image:



One of the figures in the image that was different from the rest had to be chosen. The trait that made the figure different was considered too challenging to determine. This was changed to the following image that was seemingly easier because the figure that was different had a trait that was not so difficult to determine:



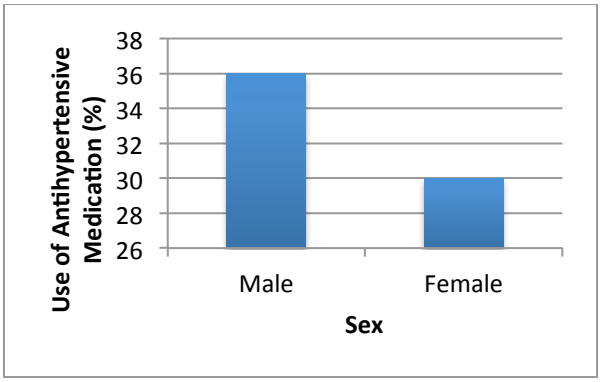
The conceptualizing question also needed images changed. The original images revolving around the fictitious concept of a "brom" were as follows:



These images showed that broms have three shared characteristics: a small circle, a tail, and four lines, and non-broms missing at least one of these characteristics. One reviewer stated that there were too many variables, so the images were modified to show broms having two shared characteristics of a tail and four lines, and non-brom missing at least one of these characteristics.

There were two concluding questions that required revisions. The first question

involving batteries and light bulbs wired in various circuits had answers that required students to make conclusions about electricity flow based on observations comparing the brightness of the light bulbs in each circuit. A reviewer commented that students would need the background knowledge that brighter light bulbs equal more electricity flow. All language concerning electricity flow in the answers was then changed to bulb brightness. The second question showed the following graph, which was the source of two issues:



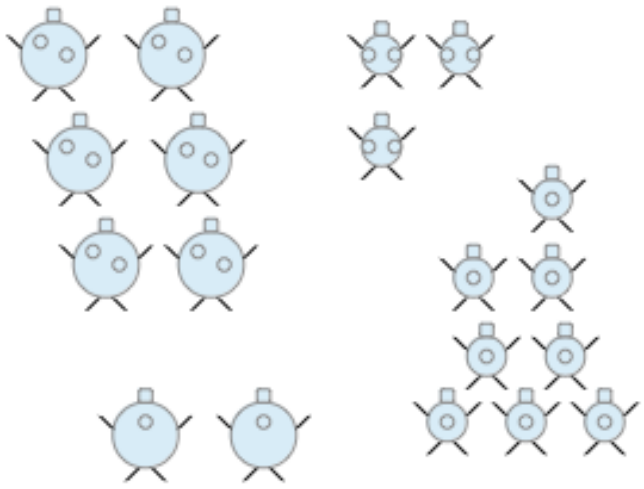
One issue was the use of the term antihypertensive. This term was changed to blood pressure, which was used in the explanation associated with the graph. The other issue was the correct answer stating that males take blood pressure medicine at a rate roughly 20% higher than females. One reviewer commented that this tested understanding of percent and rate. As a result, the answer was changed to males use blood pressure medicine 1.2 times more than females.

The problematizing question had issues within the following table that displayed the patterns of men's behavior whom had recovered from baldness within the last few months:

Percent of men with the following traits:	
15%	Have lost weight during the past year.
23%	Have gained weight over the past year.
83%	Take aspirin daily to prevent heart attack.
26%	Use a particular type of hair shampoo.
98%	State that they enjoy watching television.
12%	Have reduced their exposure to sunlight.

The language associated with taking aspirin and watching television appeared to be too vague to properly choose the correct answer of the best research question based on the data. The questions, "does aspirin cure baldness?" and "does watching television affect baldness?" both could have been justifiably correct. Both traits had a much higher percentage of men whom recovered from baldness. However, watching television was incorrect, because it is assumed to be more of a lifelong habit that should have no recent effect on baldness, an assumption that may be unreasonable without explicitly stating it. Consequently, the language of watching television in the table was changed to state this assumption. Also, the language of taking aspirin daily in the table was changed to highlight that it was a recent habit.

The correlational thinking question had answers that were not definitive enough about the link between the size of the sea turtles and number of markings shown in the following picture that depicted a collection of sea turtles:



Three of the answer choices stated that there appears to be either: a. a strong link, b. a weak link, or c. no link. This led one reviewer to ask how one would define strong vs. weak. Consequently, the answer choices were changed to state that there appears to be a relationship between size and number of markings for either: a. most turtles, b. some turtles, or c. no turtles.

Once all of the changes were made as a result of the expert review, the test consisted of 33 questions aligned to 26 scientific reasoning skills. All skills had a minimum of one question with four of the skills having two questions: concluding, using correlational thinking, using causal reasoning to distinguish co-incidence from cause and effect, and using proportional reasoning to make predictions; and one skill having four questions: interpreting quantifiable data to establish laws using logic. The reason there was a relatively higher number of interpreting data to establish laws using logic questions was because at the time of the review and first round of testing these four questions were addressing two very similar scientific reasoning skills that were merged into one skill. The merge of these skills took place following the pilot test data analysis.

Data Collection

Before any data were collected, the Research and Ethics Committee (REC) at Illinois State University deemed that there was no need to obtain an approved research protocol from the Institutional Review Board (IRB), because the research was based on the statistical characteristics and clarity of the assessment questions, and not evaluating human subjects.

Following the expert review, high school physics teachers in Illinois found in the Illinois Section of American Association of Physics Teachers (ISAAPT), Illinois Science Teachers Association (ISTA), and Illinois State University Physics Education email databases were contacted by email to solicit participation. Teachers who could administer 100-125 tests each were chosen so that only five to six teachers were needed. Thirty copies of the assessment and 100-125 Opscan forms with a cover letter were delivered to these teachers. The cover letter indicated that the data collected would determine how well the questions of the test perform and would not be based on student performance, so it was imperative that all results remain anonymous. Consequently, the cover letter instructed the teachers on how students should fill out their Opscan forms. No student should write his or her name anywhere on the form. Each student should write and bubble in the provided teacher code in the MISC. section, and the class period in the DEPT. section. The cover letter also informed the teachers that no calculators were permitted, and the test should take no more than one class period to complete. To help motivate the teachers, they were informed that feedback concerning the results would be provided. Once the teachers were chosen to participate, the pilot test was administered to students during September of 2015. A second test based on modifications as a result of the pilot

test data analysis was administered to a new set of students during October of 2015.

Six science teachers from Lanphier High School in Springfield, Illinois administered the pilot test to a total 540 students in 9th, 10th, 11th, and 12th grades. Two science teachers from Lanphier High School, one science teacher Springfield High School in Springfield, Illinois, one science teacher from Southeast High School in Springfield, Illinois, one science teacher from University High School in Normal, Illinois, and one science teacher from Clinton High School in Clinton, Illinois administered the second test to a total of 379 students in 9th, 10th, 11th, and 12th grades.

Statistical Measures

The pilot test and its multiple-choice questions were analyzed using various statistical methods performed by Measurement and Evaluation Services at Illinois State University. The analysis of the test overall included range of scores, mean score, standard deviation, variance, and Kuder-Richardson Formula 20 (KR-20). The mean score, standard deviation, and variance were utilized for norming purposes. A mean score with a value that reflects a percent score of approximately 50% was ideal because the assessment is designed to produce the maximum possible spread among scores. This in turn would produce higher standard deviation and variance. The KR-20 value was the main indicator of an acceptable assessment instrument. The KR-20 is a measure of the extent to which the items on a test provide consistent information about students' level of knowledge of the content of the test. KR-20 values for professionally developed and widely administered tests such as SAT and GRE are expected to be greater than or equal to .80, which was the benchmark for this test (Office of Measurement and Evaluation of

Teaching, 2015).

An analysis was also conducted on each question looking at the item difficulty index and point-biserial discrimination index. The item difficulty index is a measure of the proportion of students who answered the item correctly, and typically has a value between .40 and .60 for norm referenced tests (Professional Testing, 2015). The point-biserial discrimination index indicates how well an item serves to discriminate between students with higher and lower levels of knowledge, and as a general rule is considered desirable with values of .20 and above (Office of Measurement and Evaluation of Teaching, 2015). These indices helped determine whether any question needed to be revised or rejected.

CHAPTER IV

ANALYSIS OF THE DATA

Statistical Analyses

The pilot test with 33 questions had a range of scores of 28, mean score of 9.90 (30%), standard deviation of 4.34, variance of 18.80, and KR-20 of .68. The item difficulty index and point-biserial discrimination index for each question are found in Table 9. Table 9 also includes the scientific reasoning skills operationally defined by Wenning and Vierya (2015) aligned to each question and question numbers on the test. Because the pilot test had a mean score value that was well below 50% and KR-20 value less than .80, changes needed to be made in order to increase the performance of the test. The questions that were involved in these changes are noted in Table 9 as well.

