SCIENCE AND ENGINEERING PRACTICES IN SECONDARY SCIENCE

by

Derek Kelley Engebretsen

A professional paper submitted in partial fulfillment
of the requirement for the degree

of

Master of Science

in

Science Education

MONTANA STATE UNIVERSITY
Bozeman, Montana

July 2018

ACKNOWLEDGEMENT

Throughout my time in the MSSE program, I have received a great deal of help from a number of outstanding individuals. Chief among these would be Diana Paterson and John Graves, without whom my time in the program would have been entirely different. The professionalism, organization and dedication these two have displayed has defined my time in the program. I would like to thank Lindsey Hall and Dave Willey for their thoughtful feedback and insight on this work. I would also like to thank Dave and the rest of the faculty who taught elective science classes for providing me with an unparalleled field-based science education.

My parents, Arne and Kathy Engebretsen, both spent their entire careers as public-school teachers. I began learning what excellence looked like in education far before I knew I would become a teacher myself. The longer I teach, the more grateful I become for the environment they provided for me growing up. Finally, I would like to thank my wife Katie Twitchell for joining me on this journey through the MSSE program. Being able to share all of these experiences with the one you love has added a depth to the program and our relationship that I am truly grateful for.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This study investigated the effect of developing and using rubrics to assess students' abilities to plan and carry out investigations and engage in argument from evidence as defined by the Next Generation Science Standards. The intervention was carried out with a group of eighth-grade students in science and a high school physics class. A control group of similar eighth-grade students was also established in another class. Modest gains were seen with the eighth-grade intervention subgroup in a test of scientific reasoning skills, but the real value in the intervention was the ability for the assessment tools to communicate expectations for these practices.

INTRODUCTION AND BACKGROUND

Telluride Middle/High School is located in the San Juan Mountains of southwestern Colorado. The school serves a resort community of a few thousand full-time residents and has an enrollment of 254 students in the high school (THS Senior Profile, 2017). About a quarter of the students are Latino, with the majority of the rest of the school being Caucasian. A similar fraction of the school qualifies for free and reduced lunch. Southwestern Colorado is a sparsely populated part of the state, and locals often like to point out that there are no stoplights in the county. However, the school's rural nature should not be confused for a lack of opportunity, as it is consistently ranked one of the top public high schools in the state (US News & World Report, 2018).

I have been teaching science in the combined middle/high school for seven years. Each of those years has included some middle school classes, and my background as a physics education major eventually earned me the spot of teaching the physics and AP Physics courses as well. My teaching assignment also often includes a new science elective in the high school, and occasionally a math or even a PE class. Such is the nature of teaching in a small school. This past year I had four different classes: physics, AP Physics, a STEM elective class, and two sections of eighth-grade grade science. I split the eighth grade science classes with a colleague, who was gracious enough to allow me to use his students as a control group in this study. I was the only teacher for the rest of my classes, which is typical for most other secondary teachers in the district. For the past four years I have also served as the science department head in our school.

Most Telluride alumni attend college after graduation and are generally bought in to what we were trying to accomplish as an institution. In my experience as a science teacher, when I pass out a sheet of instructions for a lab, most students would get right to work following those instructions. However, students are often less clear on what the purpose of a laboratory investigation might be, and also how they could design an investigation themselves to provide evidence for or against a claim. While planning an actual experiment and answering multiple choice questions about designing an experiment are different things, the observation that my students are not as versed with experimental design also shows up in our standardized testing data. As the department head, I can say with some confidence that this situation exists despite students spending considerable time in a laboratory environment in every one of the science classes we offer. Colorado has not officially adopted the Next Generation Science Standards (NGSS), but teachers in our department still give the NGSS at least equal consideration to our state standards when discussing our approach in department meetings.

A similar problem existed with respect to engaging in argument from evidence. Students frequently argue about their ideas in science class, they often do not need to be prompted to do so. While passionate, I have found these arguments often lack a solid grounding in evidence, or a clear link between the evidence and the point they are trying to support. I have found this to be true with my students in arguments large and small, from providing a rationale for explaining why they think another classmate's work might be wrong to the arguments found in more formal, edited writing samples. This problem transcends the science department in our school. I know from conversations I've had with

other colleagues that this is a challenge that the English and Social Studies departments are working on as well.

Taken together, planning and carrying out investigations and engaging in argument from evidence are practices that I hold as high priorities in what I want to get across in my teaching. I think of my practice as one that is grounded in an inquiry-based approach, and I want students to learn to argue in an informed, logical fashion. However, when I examine what ends up being factored in to my final grades, these practices only play a minor role, and some are entirely absent. Students know that I care about these practices, but I have not leveraged my grading to increase student performance. With the college-bound students in Telluride, this amounts to a missed opportunity.

For my classroom research based on the action research model, I intended to investigate whether utilizing rubrics designed to assess the different aspects of developing and carrying out investigations and engaging in argument from evidence, as defined by the NGSS, would improve student performance with these practices. My intent was both to share the rubrics that I develop with the students to help articulate what exactly is being expected of them during these activities, and also to give students the opportunity to use the rubrics to assess their own performance.

The main question that drove my research was, Will the development and implementation of rubrics specifically designed to assess students' competence in planning and carrying out investigations and engaging in argument from evidence improve performance in these areas? I was also interested in a secondary question, Will

the development and implementation of these rubrics also have a measurable effect on student performance on standardized test questions that cover experimental design?

CONCEPTUAL FRAMEWORK

In the summer of 2011, the National Research Council published *A Framework for K-12 Science Education*, a report intended to provide guidance and recommendations for improving science education from kindergarten to 12th grade (Keller & Pearson, 2012). Outlining a three-dimensional approach to science education, this report defined these dimensions as Cross-Cutting Concepts, Disciplinary Core Ideas, and the Science and Engineering Practices. The Next Generation Science Standards (NGSS) followed shortly thereafter, fleshing these dimensions out into a set of standards that could be used and adopted by schools (NGSS Lead States, 2013). In contrast to other sets of science standards, the NGSS were clear about what students should be doing in a science class, as opposed to solely describing the content that should be taught. Each practice was unpacked and progressions were delineated, so that educators were clear on what exactly could be expected of a student for each practice at different points in a student's education.

With respect to planning and carrying out investigations, the standards focus on students' abilities to formulate a testable question, as well as a hypothesis that answers that question. Students should be able to determine what data should be collected, what tools would be needed to collect that data, and how this data should be recorded. Students need to be conscious of how much data should be collected for the results to be valid, and how sources of error might impact that data. Finally, students need to be able to identify

dependent and independent variables, and controlling for additional variables (NGSS Lead States, 2013). In addition to defining these characteristics, the NGSS also provides guidance on the progression of these skills throughout their K-12 years. For example, "Older students should be asked to develop a hypothesis that predicts a particular and stable outcome and to explain their reasoning and justify their choice. By high school, any hypothesis should be based on a well-developed model or theory" (NGSS Lead States, 2013, p.61). The emphasis in these practices, as with much of the NGSS, paints a clear picture of what students should be doing regularly in a science classroom.