The second test developed as a result of these changes contained 26 questions that had a range of scores of 26, mean score of 11.18 (43%), standard deviation of 5.61, variance of 31.45, and KR-20 of .85. The item difficulty index and point-biserial discrimination index for each question are found in Table 10. Table 10 also includes the scientific reasoning skills operationally defined by Wenning and Vierya (2015) aligned to each question and question numbers on the test. Lower performing questions that could possibly revised are noted in Table 10 as well.

Table 9

Thirty-Three Pilot Test Questions Aligned to the Scientific Reasoning Skills Operationally Defined by Wenning and Vierya (2015)

Category	Scientific Reasoning Skill	Test Question	Question Number	Item Difficulty	Point-Biserial
Rudimentary	Classifying	Five figures are shown with one different from the rest. The different figure is chosen.	2	.65	.24
	Conceptualizing	One set of figures show the characteristics of "broms" and another set shows similar characteristics but are not "broms." Broms are chosen from a third set.	3^	.34	.17
	Concluding	Observations are made about four different electrical circuits. A conclusion is chosen based on the observations.	5	.53	.34
		A graph displays the percentage of males and females taking blood pressure medication. The conclusion that can be drawn from the data shown in the graph is chosen.	7*	.11	.26
	Contextualizing	Students observe a variety of demonstrations dealing with static electricity. Where else we see the effect of static electricity is chosen.	1	.58	.30
	Generalizing	A group of gray and black objects is shown. The correct general statement about the group is chosen.	4	.45	.45
	Ordering	Various quantities of planets are shown in a data table. The order of the planets' distances from the sun is chosen.	6	.34	.33
	Problematizing	Data concerning the traits of men with baldness is shown. The best research question based on the data is chosen.	9^	.20	.25

Table Continues

Category	Scientific Reasoning Skill	Test Question	Question Number	Item Difficulty	Point-Biserial
Basic	Estimating	The best estimate for how many heartbeats are made in a span of 25 years is chosen.	8^	.21	.14
	Explaining	Explanations of what occurs during the burning of steel wool are compared to a figure of steel wool on a balance before and after it is burned. The best explanation that supports the figure is chosen.	12^	.20	.23
	Predicting	A graph represents the relationship between the weight and age. The weight at a certain age beyond the data in the graph is predicted.	10	.33	.27
	Using conditional thinking	The line of reasoning that all apples are either red or green, and all green apples are hard is given. The correct conclusion about all hard apples is chosen.	11	.30	.23
Intermediate	Applying information	A capped bottle filled with water that contains an eyedropper that sinks and floats as the bottle is squeezed and released. Knowing how density is related to floating and sinking, the reason why the eyedropper floats and sinks is chosen.	13	.31	.31
	Describing relationships	A graph shows the relationship between mass and volume of two substances. The statement that best describes the relationship is chosen.	18^	.29	.27
	Making simple sense of quantitative data	A data set of distance and speed is shown. The graph that best represents the relationship between distance and speed is chosen.	17^	.19	.13
	Using combinatorial thinking	A data set of mass, density, volume submerged, and buoyant force is shown. The related variables are chosen.	16^	.23	.29
	Using correlational thinking	A picture depicts sea turtles that are either small or big and have either one or two markings on their backs. The relationship between the size of the sea turtle and number of markings is chosen.	14	.29	.29
		Graphs show the relationship of height vs. weight of three different groups of children. The group that displays the strongest relationship is chosen.	15*	.55	.36

Table Continues

Category	Scientific Reasoning Skill	Test Question	Question Number	Item Difficulty	Point-Biserial
Integrated	Defining precisely a problem to be studied	A pair of identical springs used to pull a cart up a hill is connected first side by side then one after another. The springs stretch a given distance in the first arrangement and twice as much in the second arrangement. The problem that might be studied based on these arrangements is chosen.	19	.34	.39
	Defining precisely the system to be studied	A piece of paper and a round stone are released simultaneously from rest at the same height above a floor to test the claim that heavier objects fall faster than lighter objects. The correct student statement about the observation is chosen.	20^	.18	.18
	Designing and conducting controlled scientific investigations	A student designs an experiment to determine if weight, shape, and color effect how quickly objects sink to the bottom of a container filled with water. The group of objects that determine if shape has an effect on the sinking rate is chosen.	21	.31	.48
	Interpreting quantifiable data to establish laws using logic	Graphs show how weight, age, and environment temperature are related to food eaten by a newly discovered species. The correct single combined relationship for these variables is chosen.	22^	.12	.27
		A data table shows data collected by a scientist trying to find the relationship between food eaten by a newly discovered species and the size of the creature within the species and temperature of the environment. The correct relationship that the scientist found is chosen.	23*	.21	.24
		A data set and graph of position and time of a motorized car is shown. The velocity of the motorized car is chosen.	24*	.29	.37
		A data set and graph of position and time of a motorized car is shown. The correct mathematical model of the data and graph is chosen.	25*	.18	.34

Table Continues

Category	Scientific Reasoning Skill	Test Question	Question Number	Item Difficulty	Point-Biserial
Culminating	Determining if an answer to a problem or question is reasonable including size and/or units	A scientist calculates the number of kilometers in a light year and arrives at 3 x 10^7 seconds. The statement of whether or not the answer is reasonable and why or why not is chosen.	33^	.11	.34
	Summarizing for the purpose of logically justifying a conclusion on the basis of empirical evidence	A graph shows a scientist's count of the number of electrons emitted from a radioactive sample as a function of time. The conclusion that the scientist can properly draw from the data is chosen.	30^	.15	.07
	Using causal reasoning to distinguish co-incidence from cause and effect	A person crosses paths with a black cat, and later is involved in an accident. The reason why this occurred is chosen.	27	.44	.45
		The patterns cold weather of winter follows bears hibernating in the autumn, and hot weather of summer follows birds migrating north in the spring are given. The correct statement based on these patterns is chosen.	28*	.21	.02
	Using causal reasoning to distinguish correlation from cause and effect	A swimmer at a beach notes that ice cream sales affect the number of shark attacks on swimmers, because the higher the ice cream sales, the greater number of shark attacks on swimmers. The problem with this statement is chosen.	29^	.21	.39
	Using data and math in the solution of real-world problems	A data table shows how many days a patient is cured after taking a certain dosage of new medication. The conclusion that can be drawn from this data is chosen.	26	.49	.42
	Using proportional reasoning to make decisions	Knowing the ratio between cups of flour and loaves of bread, the amount of flour needed to make three loaves of bread is chosen.	31*	.49	.46
		The relationship U = kQq/r is considered where k and Q are constants. The statement of U increasing or decreasing due to q and r being changed by various factors is chosen.	32^	.29	.30

Note. Questions that were eliminated following the statistical analysis are marked with "*". Questions that were revised or replaced following the statistical analysis are marked with "^".

Table 10

Twenty-Six Second Test Questions Aligned to the Scientific Reasoning Skills Operationally Defined by Wenning and Vierya (2015)

Category	Scientific Reasoning Skill	Test Question	Question Number	Item Difficulty	Point-Biserial
Rudimentary	Classifying	Five figures are shown with one different from the rest. The different figure is chosen.	2	.64	.38
	Conceptualizing	One set of figures show the characteristics of "broms" and another set shows similar characteristics but are not "broms." Broms are chosen from a third set.	3+	.16	.18
	Concluding	Observations are made about four different electrical circuits. A conclusion is chosen based on the observations.	5	.53	.56
	Contextualizing	Students observe a variety of demonstrations dealing with static electricity. Where else we see the effect of static electricity is chosen.	1+	.69	.12
	Generalizing	A group of gray and black objects is shown. The correct general statement about the group is chosen.	4	.65	.58
	Ordering	Various quantities of planets are shown in a data table. The order of the planets' distances from the sun is chosen.	6	.43	.48
	Problematizing	Data concerning the traits of men with baldness is shown. The best research question based on the data is chosen.	8+	.23	.40
Basic	Estimating	The number of chirps a cricket will make over the course of 24 hours is estimated.	7	.43	.54
	Explaining	Explanations of what occurs during the burning of steel wool are compared to a figure of steel wool on a balance before and after it is burned. The best explanation that supports the figure is chosen.	12+	.25	.49
	Predicting	A graph represents the relationship between the weight and age. The weight at a certain age beyond the data in the graph is predicted.	9	.39	.41
	Using conditional thinking	The line of reasoning that all apples are either red or green, and all green apples are hard is given. The correct conclusion about all hard apples is chosen.	10	.37	.35

Table Continues

Category	Scientific Reasoning Skill	Test Question	Question Number	Item Difficulty	Point-Biserial
Intermediate	Applying information	A capped bottle filled with water that contains an eyedropper that sinks and floats as the bottle is squeezed and released. Knowing how density is related to floating and sinking, the reason why the eyedropper floats and sinks is chosen.	11	.39	.49
	Describing relationships	A graph shows the relationship between mass and volume of two substances. The statement that best describes the relationship is chosen.	16	.49	.46
	Making simple sense of quantitative data	A data set of distance and speed is shown. The graph that best represents the relationship between distance and speed is chosen.	15+	.19	.23
	Using combinatorial thinking	A data set of mass, density, volume submerged, and buoyant force is shown. The directly proportional variables are chosen.	14	.36	.44
	Using correlational thinking	A picture depicts sea turtles that are either small or big and have either one or two markings on their backs. The relationship between the size of the sea turtle and number of markings is chosen.	13	.39	.41
Integrated	Defining precisely a problem to be studied	A pair of identical springs used to pull a cart up a hill is connected first side by side then one after another. The springs stretch a given distance in the first arrangement and twice as much in the second arrangement. The problem that might be studied based on these arrangements is chosen.	17	.50	.54
	Defining precisely the system to be studied	A piece of paper and a round stone are released simultaneously from rest at the same height above a floor to test the claim that heavier objects fall faster than lighter objects. The correct student statement about the observation is chosen.	18	.36	.45
	Designing and conducting controlled scientific investigations	A student designs an experiment to determine if weight, shape, and color effect how quickly objects sink to the bottom of a container filled with water. The group of objects that determine if shape has an effect on the sinking rate is chosen.	20	.48	.55

Table Continues

Integrated	Interpreting quantifiable data to establish laws using logic	Graphs show how weight, age, and environment temperature are related to food eaten by a newly discovered species. The correct single combined relationship for these variables is chosen.	21	.33	.51
Culminating	Determining if an answer to a problem or question is reasonable including size and/or units	A shooter concludes that a bullet traveling 300 feet per second takes a time of $3^1/_3$ feet to reach a target 1000 feet away. The statement of whether or not the conclusion is correct and why or why not is chosen.	25	.40	.55
	Summarizing for the purpose of logically justifying a conclusion on the basis of empirical evidence	A graph shows a scientist's count of the number of electrons emitted from a radioactive sample as a function of time. The conclusion that the scientist can properly draw from the data is chosen.	23	.43	.52
	Using causal reasoning to distinguish co-incidence from cause and effect	A person crosses paths with a black cat, and later is involved in an accident. The reason why this occurred is chosen.	19	.60	.52
	Using causal reasoning to distinguish correlation from cause and effect	A swimmer at a beach notes that ice cream sales affect the number of shark attacks on swimmers, because the higher the ice cream sales, the greater number of shark attacks on swimmers. The problem with this statement is chosen.	26	.44	.50
	Using data and math in the solution of real-world problems	A data table shows how many days a patient is cured after taking a certain dosage of new medication. The conclusion that can be drawn from this data is chosen.	22	.62	.53
	Using proportional reasoning to make decisions	The relationship U = q/r is considered. The statement of U increasing or decreasing due to q and r being changed by various factors is chosen.	24	.44	.53

Note. Lower performing questions that could possibly be revised are marked with "+". Question numbers correspond to the test questions found in Appendix A.

Findings and Results

Pilot Test

As was stated previously, the pilot test failed on both the mean score and KR-20 values. The 30% mean score was well below 50%. The KR-20 of .68 was less than .80. This meant that the analyses of individual questions had to be considered in order to seek out underperforming questions that were decreasing the overall test values. As a result, seven questions were eliminated and 13 were revised or replaced.

Eliminated Questions. The eliminated questions were associated with multiple question scientific reasoning skills. By removing these questions, every skill had one aligned question. Having one question was the intended goal before the pilot test was administered. However, the decision to choose which of the multiple questions was best for its scientific reasoning skill could not be made without knowing which questions were statistically acceptable. Consequently, it was deemed necessary to run the multiple questions through the pilot test analysis to obtain their statistical values. Questions with higher statistical values were then given more weight in the decision making process, but these values were not the only factor that influenced the decision.

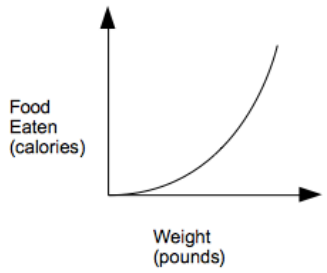
Concluding had one question outperform the other question on both statistical measures. Furthermore, this higher performing question was considered acceptable with respect to these measures. Question 5 in Table 9 had an item difficulty index of .53, which fell between the.40 and .60 values typically found on norm-referenced tests, and a point-biserial discrimination index of .34, which was above the desirable .20 value. In comparison, question 7 in Table 4 had an item difficulty index of .11, which did not fall between .40 and .60, and a point-biserial discrimination index of .26. As a result,

question 5 was chosen strictly on the basis of having higher and more acceptable statistical values.

Like concluding, using correlational thinking had one question outperform the other question on both statistical measures. Also, this higher performing question was considered acceptable with respect to these measures. Questions 14 and 15 in Table 9 had item difficulty indices of .29 and .55 and point-biserial discrimination indices of .29 and .36 respectively. These values showed that question 15 performed better, and that it could be considered statistically acceptable with respect to both measures. Question 14 on the other hand had a low item difficulty index value. Regardless, the decision was made to keep question 14 with its acceptable point-biserial index value, because the question better represented using correlational thinking. Although question 15 focused on correlation, the question was tailored more for precision in measurement with its graphical comparisons.

Interpreting quantifiable data to establish laws using logic had four questions with acceptable point-biserial indices above .20, but unacceptable difficulty indices below .40. Questions 22, 23, 24, and 25 in Table 9 had item difficulty indices of .12, .21, .29, and .18 and point-biserial discrimination indices of .27, .24, .37 and .34 respectively. Although questions 24 and 25 had the highest point-biserial discrimination index values, these questions were eliminated because they did not align to the skill as well as questions 22 and 23. Both questions did not require any establishment of laws from the data interpretation. Whereas questions 22 and 23 specifically asked for combined relationships of variables from data that were essentially laws. After eliminating questions 24 and 25, the choice was made to keep question 22 due to its higher point-

biserial index value. This meant that question 22 had to be revised in order to increase its relatively low item difficulty index value. This modification entailed stating the proportional relationship of each graph, so there was no need take this step before combining the proportional relationships. For example, $F \propto W^2$ was written above the following graph:



Using causal reasoning to distinguish co-incidence from cause and effect had one question that was considered acceptable and another considered unacceptable with respect to both statistical measures. Questions 27 and 28 in Table 9 had item difficulty indices of .44 and .21 and point-biserial discrimination indices of .45 and .02 respectively. Question 27 was chosen strictly on the basis of having the acceptable statistical values.

Using proportional reasoning to make decisions had one question outperform the other question on both statistical measures like was the case with concluding and using correlational thinking. Once again, this higher performing question was considered acceptable with respect to these measures. Questions 31 and 32 in Table 9 had item difficulty indices of .49 and .29 and point-biserial discrimination indices of .46 and .30 respectively. These values showed that question 31 performed better, and that it could be considered statistically acceptable with respect to both measures. Question 32 on the

other hand had a low item difficulty index value. Regardless, the decision was made to keep question 32 with its acceptable point-biserial index value and make a relatively minor revision, because the question was better suited scientifically for using proportional reasoning to make decisions. Question 31 was based more on mathematical ratios instead of how forecasting consequences of variable changes in a mathematical law. The minor revision to question 32 involved simplifying the relationship $U = kQq/r$ where k and Q are constants to $U = q/r$ without constants.

Revised and Replaced Questions. In addition to the elimination of the seven questions and revision of the two questions that were chosen, 11 other questions needed to be revised or replaced with the intent of increasing their statistical values. Question 3 in Table 9, which had an item difficulty index of .16 and a point-biserial discrimination index of .18, appeared to be too complex for students. The simplification of broms sharing three characteristics being reduced to two shared characteristics during the expert review process was apparently not substantial enough. The question was simplified further by reducing the number of sample broms from five to three, non-broms from five to three, and possible broms to choose as part of the answer from six to four.

Question 8 in Table 9, which had an item difficulty index of .21 and a point-biserial discrimination index of .14, was completely replaced. The best estimate for how many heartbeats in 25 years was most likely too challenging because making the necessary conversions from years to minutes required too many calculations that students had trouble making without a calculator. Estimating cricket chirps in 24 hours alleviated the need for a calculator by reducing the number of conversions.

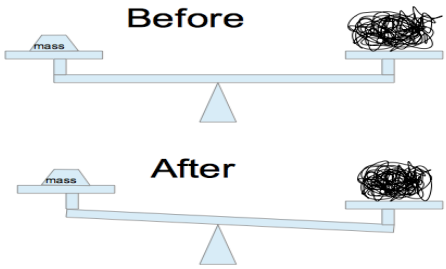
Question 9 in Table 9 had an item difficulty index of .20 and a point-biserial

discrimination index of .25. The cause for a low item difficulty index was mostly due to a distractor being chosen on 31% of the tests. This distractor could easily be interpreted as correct if students focused merely on percentages when choosing the best research question. The table that showed the patterns of behavior of men who recovered from baldness in the last few months was as follows:

Percent of men with the following traits:	
15%	Have lost weight during the past year.
23%	Have gained weight over the past year.
83%	Have recently taken aspirin daily to prevent heart attack.
26%	Use a particular type of hair shampoo.
98%	State that they enjoy watching television as a lifelong habit.
12%	Have reduced their exposure to sunlight.