These new standards were embraced by science teachers across the country, including by schools and teachers in states that had not formally adopted the standards. However, as the director of the National Science Teachers Association pointed out, "significant effort is needed to use these performance expectations to modify and guide the development of classroom/ formative and high-stakes summative assessments at all levels" (Evans, 2014, p.3). While the Science and Engineering Practices had been clearly defined, assessing students' performance in these practices remains a challenge.

If authentically assessing student progress towards what has been set out in the Science and Engineering Practices is challenging, it isn't just a problem of accurate record keeping for the teacher. For at least the past several decades, assessment has also been seen as a vehicle for additional learning opportunities, often referred to as assessment for learning. "Assessment for learning is any assessment for which the first priority in its design and practice is to serve the purpose of students' learning" (Black, Harrison, Lee, Marshall & William, 2004, p.10). To achieve this aim, the assessments of

the practices should also provide additional clarity to the student as to what is being expected of them. To fail to do so would involve missing out on a significant opportunity for teaching and learning.

One way of accomplishing this aim is with the usage of rubrics. "Assessments can serve these purposes when they are clearly linked to standards that are reflected in the rubrics used for scoring the work; when these criteria are made available to students as they are developing their work; and when students are given the opportunity to engage in self- and peer assessments using these tools" (Conley & Darling-Hammond, 2013, p.29). Not only do rubrics provide a way of accurately assessing a multifaceted objective like a NGSS Science and Engineering Practice, they also make clear exactly what and how the task is being assessed, and when put in the hands of students they provide a means for students to assess their own work and the work of their peers. This self- and peer-assessment is critical for students to achieve their learning goals, since using the rubric will make clear to them exactly what is being expected and allows them to set clear goals for future performance (Black, Harrison, Lee, Marshall, & Wiliam, 2004; Chappuis & Stiggins, 2002).

When students use rubrics for self-assessment, they are not just an effective tool for making objectives and learning goals clear. Hafner and Hafner (2003) showed that student use of a rubric to assess their peers' presentations were not only consistent with each other but were also strongly correlated with the professor's own grading of the presentations. This suggests that student scores can also be used somewhat reliably as a component of a student's grade, adding to the legitimacy of the exercise. Another

surprising result of the study was that the results were generalizable with as few as five to ten ratings, meaning that large numbers of peer assessments did not necessarily need to be carried out and compiled for the results to be reliable. In another study, Eyster (1997) showed that two different teachers with different levels of familiarity with a rubric designed to assess student performance in lab were able to generate scores which were highly consistent with one another. This suggests that teachers could collaborate on scoring student performance with a rubric without necessarily first going through an extensive process to normalize their scores.

The Framework for K-12 Science Education and the NGSS set out clear performance expectations for what students should be doing in their science classes, known as the Science and Engineering Practices. Assessing these Practices remains as an open problem in science education, but the use of rubrics appears to be a promising tool in meeting these challenges. Rubrics provide clear and tangible descriptions of what is being expected of students, and also allow students to assess themselves and their peers using the same tool. These student and peer assessments have been shown to be consistent enough that they can be used in final scoring, honoring student's effort in joining in on the assessment without detracting from the legitimacy of the final grades. This creates an authentic assessment-for-learning opportunity that can meet rigorous and meaningful standards, accurately measure progress toward those standards, and help students understand what is being expected of them all at once.

METHODOLOGY

This study was carried out in two science classes at Telluride Middle/High School. One was a middle school science class divided into two periods (*n*=37). The other was a relatively small physics class for juniors and seniors (*n*=8). A control group was established using the students from my colleague Chris Loew's eighth-grade science students, who took part in the pre- and post-testing but none of the interventions (*n*=38). The overall participation rate for the study was 82%, though only 57% in the physics class. The numbers above represent full participants. The backgrounds of the students in these classes generally mirrored the demographics of Telluride as a whole. The students in each class were predominantly white, and evenly populated with girls and boys. Four of the students in my physics class did not speak English as their first language, three of which were European students on a Rotary study abroad program. The intervention for the eighth-grade students and the physics students was similar, though the data was sometimes reported separately for the middle and high school sub-groups (*N*=83). The research methodology for this project received an exemption by Montana State University's Institutional Review Board and compliance for working with human subjects was maintained (Appendix A).

The aim of this study was to improve student performance in the NGSS science and engineering practices of planning and carrying out investigations and engaging in argument from evidence. The intervention primarily involved assessing these practices more explicitly, using rubrics to assess student performance. The Investigations Rubric was developed directly from a section within the Framework for K-12 Science Education,

where the expectations for planning and carrying out investigations was described (Appendix B). The Investigations Rubric used a four-point scale, where possible the top score corresponded to 12th grade expectations as defined in the Framework, a three equaled middle school expectations, two indicated partial proficiency, and a one did not meet any substantial expectations. Performance descriptions were copied directly from the Framework or reduced for brevity. Students could score up to 16 points on this rubric. The Argumentation Rubric was found online and used a three-point scale to assess student proficiency in scaffolding arguments (Appendix C). This rubric and all instruction around it relied on the Claim-Evidence-Reasoning structure. The maximum possible score on the Argumentation Rubric was six, as the low score in any component earned zero points.

Explicit instruction related to the use of these rubrics and structure occurred over a two-month period. During this time, the eighth-grade students studied weather and climate, while the physics students completed units on projectile and uniform circular motion. In an effort to make the intervention more robust, both classes departed from the normal curriculum near the end of the treatment to complete more labs and arguments than were normally called for during this period.

During the intervention, students were assessed using the Investigations Rubric in each lab that took a full class period or longer to complete and were periodically asked to fill out the rubric to self-assess their performance. This was largely to improve their familiarity with the practices themselves, and also served as a self-assessment of their progress. It became immediately clear that some labs would need to be reframed to allow

students more control in designing their own investigations. In total, students used the Investigations rubric in five different labs during the intervention period. Mini-lessons on various skills were interspersed between these labs to address patterns of weakness found in the rubric results, making use of the rubrics as formative assessments.

In a similar fashion, students were also asked to construct arguments using the Claim-Evidence-Reasoning structure and assessed with the Argumentation Rubric. These arguments were necessarily given in a written, worksheet form, so as to score the work in a standardized way. However, especially in my physics class where I have the students frequently present and discuss homework problems, the common language of the Claim-Evidence-Reasoning structure was stressed during the presentations and students were encouraged to construct their arguments in a similar fashion. Students were asked to complete eight of these argumentation worksheets, and again mini-lessons were interspersed to address patterns of weakness observed in the data.