Naturally, 31% of students looked at the 98% connected to watching television, and gravitated toward the best research question being, “Does watching television affect baldness?” In other words, these students did not consider the patterns of behavior that could realistically cause recovery from baldness, such as recently taking aspirin daily. To reduce the chances of this distractor being chosen the following statement was included in the question: Caution: Be certain to consider the connection between possible cause and effect, and not just percentages.

Question 12 in Table 9, which had an item difficulty index of .20 and a point-biserial discrimination index of .23, was most likely affected by an image that was not explicit enough, and students choosing answers based on their own preconceived explanations. The image that showed the mass of steel wool before and after burning was as follows:

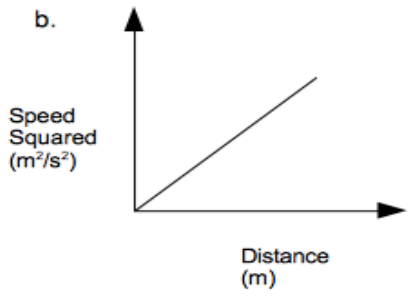


This image was supposed to be a representation of steel wool increasing mass through the burning process, which may have not been completely clear. To improve the clarity of what was occurring, wool = mass was written next to the before balance and wool > mass next to the after balance. To hinder students from choosing the explanation of what occurs during the burning of steel wool was correct based on their own preconceived explanations, the following statement was included in the question: Caution: A correct statement is not always the answer to a given question.

Question 16 in Table 9 had an item difficulty index of .23 and a point-biserial discrimination index of .29. The low item difficulty index probably stemmed from the question being too vague. Asking which variables were related without knowing what constituted a relationship between variables added a step to the question that may have troubled students. Instead of asking which variables were related, the question was changed to asking which variables were directly proportional to one another in a mathematical sense.

Question 17 in Table 9 had an item difficulty index of .19 and point-biserial discrimination index of .13. These low values provided thoughts of replacing the question. However, simplifying the graphs in the answer choices seemed to be a viable option. The graph choices were straight-line graphs that included squared variables from

the data set. For example one of the graphs was as follows:



In doing so, students had to figure out how to manipulate the values of the variables so that they would be plotted linearly, a process that many have probably not attempted previously. Instead having answers that required this process, all of the squared variables were eliminated and the straight lines were replaced with curves that reflected how the variables were graphically related.

Question 18 from Table 9 had an item difficulty index of .29 and a point-biserial discrimination index of .27. The reason for the low item difficulty index was that one of the incorrect answer statements was confusing and 21% of students chose it. This incorrect answer stated that if the mass of both of the substances (A and B) is increased by the same amount, the volume of substance A would increase more. This may have been too counter-intuitive for students, because substance A had a higher slope line on the mass-volume graph, meaning the mass of substance A increased at a greater rate, but the volume increased at a lesser rate than substance B. The incorrect answer was changed to state that if the volume of either substance is increased, the mass of substance B would increase at a greater rate. Then, students could more easily correlate the lower slope of B with a lesser rate of mass increase.

Question 20 in Table 9, which had an item difficulty index of .18 and a point-

biserial discrimination index of .18, was also a candidate to be replaced. However, changes were made to an incorrect answer and the correct answer. The incorrect answer stated that heavier objects fall faster than lighter objects, which is a misconception that students accept as true. To deter students from selecting this answer based on their misconception of falling objects, the statement was changed to state that lighter objects fall faster than heavier objects. The correct answer stated that the experiment does not prove anything because it is not a proper experiment. This answer assumed that students would understand that air resistance not being controlled equated to an improper experiment. To make it clearer as to why the experiment did not prove anything, "it is not a proper experiment" was changed to "does not take into account wind resistance".

Question 29 in Table 9 had an item difficulty index of .21 and a point-biserial discrimination index of .39. The relatively high point-biserial index could have made this question acceptable, but a revision was made to increase the item difficulty index. One of the answers stated that there is no relationship between ice cream sales and the number of shark attacks, which was incorrect. However, students could have interpreted that there was a relationship, because the swimmer in the question noted that higher ice cream sales meant a greater number of shark attacks. This answer was changed to such a situation is impossible so that a misinterpretation of the relationship would not direct students to the answer.

Question 30 in Table 9, which had an item difficulty index of .15 and point-biserial discrimination index of .07, was collectively the worst performing question on the test. Once again, a replacement was considered, but seeing that a distractor was chosen on 32% of the tests, a revision was a sensible option. The distractor stated that the

count is completely random, which could be considered correct because the count fluctuated between higher and lower values as a function of time. Although there was this fluctuation, the count was fairly centered around a 10100 value, which was best described by the answer that stated that the count is chaotic but generally constant. Unfortunately, there were too many students who thought the fluctuation was random instead of chaotic but generally constant. Two changes were made as a result. The distractor was changed to state that none of the above answers describes this situation. The correct answer was changed to state more descriptively that the count is somewhat chaotic but fairly well centered around 10100 counts.

Question 33 in Table 9, which had an item difficulty index of .11 and point-biserial discrimination index of .34, was replaced due to its low item difficulty index. In fact, the correct answer had the lowest percent of responses. There appeared to be two issues that were cause for concern. First, students were required to perform the necessary conversions in determining the number of seconds in a year before they could arrive at a comparative number to 3×10^7 seconds. Second, students may have thought of a light year as a unit of time, and accepted seconds as being correct. Developing a question with minimal converting math and less confusing units than a light year seemed to be the correct course of action in this case.

Second Test

Judging by the statistical measures of a valid and reliable test, the question eliminations, revisions, and replacements were considered a success. The second test had a mean score of 11.18 out or 26, or 43%. This percentage was relatively close to the ideal 50%. Having all but four questions fall within or close to the item difficulty index range

of .40 to .60 found on typical norm-referenced tests was most likely the cause for a satisfactory percentage. More importantly than the mean score percentage, the KR-20 of .85 was above the .80 value expected for the SAT and GRE. This elevated KR-20 was most likely attributed to all but three questions with point-biserial discrimination indices significantly above .20. Although the mean score percentage and KR-20 reflected a valid and reliable test, there was room for improvement with those lower performing questions that could make for an even better performing test.

The item difficulty indices for each question fell into three categories: within range, close to range, and unacceptable. Within range meant that the item difficulty index fell within .40 to .60. Eleven questions (5, 6, 7, 16, 17, 19, 20, 23, 24, 25, and 26 in Table 10) had indices within range. Close to range represented an item difficulty index between .30 and .40, and .60 and .70. Eleven questions (1, 2, 4, 9, 10, 11, 13, 14, 18, 21, and 22 in Table 10) had indices were close to range. Unacceptable was all other item difficulty indices below .30 and above .70. Questions 3, 8, 12, and 15 in Table 10 had unacceptable indices. Ideally, all questions should fall within range in order to be deemed acceptable, but questions that were close to range were considered acceptable if the point-biserial discrimination index was substantially above .20. Knowing that only questions 1, 3, and 15 in Table 10 were below or slightly above .20, and the rest of the questions were at least .35, all close to range questions with the exception of question 1 were acceptable. Regardless, a closer look at the lower performing questions should shed some light on the issues and possible changes.

Question 1 in Table 10 was somewhat of an outlier with respect to its item difficulty and point-biserial discrimination indices. The item difficulty of .69 was the

highest on the test while the point-biserial of .12 was the lowest. These extreme high and low values alone did not make this question an outlier however. No other question followed this inverse type relationship. All questions with elevated item difficulty values had higher point-biserial values, and those with lower point-biserial values had depressed item difficulty values. Ultimately, the high item difficulty value in this case represented a question that was too easy to properly discriminate between higher and lower performing students. These results are somewhat of surprise considering that this question had a point-biserial index of .30 on the pilot test coupled with a within range item difficulty index of .58. At this point, it may make sense to change the answer choices so that the correct answer is not as obvious.

Question 3 in Table 10 had an item difficulty index of .16. This question required revisions in every step of the development process. As a result of the expert review, the broms were reduced from three shared characteristics to two. The pilot test analysis was the cause for the number of broms, non-broms, and possible brom answer choices to be reduced. Ironically, the reductions that were made after the pilot test resulted in the item difficulty index lowering from .34 to .16. This question might need to be replaced due to the low item difficulty index coupled with a low point-biserial discrimination index of .18 after all of the revisions. The only other option might be to reduce the number of brom shared characteristics or the number of broms, non-broms, and possible brom answer choices. Either way, this question needs to be addressed in some manner.