To measure the impact of this intervention, several instruments were used. To assess whether students improved their ability to successfully analyze experimental situations, Lawson's Classroom Test of Scientific Reasoning was administered as a pre- and post-test (Appendix D). This 24-question multiple choice test presented students with common laboratory situations, and asks students about conclusions, explanations and/or rationales for certain results. Results from these tests were compared graphically, with a calculation of normalized gain, and a matched paired t-test. Students also completed the Student Attitudes Survey before and after the treatment which probed their attitudes towards these practices and science generally (Appendix E). Statements were given, and

students selected answers using a Likert scale ranging from 4- Strongly Agree to 1- Strongly Disagree. The differences between pre- and post-test responses were calculated and compared with the CTSR growth scores with a correlation coefficient. Any shifts in survey responses were also subjected to a t-test.

Finally, twenty-five randomly selected students from the treatment groups were interviewed on their experience following the intervention (Appendix F). The questions focused on whether students felt that their approach to labs or constructing arguments had changed at all after the rubrics were introduced. Responses were recorded and cross-referenced with any apparent patterns in the data. The ways in which each of these instruments addressed the primary and secondary question for the research is summarized in the Data Triangulation Matrix (Table 1).

Table 1
Data Triangulation Matrix

Data Source	Questions	
	Primary Question: Does assessing student performance in SEP 3 and 7 using rubrics improve student performance with those practices?	*Secondary Question:* Does an increased focus on SEP 3 and 7 improve student performance on standardized tests?
Lawson's Classroom Test of Scientific Reasoning	X	X
Student Attitudes Toward Science Survey	X	
Interview Questions	X	X
Investigations Rubric	X	
Argumentation Rubric	X	

DATA ANALYSIS

Each student in the study completed Lawson's Classroom Test of Scientific Reasoning (CTSR) as a pre- and post-test. The table below shows the results from the eighth-grade control group, the eighth-grade treatment group, and the physics students (Figure 1). Median scores rose in both of the eighth-grade groups, moving from 37.5% to 41.7% in the control group and 45.8% to 58.3% in the treatment group. On average, scores fell in the physics class, with the median score going from 70.8% to 64.6%. In both eighth-grade groups, the standard deviations increased for the post-test, while narrowing slightly in the physics group. A matched paired one tailed t-test returned a p-value for the eighth-grade intervention group of .0004, indicating it was unlikely that the increase in post-test scores was due to chance. This same test returned a p-value of .12 for the physics tests, which failed to reach the .05 standard needed to reject the null hypothesis. The eighth-grade control group's data was not compiled by student, so an unpaired t-test was run instead, and returned a p-value of .17, again failing to reach the significance level to reject the null hypothesis.

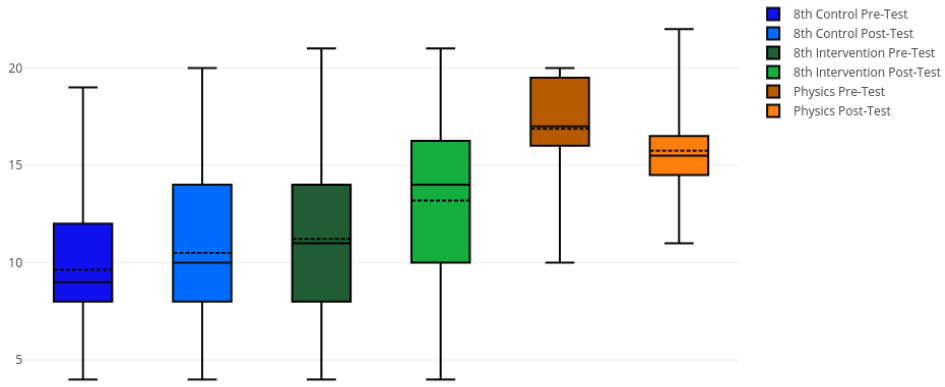


Figure 1. Lawson's Classroom Test of Scientific Reasoning data for the eighth-grade control group (*n*=38), the eighth-grade intervention group (*n*=37), and the physics class (*n*=8), (*N*=83).

Normalized gains were also calculated for each group. The control group saw an overall normalized gain of six percent, indicating that students in this group correctly answered six percent of the questions that were originally missed on the pre-test. The eighth-grade students in the treatment group saw a normalized gain of 15.4%, while the students in the physics class managed a negative normalized gain, as the average post-test score was lower than the pre-test score. A plot of pre- vs. post-test scores for both treatment groups is shown below, with eighth-grade scores shown in green and physics scores in orange (Figure 2). The control group was not sorted by student, so those data do not appear in the chart. A line with a slope of 1 has also been included. Points falling above this line represent students who improved their score on the post test, while points below show students whose scores fell.

Figure 2. Classroom Test of Scientific Reasoning pre- vs. post-test scores, (*n*=45).

For students in the treatment groups, the number of incorrect scores on each question were tabulated. This was useful in identifying areas of relative strength and weakness for the students, as well as where the most growth took place. Of the 24 questions on the test, 54% of the growth came from just five questions: numbers five, six, nine, ten, and twenty. Questions five and six were proportional reasoning questions. Nine and ten covered a pendulum lab and asked about isolating variables. Question twenty involved justifying a claim that had been made about a situation in the previous question. Question 21 saw more incorrect responses on the post-test than the pre-test by the largest margin. This question also involved testing a claim, this time proposing a further test to gain additional evidence. It was also the longest question on the test, requiring students to read about three quarters of a page to make sense of the question.

Another aspect of my analysis was to compare the scores or gains from the CTSR to other aspects of the study. A number of comparisons were made between either raw scores or gain scores on the CTSR to the responses on various survey questions. Correlation coefficients were also calculated between these measures, but few patterns emerged. One notable lack of correlation is shown in the graph of CTSR post-test scores against the survey question, “I am interested in science” (Figure 3). I predicted that students who liked science more would score higher on a test like the CTSR, or vice versa. The correlation coefficient for these two measures was .13, indicating essentially no correlation. The general interest in science indicated by both the pre- and post-test surveys mirrors a general positivity seen during the interviews, and during the year as a whole with my students. When asked if they had anything else they wanted to say, most

students declined, but when they did volunteer a response it was most often to say that they liked my class, or science generally.

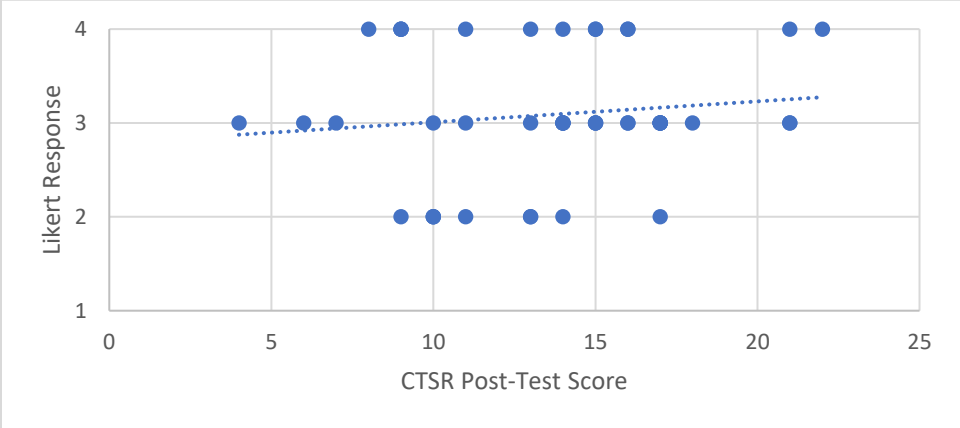


Figure 3. Classroom Test of Scientific Reasoning post-test score vs. Likert responses to the question "I am interested in science." 1=Strongly Disagree, 2=Disagree, 3=Agree, 4=Strongly Agree, (*n*=45).

Responses from the survey questions were compiled and compared from before and after the intervention (Figure 4). Generally, post-test survey responses were quite similar to the responses given before the treatment. In a few cases, students did not answer a question or gave a response that was between two values. These numbers were omitted individually, and account for the reduced number of responses in some of the questions.

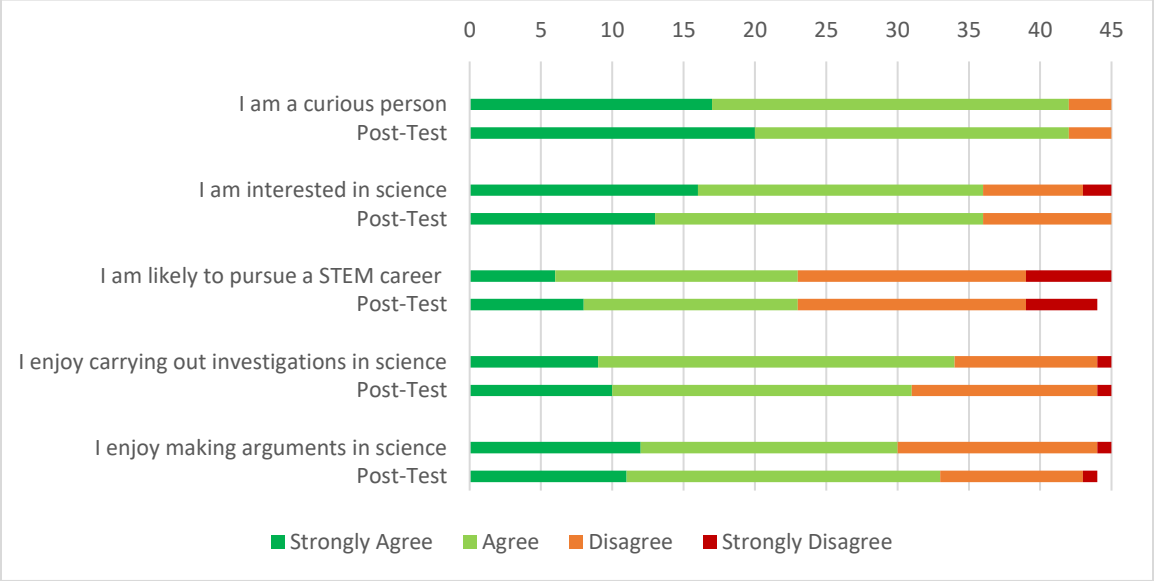


Figure 4. Likert responses to select survey questions before and after treatment, (*n*=45).

Most of the other survey questions showed similarly small shifts in student perception over the course of the treatment period. Of the eleven questions on the survey, the greatest shifts occurred in the final question, When I'm evaluating some else's argument, looking at their evidence is an important part of my analysis (Figure 5). Before the treatment, 84% of students gave an affirmative response, while after the treatment only 71% of students did so. This decrease was seen in both the eighth-grade and physics subgroups. A matched paired two-tailed t-test for the two classes combined returned a p-value of .07, narrowly missing the standard for rejecting the null hypothesis. If the students actually felt less inclined to look at evidence when evaluating another argument, the interviews indicate that many still feel that it is an important part of their own argumentation. Many students indicated that one of the benefits of the CER framework was being asked to explicitly cite evidence, as typified by this response, "Yeah, because having to write down your evidence changes your claim sometimes, having to write it down made it more clear."

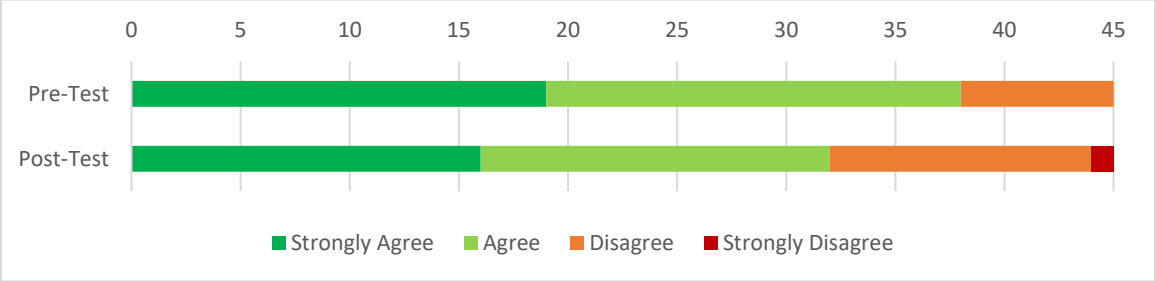


Figure 5. Likert responses to the statement, "When I'm evaluating someone else's argument, looking at their evidence is an important part of my analysis," (*n*=45)

INTERPRETATION AND CONCLUSION

Student gains on the Classroom Test of Scientific Reasoning (CTSR) were modest in the case of the eighth-grade group and disappointing in the case of my physics class. I was pleased to see that in terms of normalized gain, the students in the intervention group showed about three times more growth than those in the control group. Of course, attributing this growth to the intervention would be difficult, and even if I were able to show causation, I would then have to explain why that same intervention caused my physics students to regress. If I had to guess, I would say that perhaps giving the post-test for the physics students right before spring break did not help the scores, but that is speculation. In any case, the t-test results prevent me from making too many inferences about the changes in anything beyond the eighth-grade treatment group. The average physics CTSR scores did help put the growth in the eighth-grade groups in context. The intervention group gained back just 15.4% of the questions they originally missed, but this was over a third of the gap between the eighth-grade pre-test average and my physics students' pre-test scores, a class of juniors and seniors in high school. If we could maintain growth at this rate, these eighth-grade students would be in great shape by the time they were upper-classmen.