Question 8 in Table 10 had an item difficulty index of .23. Like question 3, this question needed revisions following the expert review and pilot test. The difference between these questions is that there was a small improvement over the pilot test item

difficulty index of .20. Also, it is possible that this question could remain as is based on its high point-biserial discrimination index of .40. Regardless, the biggest problem was most likely a distractor being chosen on 30% of the tests. This distractor stated that the best research question for patterns of behavior of men who recovered from baldness was, "Does the use of a particular type of shampoo affect baldness?" In a sense, this was a welcome issue, because the distractor on the pilot test stated, "Does watching television affect baldness?" Regardless, this new distractor is problematic in its own manner. The table stated that 26% of men who have recovered from baldness use a particular type of shampoo. This low percentage should have veered students away from the distractor. However, the added statement, "Caution: Be certain to consider the connection between possible cause and effect, and not just percentages," may have had students thinking that shampoo was the best answer because it has the most direct effect on hair growth. As a result, changing this statement so that students would not be so prone to merely focus on cause and effect could be the best course of action.

Question 12 in Table 10 had an item difficulty index of .25, which was also a small improvement over the pilot test value of .20. As was the case for question 8, a high point-biserial index of .49 was an indication that the question could remain unchanged. Keeping the question unrevised might need to be the case because there are no glaring problems, such as a distractor.

Question 15 from Table 10 was similar to question 3 in that it had low item difficulty and point-biserial discrimination index values. A replacement for this question may be necessary as well. As part of the pilot test revisions, the graphs were changed from straight-line graphs that included squared variables to curve function graphs in

hopes of bringing up the relatively poor statistical values. Unfortunately, this revision kept the item difficulty index at .19, and only moved the point-biserial discrimination index from .13 to .23. Regardless, a possible change that could be made to this question instead of replacing it is simplifying the data set to reflect a linear relationship.

CHAPTER V

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Summary of the Research Problem, Methods, and Findings

The daunting task of achieving scientific literacy and, in turn, developing scientifically literate citizens is the main goal of science education. What makes reaching this goal so daunting is that it is based on a term that has many loaded definitions. Fortunately, among these definitions of scientific literacy are dimensions that exemplify a "truly" scientifically literate person. These dimensions break scientific literacy into digestible pieces, which is much needed if we want to assess our progress toward the goal. Having one assessment instrument that encompasses scientific literacy, as a whole would be much too long. A basket of assessments, on the other hand, that individually address each dimension is much more palatable. One dimension that has been addressed through this research is scientific reasoning or what Shamos (1995) calls, the use of logic for induction and deduction.

Previous work by Lawson (1978, 2000) and Han (2013) brought forth tests associated with scientific reasoning. These tests were based on a list of six to eight reasoning dimensions not entirely science related. When utilizing Wenning and Vierya's framework of intellectual process skills and scientific practices, one can see that the six to eight reasoning dimensions can be expanded to 31 scientific reasoning skills ranging

from the most rudimentary to those of a hypothetical scientist. Consequently, there appeared to be the need to develop a test that is based on this framework that includes these skills. Then, scientific reasoning can be assessed more systematically and comprehensively.

The scientific reasoning test was developed for this very reason. The test began as a pool of 36 multiple-choice questions aligned to 31 scientific reasoning skills found in the six different levels of increasing intellectual sophistication (rudimentary, basic, intermediate, integrated, culminating, and advanced) in Wenning and Vierya's framework. These test questions as well as the framework were sent to a panel of five expert reviewers. The reviewers determined if the test had construct validity by ensuring that the framework was comprehensive and properly defined, and content validity by checking to see if the questions aligned to the skills found in the framework. The reviewers also provided feedback concerning questions that were inaccurate, incomplete, confusing, poorly worded, illogical, and/or had multiple or no correct answers. Based on the commentary from these reviewers, the number of questions was reduced to 33 aligned to 26 scientific reasoning skills. The reduction in skills was attributed to the advanced level of the framework no longer being part of the test.

Following the expert review, a pilot test with these 33 questions was administered 540 students in 9^{th}, 10^{th}, 11^{th}, and 12^{th} grade science classes. The pilot test had a range of scores of 28, mean score of 9.90 (30%), standard deviation of 4.34, variance of 18.80, and KR-20 of .68. The test failed on both the mean score and KR-20 values. The 30% mean score percentage was well below the ideal 50% value designed to produce the maximum possible spread among scores. The KR-20 of .68 was less than the expected

.80 value of the SAT and GRE. The failure on both measures was attributed to a collection of lower performing questions that were eliminated, revised, or replaced. Question performance was based on the item difficulty index and point-biserial discrimination index for each question. An item difficulty index between .40 and .60 and a point-biserial discrimination index above .20 were considered the benchmark values for an acceptable question. The seven eliminated questions were attached to four scientific reasoning skills, each containing two to four questions. The intent of the elimination was to reduce the number of questions to one for each of these skills. The statistical values played a role in deciding which questions should be eliminated. However, some questions with lower values were kept because they better represented the skill being addressed. The 12 questions that were revised or replaced had a myriad of issues that needed to be rectified in order to increase their statistical values. All questions had unacceptable item difficulty indices between .11 and .29. Six of these questions had unacceptable point-biserial discrimination indices between .07 and .18. The other six questions had acceptable point-biserial discrimination indices between .23 and .39, but the low item difficulty indices made it necessary for changes.

After changes were made to the pilot test, a second test with 26 questions was administered 379 students in 9^{th}, 10^{th}, 11^{th}, and 12^{th} grade science classes. The second test had a range of scores of 26, mean score of 11.18 (43%), standard deviation of 5.61, variance of 31.45, and KR-20 of .85. The test passed on both the mean score and KR-20 values. The 43% mean score percentage was relatively close to the ideal 50% value designed to produce the maximum possible spread among scores. The KR-20 of .85 was greater than the expected .80 value of the SAT and GRE. All but four questions fell

within or close to the item difficulty index range of .40 to .60, which was most likely the cause for a satisfactory mean score percentage. The four questions that were further outside the range had item difficulty indices of .16, .19, .23, and .25. All of the other questions had item difficulty indices between .35 and .69. All but three questions had point-biserial discrimination indices significantly above .20, which was most likely the cause for an elevated KR-20. The three questions with relatively low values had point-biserial discrimination indices of .12, .18, and .23. All of the other questions had point-biserial discrimination indices of at least .35. Overall, five questions were responsible for these underperforming values. Revisions could possibly be made to increase the performance of these questions to a more acceptable level, which should consequently heighten the performance of the test even further.

Conclusions and Implications

The purpose of this study was to create a test that would answer the following questions:

- Can a valid scientific reasoning test for high school science students be created from the defined scientific reasoning skills in Wenning and Vierya's intellectual process skills and scientific practices framework?
- Can a reliable scientific reasoning test for high school science students be created from the defined scientific reasoning skills in Wenning and Vierya's intellectual process skills and scientific practices framework?
- Can a scientific reasoning test for high school science students address skills that go above and beyond the dimensions addressed by the *Lawson Classroom Test of*

Formal Reasoning and *Inventory for Scientific Thinking and Reasoning Assessment*?

With respect to the question concerning validity, the answer is yes. The test is based on a framework of skills vetted by a panel of five expert reviewers to ensure that these skills are properly defined and collectively comprehensive enough. The reviewers also aligned the pool of scientific reasoning test questions to the skills. The combination of these actions provides the test with construct and content validity.

With respect to the question concerning reliability, the answer is yes. The test questions were administered to over 800 students during two rounds of testing. The end result was a test that had a mean score percentage close to ideal and KR-20 greater than the expected value of the SAT or GRE. A large sample size with these acceptable statistical values represents a test with reliability. Even so, five of the 26 questions could be revised to make the test more reliable.

With respect to the question concerning addressed skills that goes above and beyond the dimensions of the *Lawson Classroom Test of Formal Reasoning* and *Inventory for Scientific Thinking and Reasoning*, the answer is yes. The test is based on a framework of defined scientific reasoning skills that is more comprehensive than the list of dimensions of Lawson and Han. The 26 scientific reasoning skills taken from Wenning and Vierya's (2015) framework of intellectual process skills and scientific practices are vastly more numerous than six and eight scientific reasoning dimensions of Lawson and Han, respectively.

Recommendations for Future Research

Research moving forward can come in two forms. First, further research can be done on the test. The five lower performing questions could be changed based on the suggestions discussed in the second test findings and results. Following these changes, this third generation test could be administered to another batch of high school science students, and put through the same analysis as the previous two tests. Assuming the third test changes are successful, this test could also be administered to subgroups of students to see if there are any biases with respect to gender, race, socioeconomic status, etc. Furthermore, this third generation test could be put through other measures of reliability, such as test-retest or parallel forms. Second, research can be done using the test as a research study instrument. The test could be used to validate various teaching methods in order to find out if students are effectively being taught to reason like a scientist. For example, a comparative study of two teaching methods (i.e. inquiry vs. lecture) could be completed to determine which method is better. This would involve a pre-test/post-test with the intent of seeing which method promotes more growth. Another example would be for a teacher who implements the levels-of-inquiry into his or her curriculum. This would also entail a pre-test/post-test gauging the amount of student growth using this method.

REFERENCES

American Association for the Advancement of Science. (1993). *Benchmarks for Science Literacy.* New York, NY: Oxford University Press.

Ary, D., Jacobs, L. C., & Razavieh, A. (1972). *Introduction to Research in Education*. New York: Holt, Rinehart and Winston, Inc.

Brown, J. D. (2000). What is Construct Validity? *JALT Testing & Evaluation SIG Newsletter. 4*(2). 8-12.

DeBoer, G. E. (2000). Scientific Literacy: Another Look at Its Historical and Contemporary Meanings and Its Relationship to Science Education Reform. *Journal of Research in Science Teaching. 37*(6). 582-601.

DeVellis, R. F. (1991). *Scale Development: Theory and Applications.* London: Sage Publications.

Han, J. (2013). *Scientific Reasoning: Research, Development, and Assessment* (Unpublished doctoral dissertation). The Ohio State University. Ohio.

Herr, N. (2008). *The Sourcebook for Teaching Science*. California: John Wiley & Sons, Inc.

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*. *15*(1). 11-24.

Lawson, A. E., Adi, H., & Karplus, R. (1979). Development of correlational reasoning in secondary schools: do biology courses make a difference? *The American Biology Teacher, 41,* 420-425.

Lawson, A. E. (2005). What Is the Role of Induction and Deduction in Reasoning and Scientific Inquiry? *Journal of Research in Science Teaching. 42*(6). 716-740.

National Assessment Governing Board. (2008). *Science Framework for the 2009 National Assessment of Educational Progress*. Washington DC: US Government Printing Office.

National Research Council. (1996). *National Science Education Standards.* Washington DC: National Academies Press

National Research Council (2013). *Next Generation Science Standards Framework.* Washington DC: National Academies Press

Office of Measurement and Evaluation of Teaching. (2015). *Test and Item Analysis*. Retrieved from http://www.omet.pitt.edu/docs/OMET%20Test%20and%20Item%20Analysis.pdf.

Professional Testing, Inc. (2015). *Building High Quality Examination Programs.* Retrieved from http://www.proftesting.com/test_topics/steps_9.php.

Roberts, D. A. (2007). Scientific literacy/science literacy. *Handbook of Research on Science Education*. 729-780.

Siegal, M. (2003). Cognitive Development. In A. Slater & G. Bremner (Eds.), *An Introduction to Developmental Psychology* (pp. 189-210). Malden, MA: Blackwell.

Shamos, M. (1995). *The Myth of Scientific Literacy*. New Brunswick, NJ: Rutgers University Press.

Wenning, C. J. (2006). Assessing nature-of-science literacy as one component of scientific literacy. *Journal of Physics Teacher Education Online, 3*(4), 3-14.

Wenning, C. J. (2007). Assessing inquiry skills as a component of scientific literacy. *Journal of Physics Teacher Education Online, 4*(2), 21-24.

Wenning, C.J. (2010). Levels of Inquiry: Using inquiry spectrum learning sequences to teach science. *Journal of Physics Teacher Education Online, 5*(3), 11-20.

Wenning, C. J. & Vierya, R. (2015). *Teaching High School Physics, 1*. Publisher: Authors.

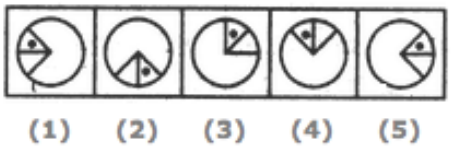
Wood, T. (1998). Alternative patterns of communication in mathematics classes: Funneling or focusing? In H. Steinbring, M. G. Bartolini Bussi, and A. Sierpinska (Eds.), *Language and Communication in the Mathematics Classroom* (pp. 167-178). Reston, VA: NCTM.

APPENDIX A

SECOND SCIENTIFIC REASONING TEST QUESTIONS

1. Students are engaged with demonstrations dealing with static electricity. Their teacher shows a variety of examples: by rubbing a balloon on his hair and showing that it sticks to the wall, by rubbing a plastic rod with a piece of fur and then using the rod to pick up tiny bits of paper, by dragging his feet on the carpet and showing he gets shocked when touching a metal door knob. When else do we see this effect in nature?
    a. When pulling wet clothing out of a washing machine.
    b. When jumping into water.
    c. When taping two pieces of paper together.
    d. When lightning strikes the ground.
    e. When putting wet clothing into a dryer.

2. Choose the figure that is different from the rest.



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

3. All of these are broms.

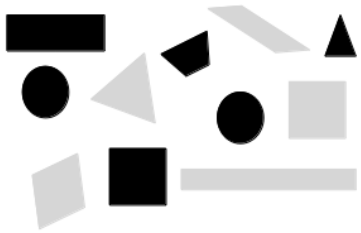
    These are not broms.

Which of these is a brom?

1 2 3 4

a. 2, 3
b. 1, 2, 4
c. 1, 3, 4
d. 2, 4
e. 1, 2

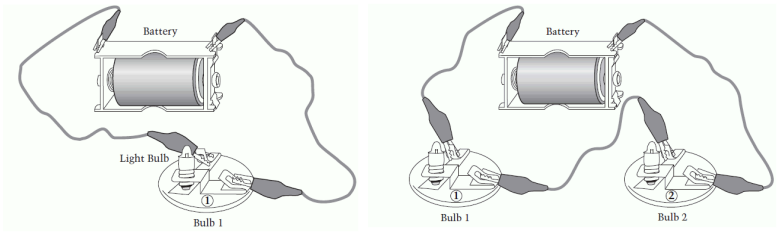
4. Which one of the following statements is correct about this group of objects?



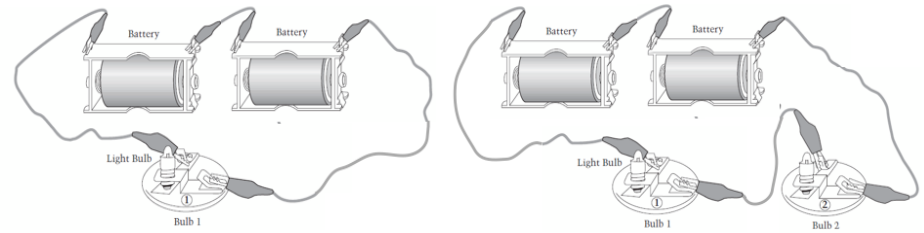
a. All square objects are black.
b. All triangular objects are gray.
c. All objects with round edges are black.
d. All objects with square corners are gray.
e. There are an equal number of black and gray objects.

5. A student has two batteries, two light bulbs, and enough wires to perform several investigations of electricity flow.

#1 #2



#3 #4



The following observations were noted:

- When comparing circuit #1 and circuit #2, the light bulb in circuit #1 was brighter.
- When comparing circuit #1 and circuit #3, the light bulb in circuit #3 was brighter.
- When comparing circuit #1 and circuit #4, the light bulbs were equally bright in both circuits.

What can the student conclude from these observations?

a. If batteries and light bulbs are added to a circuit, bulb brightness increases.
b. If batteries and light bulbs are added to a circuit, bulb brightness decreases.
c. If an unequal number of batteries and light bulbs are added to a circuit, bulb brightness stays the same.
d. If an equal number of batteries and light bulbs are added to a circuit, bulb brightness stays the same.
e. None of these conclusions are correct and more comparative investigations need to be performed.

6. Arrange the planet labels in the table in order of increasing distance from the sun. (Closest planet first to farthest planet last.)

Planet	Distance from sun (millions of km)	Time for complete trip (yr)	Radius of planet (km)
A	150	1.00	6371
B	?	12.0	69911
C	230	1.88	3397
D	58	0.241	2440
E	4500	165	55528
F	?	84.0	51118
G	?	0.698	12104

a. D, C, A, G, F, E, B
b. E, F, B, C, A, G, D
c. D, G, A, C, B, F, E
d. D, A, C, B, G, F, E
e. Unable to determine.

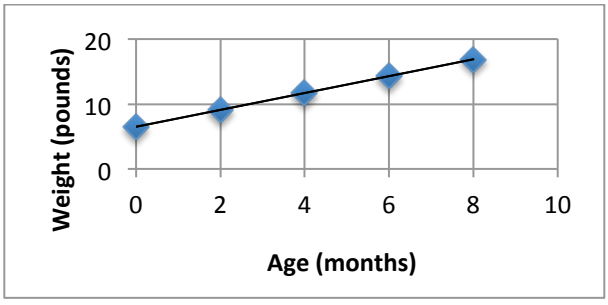
7. If a cricket chirps at a constant rate of 2 times per second, about how many chirps will it make over the course of 24 hours?
   a. 48
   b. 3,600
   c. 7,200
   d. 86,000
   e. 173,000

8. A dermatologist (skin doctor) is interested in knowing what cures baldness in men. She has observed 256 men – all of whom seem to have recovered from baldness within the last few months – with the following patterns of behavior:

Percent of men with the following traits:	
15%	Have lost weight during the past year.
23%	Have gained weight over the past year.
83%	Have recently taken aspirin daily to prevent heart attack.
26%	Use a particular type of hair shampoo.
98%	State that they enjoy watching television as a lifelong habit.
12%	Have reduced their exposure to sunlight.