Again, given the time and scope of the intervention, it might not be especially surprising that students' responses to questions such as "I am likely to pursue a STEM career later in life" did not change significantly over the course of the study. That said, I was surprised to see that other survey questions were similarly stable, especially given that such a strong majority of students interviewed stated that they found the tools of the intervention to be helpful. I was suspicious during the interviews that some students felt compelled to tell me what they thought I wanted to hear, so perhaps some of these responses should be taken with a grain of salt. It is also possible that while the students do now have a better understanding of what is required to competently carry out an investigation or base an argument in evidence, more time is needed before that understanding translates to correctly analyzing the situations presented on the CTSR. Students in both classes showed substantial growth in CTSR questions nine and ten, which deal with a pendulum. Both the physics and eighth-grade students did an investigation on a pendulum as a part of this study, and it seems likely that this experience translated to better scores on these items. And while we were frequently emphasizing investigation and argumentation skills that were related to other questions, the connection was less explicit, and likely would take longer to develop.

The strongest signal that I received from my data came from the student interviews I conducted at the end of the study. Student after student informed me that both the Investigation Rubric and the Claim Evidence Reasoning scaffold provided them with concrete direction that helped them understand what I was looking for in these tasks. The survey did not provide evidence that this helped students enjoy these activities any

more, but this was not necessarily the goal. Tellingly, one student told me in response to an interview question on how the CER framework had changed her arguments, “I don’t like it. My arguments have gotten better. Verbal arguments are better, I don’t know how to word it on paper … Writing enough to provide evidence makes it difficult.” In an early iteration of the CER framework on a lab dealing with a pendulum, she appears to be using the Claim section as a replacement for an opportunity to state a hypothesis (Figure 6). Numerous other students did something similar, some even followed by evidence and reasoning in which they provided a convincing case for why their stated claim was actually wrong. It appeared to me at the time that students were mimicking elements of a lab report, equating Claim and Reasoning with Hypothesis and Conclusion. After a mini-lesson and a few more iterations, the expectation seem to have been more clear. A few weeks later in a lab on balance and torque the student quoted earlier now made a claim that was clearly supported by her evidence and used the reasoning section to provide some quantitative backing for the claim. While ideally, I was looking for some quantitative relationship to be stated in the claim section, this student was now using each section as I had instructed, and taken together I feel she has constructed an argument that was based on evidence that she gathered for that purpose. Even in the absence of conclusive data that showed that these instruments helped improve student performance, a tool that helps better communicate what we are looking for in lab or in structuring an argument is quite useful. I will certainly continue to use the Claim Evidence Reasoning structure in future years, and I will continue to refine the rubric for assessing labs, if only

to provide another way of explaining to students what my expectations are for these investigations.

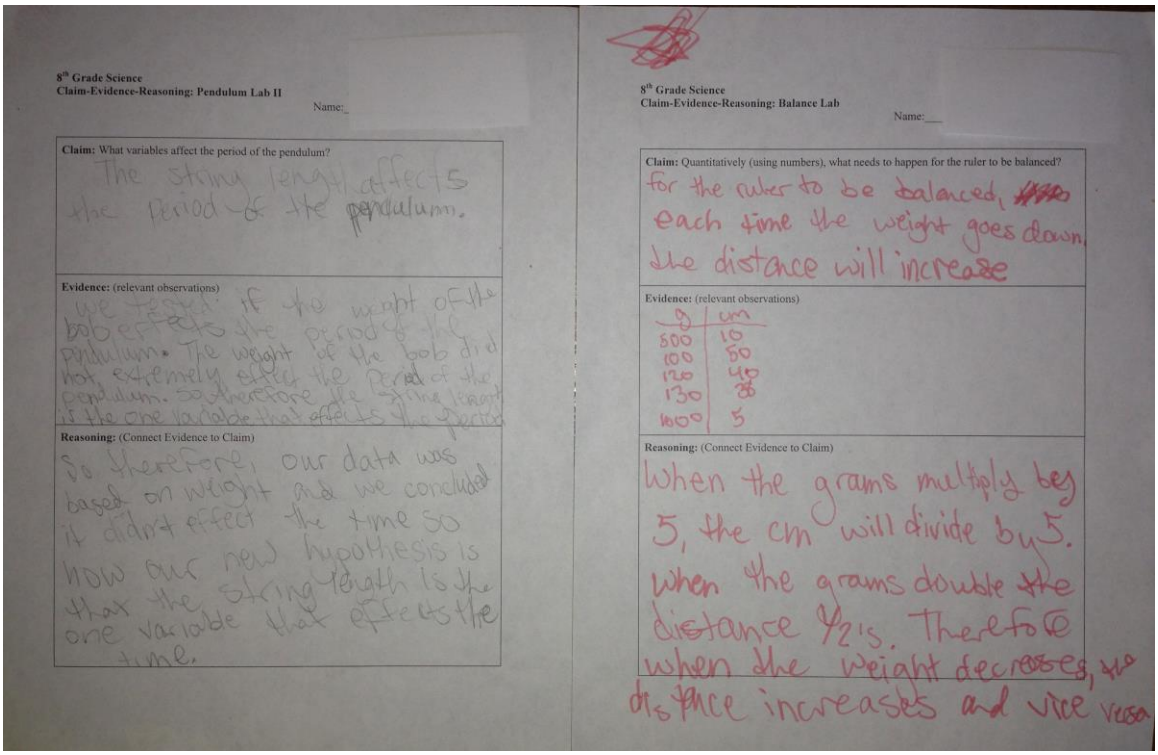


Figure 6. Student samples of Claim Evidence Reasoning arguments from early and late in the treatment.

VALUE

As explained in the introduction, to teach in a small school is to wear a lot of different hats. Since moving to Telluride seven years ago I have taught nine different classes, three or four at a time, developing curriculum for most of them. I have learned much from these experiences. Teaching so many different classes has broadened my own knowledge base and added many different activities to my toolkit as a science teacher. However, in preparing for and carrying out this project, I have realized that some of the central aspects of my science teaching have been neglected, namely developing and improving my assessment tools, and aligning my practice to the NGSS. This project gave

me an opportunity to revisit how I assess some of the aspects of my science teaching that I hold to be very important yet have been put on hold in favor of what often felt like the more immediate demand of planning lessons in new classes.

Before this project, I had heard that the elements within the Next Generation Science Standards were highly interconnected. I had even repeated this idea myself but didn't really appreciate the depth until I got into this project. There were a number of instances where I was helping a lab group improve their performance with the designing and carrying out investigations practice, but I found myself talking about other practices instead. For example, I was helping a group improve their score on the hypothesis component of my Investigations Rubric and found myself emphasizing the importance of developing and using models. Similarly, I would be helping a group wrap up and make sense of their investigation and realized that we had transitioned from the investigation to analyzing and interpreting data, as well as using mathematics and computational thinking. I had originally chosen to focus this project on the investigations and argumentation practices, thinking of them as important but discrete. Early on I realized that these two practices complimented each other well. Resolving our investigations involved developing an argument or grounding our arguments in evidence required another investigation. I had been an advocate of the NGSS before this project, but doing this research really helped me understand how the NGSS practices aren't just a collection of solid priorities, but a package of interrelated skills that shouldn't be thought of as individual entities.

I also found that when I was planning the lessons with assessment tools in mind, my approach to framing the lab changed. With the eighth-grade students, I had a bit of difficulty early on finding labs that students could take a meaningful role in designing themselves, since the content at the time was covering large systems like weather and climate. When I moved on to general investigations outside of the curriculum, I found that when my main goal was to have students construct an argument at the end of a lab, I framed the whole activity differently than I previously would have. Before the study, I certainly would have described my approach to science instruction as primarily inquiry based, and I still feel that it was. However, this experience allowed me to try an approach where the argument that I wanted students to eventually construct took on a much larger role than it had before. I spent more time clarifying the question that would guide this inquiry and setting the lab up so that I could release students to whatever they deemed to be a worthwhile investigation during the work time. When students asked questions like, "how much data do we need to collect?" it felt a lot more authentic to ask them about what they felt like they would need to confidently back up their claim, instead of just telling them that a scatterplot should have at least eight different data points as I might have said previously. While I spent more time working with students developing the reasoning portion of their CER arguments, it was the evidence portion that helped me think about the purpose of our laboratory investigation with more focus than I had before. It is entirely possible that while I was measuring what I could about students' performance, the largest changes in this study were happening to my own practice, in how I was conceiving the lab from the start.

Assessing these practices through rubrics or other means provided an effective way of communicating these priorities to students, not least because it ties the priority to their grade. The students at my school frequently look for ways to improve their grade, and teachers can grow weary when students seem more interested in points than learning. If I can close the gap between what gets points and what is important in their science education, I can at least better leverage the students' desire for better grades into more learning opportunities. The students have already told me that the tools developed in this study helped them understand what was expected. One challenge I have going forward is to refine how these assessment tools can occupy a larger part of their grade, and how more can be developed so that the gradebook and the NGSS start to have more in common.

Given the small sample size, low growth, and weak correlations that my own study produced, it would be difficult for me to make the case that these assessment tools must be taken up by other science teachers. I might instead advocate for others to try to use the action research process as a way of reflecting on their own practice and subjecting new ideas to more rigorous testing. There were certainly times during this process where I despaired at the number of uncontrollable variables related to students and performance within a school. It is tempting then to fall back on instinct and qualitative judgement and given the number of decisions that a teacher must make in any period of time, this approach will always be necessary. However, if we resort exclusively to these instincts, as I feel I have over the past few years, we can become lost. As Carl Sagan said, "Science is far from the perfect instrument of knowledge. It's just the best we have." As science

teachers, we shouldn't just be teaching about the scientific process, we should take steps to make sure that the scientific process in informing the decisions we make, even if what we are left with at the end of our inquiry falls short of a definite conclusion.

REFERENCES CITED

Anderson, K., (2016, July 25). Creating Rubrics for Performance Tasks Aligned to NGSS – Part 1 [Web log post]. Retrieved from http://wisdpiscience.blogspot.com/2016/07/creating-rubrics-for-performances-tasks.html

Best High Schools in Colorado. (2018). Retreived May 10, 2018, from https://www.usnews.com/education/best-high-schools/colorado

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working Inside the Black Box: Assessment for Learning in the Classroom. *The Phi Delta Kappan, 86*(1), 8-21. Retrieved from http://www.jstor.org.proxybz.lib.montana.edu/stable/20441694

Chappuis, S., & Stiggins, R. J. (2002). Classroom Assessment for Learning. *Educational Leadership, 60*(1), 40.

Chen, H., She, J., Chou, C., Tsai, Y., & Chiu, M. (2013). Development and Application of a Scoring Rubric for Evaluating Students' Experimental Skills in Organic Chemistry: An Instructional Guide for Teaching Assistants. *Journal of Chemical Education, 90*(10), 1296.

Conley, D.T., & Darling-Hammond, L. (2013). Creating systems of assessment for deeper learning. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Evans, D. L., PhD. (2014). Assessing for NGSS. *NSTA Reports, 25*(7), 3. Retrieved from https://search-proquest-com.proxybz.lib.montana.edu:3443/docview/1504546117?accountid=28148

Eyster, Linda S. (1997). A comprehensive rubric. (science students; laboratory investigation assessment). *The Science Teacher, 64*(9), 18.

Hafner, John, & Hafner, Patti. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education, 25*(12), 1509-1528.

Keller, T. E., & Pearson, G. (2012). A Framework for K-12 Science Education: Increasing Opportunities for Student Learning. *Technology & Engineering Teacher, 71*(5), 12-18.

Krajcik, J. S., & McNeill, K. L. (2011). CER Rubric. Retrieved from https://s3.amazonaws.com/NSTA1/1206019/CER_Rubric.pdf?AWSAccessKeyId=AKIAIMRSQAV7P6X4QIKQ&Expires=1531062054&Signature=GpWXH0EK2ZfwZuBBKya9rSsk66s=

NGSS Lead States. 2013. *Next Generation Science Standards: For States, By States.* Washington, DC: The National Academies Press

Padilla, M., & Cooper, M. (2012). GUEST EDITORIAL: From the Framework to the Next Generation Science Standards: What Will It Mean for STEM Faculty? *Journal of College Science Teaching, 41*(3), 6-7. Retrieved from http://www.jstor.org.proxybz.lib.montana.edu/stable/43748314

THS Senior Profile. (n.d.) Retrieved December 9, 2017, from http://www.telluride school.org/UserFiles/Servers/Server_4072017/File/School%20Profile%202017_2018%20paper%20copy.pdf

APPENDICES

APPENDIX A

MONTANA STATE UNIVERSITY IRB EXEMPTION

MONTANA
STATE UNIVERSITY

INSTITUTIONAL REVIEW BOARD
For the Protection of Human Subjects
FWA 00000165

960 Technology Blvd. Room 127
c/o Microbiology & Immunology
Montana State University
Bozeman, MT 59718
Telephone: 406-994-6783
FAX: 406-994-4303
E-mail: cherylj@montana.edu

Chair: Mark Quinn
406-994-4707
mquinn@montana.edu
Administrator:
Cheryl Johnson
406-994-4706
cherylj@montana.edu

MEMORANDUM

TO: Derek Engebretsen and John Graves

FROM: Mark Quinn Mark Quinn CJ
Chair, Institutional Review Board for the Protection of Human Subjects

DATE: December 4, 2017

RE: *"Improved Assessment of the NGSS Practice Standards: Designing and Carrying Out Investigations and Engaging in Argument from Evidence"* [DE120417-EX]

The above research, described in your submission of December 4, 2017, is exempt from the requirement of review by the Institutional Review Board in accordance with the Code of Federal regulations, Part 46, section 101. The specific paragraph which applies to your research is:

_X_ (b) (1) Research conducted in established or commonly accepted educational settings, involving normal educational practices such as (i) research on regular and special education instructional strategies, or (ii) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods.