   Which of the following is the *best* research question to ask based on these data? Caution: Be certain to consider the connection between possible cause and effect, and not just percentages.
   a. "Does aspirin cure baldness?"
   b. "Does weight loss or gain affect baldness?"
   c. "Does exposure to sunlight affect baldness?"
   d. "Does the use of a particular type of shampoo affect baldness?"
   e. "Does watching television affect baldness?"

9. The following graph represents the relationship between the weight of a baby and its age:



   Slope = 1.3 pounds/month Y-intercept = 6.5 pounds

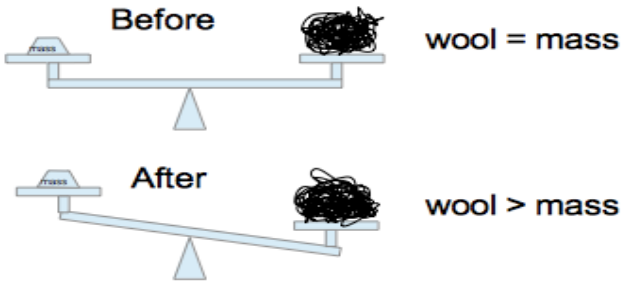
Predict the weight of the baby at 15 months, assuming the rate of growth remains constant.

a. 19.5 lbs
b. 24.7 lbs
c. 26.0 lbs
d. 32.5 lbs
e. None of these are correct.

10. What is the correct conclusion given the following line of reasoning? All apples are either red or green. All green apples are hard. These apples are all hard; therefore,
    a. all of these apples are green.
    b. all of these apples are red.
    c. some of these apples are red.
    d. there are more green apples than red apples.
    e. none of the above conclusions can be correctly drawn.

11. A capped bottle filled with water is shown to the students. (See the above figure.) Inside the bottle is an eyedropper floating just beneath the surface. The eyedropper is partially filled with water and partially filled with air. When the water bottle is squeezed, the eyedropper sinks to the bottom, but will not turn over. When the bottle is released, the eyedropper floats to the top. Water will not significantly change its volume under pressure, but air will. What accounts for the eyedropper sinking and floating?
    a. The water in the bottle compresses when the bottle is squeezed thus making the eyedropper denser than the water in the bottle.
    b. The air in the eyedropper is compressed when the bottle is squeezed causing the eyedropper to become denser than the water in the bottle.
    c. The air in the eyedropper is pushed out when the bottle is squeezed causing the eyedropper to become denser than the water in the bottle.
    d. The water in the eyedropper compresses when the bottle is squeezed thus making the eyedropper denser than the water in the bottle.
    e. None of these explanations is correct.

12. The following figure shows a balance that is measuring the mass of steel wool before and after it has burned for a short period of time. In the figure, the balancing mass is on the left side and the steel wool is on the right side.



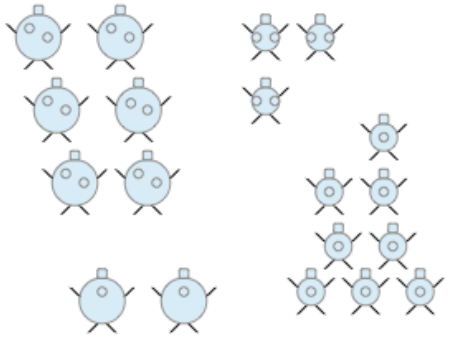
Three students have different explanations as to what occurs during the burning of steel wool:

- Student 1: Oxygen from the atmosphere combines with the steel wool, because burning is a chemical reaction that always involves oxygen.
- Student 2: Carbon dioxide from the steel wool is released into the atmosphere, because burning is a process that always involves smoke coming from the substance.
- Student 3: There is no exchange of gases between the steel wool and the atmosphere, because steel wool does not "burn" like other substances.

The figure best supports which of the students' explanations? Caution: A correct statement is not always the answer to a given question.

a. Student 1
b. Student 2
c. Student 3
d. Student 1 and Student 2
e. There is not enough evidence in the figure to support any of the students.

13. The following picture depicts a collection of sea turtles with different traits. All of the sea turtles are either big or small and have either one or two circular markings on their backs.



What can you say about the relationship between the size of the sea turtle and number of markings?

a. For most sea turtles there appears to be a relationship between size and number of markings.
b. For some sea turtles there appears to be a relationship between size and number of markings.
c. There appears to be no relationship between size and number of markings.
d. The relationship would be stronger if there were more big sea turtles with one marking on their backs.
e. The relationship would be stronger if there were more small sea turtles with two markings on their backs.

14. The following data were collected of various materials placed in a container of fluid.

Mass (g)	Density (g/cm^3)	Volume Submerged (cm^3)	Buoyant force (N)
8	2.0	2.67	80
20	5.0	4.00	120
6	1.5	2.00	60
36	9.0	4.00	120
2	0.5	0.67	20

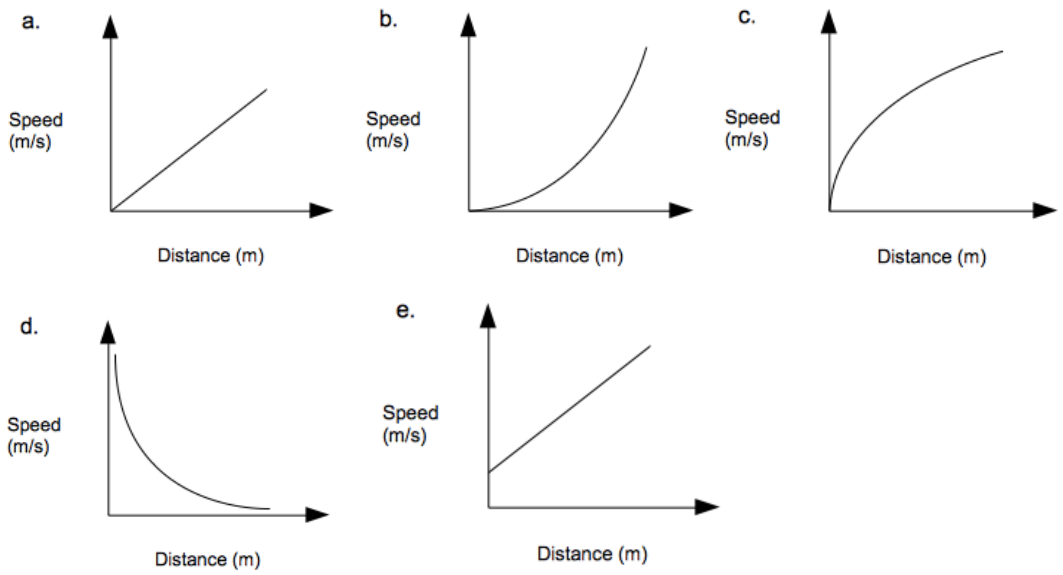
From the above data, which of the following appear to be directly proportional to one another in a mathematical sense (e.g., double X and Y doubles)?

1. Mass and density
2. Mass and volume submerged
3. Mass and buoyant force
4. Density and volume submerged
5. Density and buoyant force
6. Volume submerged and buoyant force

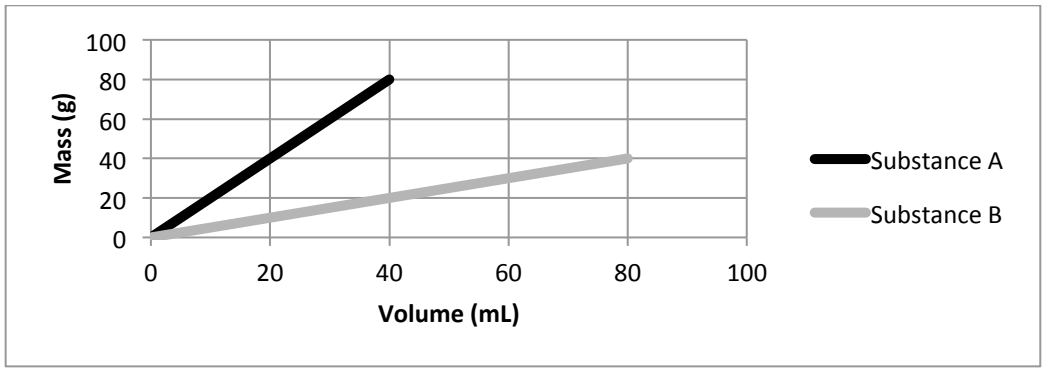
a. 1, 2
b. 3, 4
c. 5, 6
d. 1, 6
e. 2, 5

15. Which graph best shows the relationship between distance and speed in the following data set?

Distance (m)	Speed (m/s)
4	2
9	3
16	4
25	5
36	6



16. The following graph shows the relationship between mass and volume of two substances:



Which statement best describes the mass and volume relationships of the substances?

a. If the mass of either substance is increased, the volume of either substance will also increase.
b. If the volume of either substance is increased, the mass of either substance will also increase.
c. If the volume of either substance is increased, the mass of substance B will increase at a greater rate.
d. Answers a and b are correct.
e. Answers a, b, and c are correct.