_X_ (b) (2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless: (i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability, or be damaging to the subjects' financial standing, employability, or reputation.

___ (b) (3) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior that is not exempt under paragraph (b)(2) of this section, if: (i) the human subjects are elected or appointed public officials or candidates for public office; or (ii) federal statute(s) without exception that the confidentiality of the personally identifiable information will be maintained throughout the research and thereafter.

____ (b) (4) Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available, or if the information is recorded by the investigator in such a manner that the subjects cannot be identified, directly or through identifiers linked to the subjects.

____ (b) (5) Research and demonstration projects, which are conducted by or subject to the approval of department or agency heads, and which are designed to study, evaluate, or otherwise examine: (i) public benefit or service programs; (ii) procedures for obtaining benefits or services under those programs; (iii) possible changes in or alternatives to those programs or procedures; or (iv) possible changes in methods or levels of payment for benefits or services under those programs.

____ (b) (6) Taste and food quality evaluation and consumer acceptance studies, (i) if wholesome foods without additives are consumed, or (ii) if a food is consumed that contains a food ingredient at or below the level and for a use found to be safe, or agricultural chemical or environmental contaminant at or below the level found to be safe, by the FDA, or approved by the EPA, or the Food Safety and Inspection Service of the USDA.

Although review by the Institutional Review Board is not required for the above research, the Committee will be glad to review it. If you wish a review and committee approval, please submit 3 copies of the usual application form and it will be processed by expedited review.

APPENDIX B

INVESTIGATIONS RUBRIC

	4	3	2	1
Guiding Question	The investigation hinges on a testable question, and the investigation is completely focused around generating evidence to help answering that question	A question has been identified, and the investigation is relevant	A question exists, but the investigation only has partial relevance in finding an answer to it	Either there is no guiding question, or the investigation is unrelated to the question
Hypothesis	A hypothesis makes a prediction about the answer to the question, and is based on a scientific model or theory	A hypothesis makes a prediction about the answer to the question, and is based on some existing background knowledge in science	A hypothesis makes a prediction about the answer to the question	No hypothesis was stated
Identification and Isolation of Variables	When appropriate, an independent and dependent variable are clearly identified, and other variables are identified and controlled to avoid interference in the investigation	When appropriate, independent and dependent variables are clearly identified, and most other variables are controlled to avoid interference in the investigation	Variables are identified, but some confusion may exist in which are independent or dependent. Also, other variables may remain uncontrolled and interfere with the validity of the data	Variables are not identified, and little effort has been made to control other variables
Data Collection	Data is collected carefully, with every effort to systematically identify and reduce sources of error. By the end of the investigation, the data is useful in providing evidence for the claim that will be made about the question	Data is collected carefully and some effort has been made to identify and reduce sources of error. The data collected is useful as evidence in answering the question.	Data is collected, but for some reason (carelessness, error, disorganization) has limited usefulness in answering the question.	The data collected is not useful in answering the question.

APPENDIX C

ARGUMENTATION RUBRIC

Base Explanation Rubric (McNeill & Krajcik 2011)

Component	Level		
	0	1	2
Claim A statement that responds to the question asked or the problem posed.	Does not make a claim, or makes an inaccurate claim.	Makes an accurate but incomplete claim.	Makes an accurate and complete claim.
Evidence Scientific data used to support the claim.	Does not provide evidence, or only provides inappropriate evidence (Evidence that does not support claim).	Provides appropriate, but insufficient evidence to support claim. May include some inappropriate evidence.	Provides appropriate and sufficient evidence to support claim.
Reasoning Using *scientific principles* to show *why data count as evidence* to support the claim.	Does not provide reasoning, or only provides reasoning that does not link evidence to the claim.	Provides reasoning that links the claim and evidence. Repeats the evidence and/or includes some scientific principles, but not sufficient.	Provides reasoning that links evidence to claim. Includes appropriate and sufficient scientific principles.

APPENDIX D

LAWSON'S CLASSROOM TEST OF SCIENTIFIC REASONING

CLASSROOM TEST OF

SCIENTIFIC REASONING

Multiple Choice Version

Directions to Students:

This is a test of your ability to apply aspects of scientific and mathematical reasoning to analyze a situation to make a prediction or solve a problem. Make a dark mark on the answer sheet for the best answer for each item. If you do not fully understand what is being asked in an item, please ask the test administrator for clarification.

DO NOT OPEN THIS BOOKLET UNTIL YOU ARE TOLD TO DO SO

Revised Edition: August 2000 by Anton E. Lawson, Arizona State University. Based on: Lawson, A.E. 1978. Development and validation of the classroom test of formal reasoning. *Journal of Research in Science Teaching*, 15(1): 11-24.

1. Suppose you are given two clay balls of equal size and shape. The two clay balls also weigh the same. One ball is flattened into a pancake-shaped piece. *Which of these statements is correct?*

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

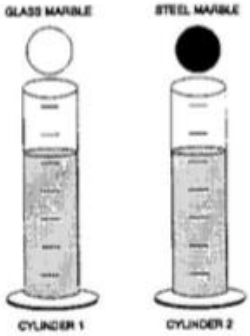
2. *because*

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

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

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

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

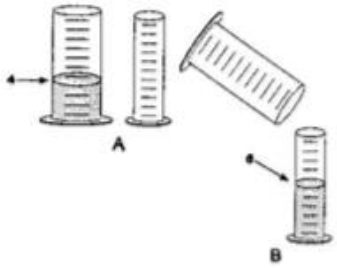


   a. to the same level as it did in Cylinder 1
   b. to a higher level than it did in Cylinder 1
   c. to a lower level than it did in Cylinder 1

4. *because*

   a. the steel marble will sink faster.
   b. the marbles are made of different materials.
   c. the steel marble is heavier than the glass marble.
   d. the glass marble creates less pressure.
   e. the marbles are the same size.