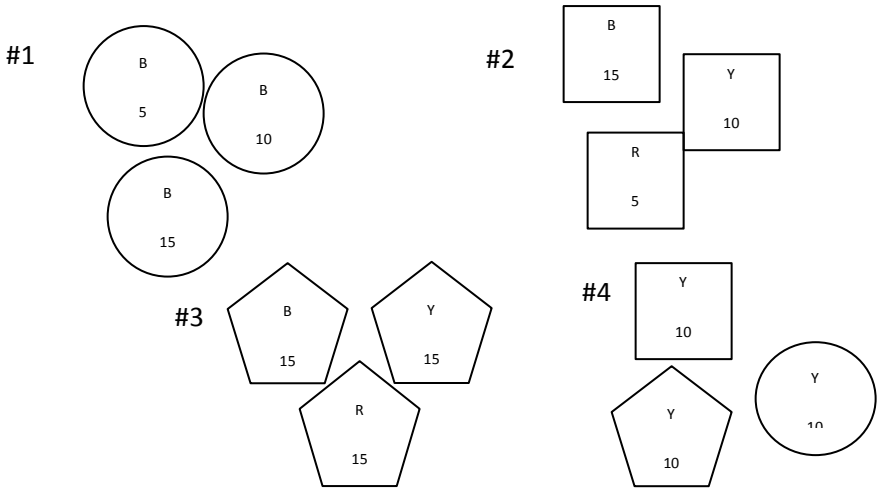
17. A pair of identical springs is used to pull a cart up a hill at a constant speed in two situations. In the first situation the springs are connected to the cart side by side. In the second situation, the springs are connected to the cart one after another. In the first situation, the springs extend a given distance. In the second situation, the springs stretch twice as much as in the first situation. What problem might be studied based on these situations?

a. How does the steepness of the hill determine the amount of spring stretch?
b. How does the material of the spring determine the amount it stretches?
c. How does the arrangement of spring determine the amount of their stretch?
d. How does the mass of the cart determine the amount of spring stretch?
e. How does the speed of the cart determine the amount of spring stretch?

18. A piece of paper and a round stone are released simultaneously from rest at the same height above the floor as a test of the claim that heavier objects fall faster than do lighter objects. The stone hits the floor long before the paper. Students discuss the observation. Which of the following student statements is correct?
    a. This is proof that lighter objects fall faster than heavier objects.
    b. This is proof that larger objects fall faster than smaller objects.
    c. Gravity is pulling harder on the rock than the paper so it must fall faster.
    d. Round objects fall faster than do flat objects.
    e. This doesn't prove anything because it does not take into account wind resistance.

19. A black cat crossed Babbs' path yesterday and, sure enough, she was involved in an accident later that same afternoon. Why did this occur?
    a. Black cats cause bad luck.
    b. One can't really say; there is no relationship between black cats and bad luck.
    c. Black cats cause accidents.
    d. An accident will always occur the day after a black cat crosses one's path.
    e. Babbs' friend had this same thing happen a year ago.

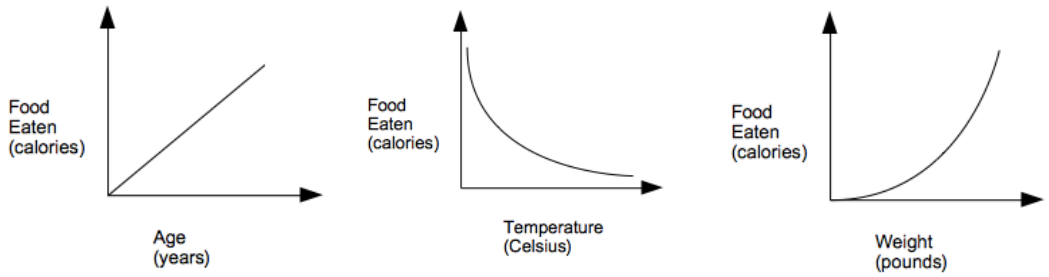
20. A student wants to design an experiment to determine which characteristics affect how quickly objects sink to the bottom of a container filled with water. The student is given a collection of objects (shown below) with various weights, shapes, and colors. The number within each shape represents the weight. The letter in each shape represents the color (R = red, B = blue, Y = yellow). Assuming all volumes are the same, which group of objects would this student need to choose to determine if shape has an effect on the sinking rate?



a. #1
b. #2
c. #3
d. #4
e. There is no group of objects found in this collection.

21. A scientist studying a newly discovered species noticed that the eating habits of this species seemed to depend on the weight (W) and age (A) of the creature, and the temperature (T) of the environment. The following three graphs shows how each variable is related to food eaten (F) by the species.

$F \propto A$ $F \propto \frac{1}{T}$ $F \propto W^2$



What is the correct single combined relationship for food eaten, weight, age, and temperature?

a. $F = AT^2/W$
b. $F = AW/T$
c. $F = AT/W$
d. $F = AT/W^2$
e. $F = AW^2/T$

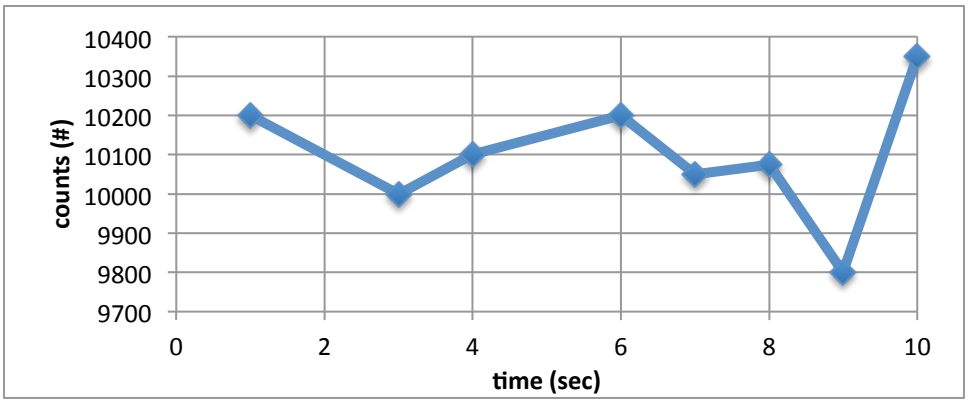
22. Doctors give patients with a common cold the following doses of a new medication based upon their body weights. Below is a table of representative data from the research.

Patient No.	Dose (milliliter per pound)	Outcome
1	1.1 ml/lb	Patient cured in 9 days
2	1.9 ml/lb	Patient cured in 7 days
3	3.9 ml/lb	Patient cured in 4 days
4	7.2 ml/lb	Patient cured in 2 days
5	9.1 ml/lb	Patient cured in 1 day

What conclusion can the researchers properly draw from these data?

a. The greater the dose the slower the cure.
b. The greater the dose the quicker the cure.
c. A dose between 3.9 ml/lb and 7.2 ml/lb is the quickest cure.
d. Any dose greater than 9.1 ml/lb will cure a patient in more than 1 day.
e. The dose does not affect how quickly the patient is cured.

23. A scientist is making a count of the number of electrons emitted from a radioactive sample as a function of time. At the end of each second, the number of electrons emitted during the past second is recorded and the following graph is generated. Which of the following conclusions can the scientist properly draw from the data for the time interval observed?



a. The count is generally decreasing.
b. The count is uniformly constant.
c. The count is generally increasing.
d. The count is chaotic but fairly centered around 10100 counts.
e. None of the above answers describes this situation accurately.

24. Consider the following relationship between variables: U = q/r. Which of the following is a correct statement given this relationship?
    a. If q doubles and r doubles, then U increases.
    b. If q doubles and r halves, then U increases.
    c. If q halves and r doubles, then U increases.
    d. If q halves and r halves, then U decreases.
    e. If q remains the same and r halves, then U decreases.

25. A bullet travels at 300 feet per second. A shooter estimates how long it takes the bullet to reach a target 1,000 feet away. She concludes "$3^1/_3$ feet". Is this answer correct or not, and why or why not?
    a. The number and unit of measure are correct.
    b. The number appears to be too small, but the unit of measure is correct.
    c. The number appears to be correct, but the unit of measure is incorrect.
    d. The number appears to be too large, but the unit of measure is correct.
    e. The number and unit of measure is incorrect.

26. A swimmer at a beach notes, “Ice cream sales affect the number of shark attacks on swimmers. The higher the ice cream sales, the greater the number of shark attacks on swimmers.” What is wrong, if anything, with this statement?
    a. Higher ice cream sales actually means a smaller number of shark attacks.
    b. Sharks do not like the taste of ice cream, so they have no reason to attack swimmers.
    c. There is nothing wrong with this statement, because ice cream sales are the cause of shark attacks.
    d. There are merely more swimmers when it is hot, and when it is hot swimmers eat more ice cream.
    e. Such a situation is impossible.