5. To the right are drawings of a wide and a narrow cylinder. The cylinders have equally spaced marks on them. Water is poured into the wide cylinder up to the 4th mark (see A). This water rises to the 6th mark when poured into the narrow cylinder (see B).



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

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

6. *because*

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

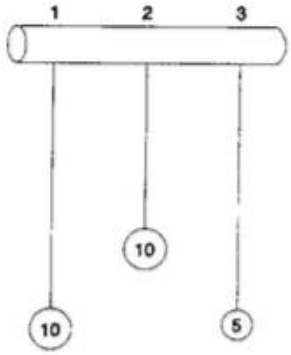
7. Water is now poured into the narrow cylinder (described in Item 5 above) up to the 11th mark. *How high would this water rise if it were poured into the empty wide cylinder?*

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

8. *because*

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

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



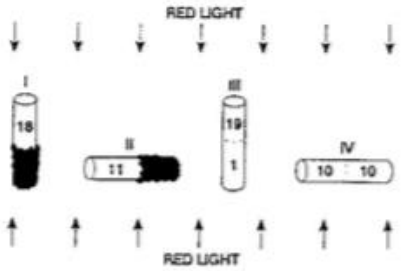
Suppose you want to find out whether the length of the string has an effect on the time it takes to swing back and forth. *Which strings would you use to find out?*

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

10. *because*

a. you must use the longest strings.
b. you must compare strings with both light and heavy weights.
c. only the lengths differ.
d. to make all possible comparisons.
e. the weights differ.

11. Twenty fruit flies are placed in each of four glass tubes. The tubes are sealed. Tubes I and II are partially covered with black paper; Tubes III and IV are not covered. The tubes are placed as shown. Then they are exposed to red light for five minutes. The number of flies in the uncovered part of each tube is shown in the drawing.



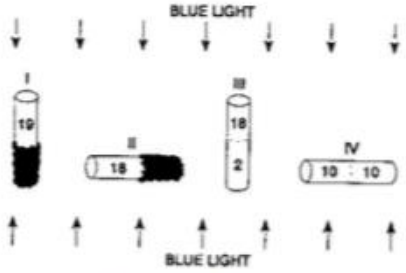
This experiment shows that flies respond to (respond means move to or away from):

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

12. *because*

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

13. In a second experiment, a different kind of fly and blue light was used. The results are shown in the drawing.



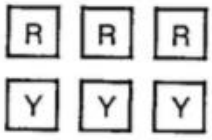
These data show that these flies respond to (respond means move to or away from):

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

14. *because*

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

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



a. 1 chance out of 6
b. 1 chance out of 3
c. 1 chance out of 2
d. 1 chance out of 1
e. cannot be determined

16. *because*

a. 3 out of 6 pieces are red.
b. there is no way to tell which piece will be picked.
c. only 1 piece of the 6 in the bag is picked.
d. all 6 pieces are identical in size and shape.
e. only 1 red piece can be picked out of the 3 red pieces.

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

What are the chances that the piece is a red round or blue round piece?

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

18. *because*

a. 1 of the 2 shapes is round.
b. 15 of the 21 pieces are red or blue.
c. there is no way to tell which piece will be picked.
d. only 1 of the 21 pieces is picked out of the bag.
e. 1 of every 3 pieces is a red or blue round piece.

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

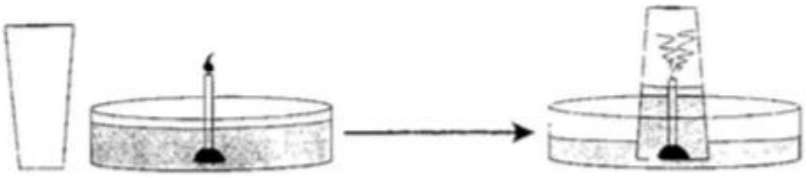
Do you think there is a link between the size of the mice and the color of their tails?

a. appears to be a link
b. appears not to be a link
c. cannot make a reasonable guess

20. *because*

a. there are some of each kind of mouse.
b. there may be a genetic link between mouse size and tail color.
c. there were not enough mice captured.
d. most of the fat mice have black tails while most of the thin mice have white tails.
e. as the mice grew fatter, their tails became darker.

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



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

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

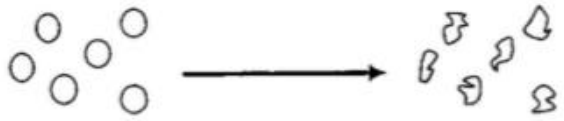
Suppose you have the materials mentioned above plus some matches and some dry ice (dry ice is frozen carbon dioxide). *Using some or all of the materials, how could you test this possible explanation?*

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

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

a. The water rises the same as it did before.
b. The water rises less than it did before.
c. The balloon expands out.
d. The balloon is sucked in.

23. A student put a drop of blood on a microscope slide and then looked at the blood under a microscope. As you can see in the diagram below, the magnified red blood cells look like little round balls. After adding a few drops of salt water to the drop of blood, the student noticed that the cells appeared to become smaller.



Magnified Red Blood Cells After Adding Salt Water

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

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

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

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

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

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

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

APPENDIX E

STUDENT ATTITUDES SURVEY

Student Attitudes Survey and Science Reasoning Test

Participation in this research is voluntary and participation or non-participation will not affect a student's grades or class standing in any way.

Directions: Rank the degree to which you agree or disagree with the following statements on a 4-point scale:

4- Strongly Agree
3- Agree
2- Disagree
1- Strongly Disagree

I am a curious person	4 SA	3 A	2 D	1 SD
I am interested in science	4 SA	3 A	2 D	1 SD
I am likely to pursue a STEM related career later in life	4 SA	3 A	2 D	1 SD
I enjoy opportunities to carry out investigations in science	4 SA	3 A	2 D	1 SD
Designing my own investigations helps me answer questions in science	4 SA	3 A	2 D	1 SD
Testing ideas is an important part of the scientific process	4 SA	3 A	2 D	1 SD
I find myself applying what I know about designing and carrying out investigations outside of science class	4 SA	3 A	2 D	1 SD
My arguments in science are always grounded in evidence	4 SA	3 A	2 D	1 SD
I enjoy making arguments in science class	4 SA	3 A	2 D	1 SD
I find myself applying what I know about engaging in argument from evidence outside of science class	4 SA	3 A	2 D	1 SD
When I'm evaluating someone else's argument, looking at their evidence is an important part of my analysis	4 SA	3 A	2 D	1 SD

Is there anything else you would like to tell me?

APPENDIX F

INTERVIEW QUESTIONS

1. Do you feel your approach to labs has changed at all since we started using the rubrics? In what way?
2. Do you feel your ability to construct arguments has changed at all since we started using the Claim-Evidence-Reasoning worksheets? In what way?
3. Is there anything else you want to tell me?