Journal of Research in STEM Education

Vol 5, No 2, December 2019, 170-188

RESEARCH REPORT

Evaluation Approach: Practice-Focused Middle School Science Modules

Sunni Newton[a1], Meltem Alemdar[a], Jayma Koval[a], Jessica Gale[a], Sabrina Grossman[a], Stefanie Wind[b], Mike Ryan[a], Marion Usselman[a]

[a]*Georgia Institute of Technology, USA;* [b]*The University of Alabama, USA*

Abstract: *Advanced Manufacturing and Prototyping Integrated to Unlock Potential (AMP-IT-UP) is a National Science Foundation (NSF) funded K-12 Math & Science Partnership (MSP) project with a goal of promoting math, science, and engineering learning through STEM integration-focused curricula. As part of this project, curriculum writers developed one-week modules providing instruction on a set of STEM practices within the context of the appropriate grade-level content. These STEM practices are grouped into strands labeled Experimental Design, Data Visualization, and Data-Driven Decision Making; the emphasis of each of these practice strands is, respectively, the collection of data, the representation of data, and the use of data to support complex decision-making. Nine one-week modules were created in the science domain, one focused on each practice strand at grade levels 6, 7, and 8. A parallel set of nine modules in the math domain were also created, resulting in a total of 18 modules. In this paper, we will focus on our evaluation of the effectiveness of these modules as they were implemented across four middle schools. We will present our methodology for evaluating this complex instructional effort. Data sources include online teacher enactment surveys, teachers' on-line posts, as well as classroom observations. Findings were compiled across these multiple data sources to provide detailed insights into curriculum functioning and teacher experiences. We will also provide some results from pre-post assessments of student learning, which were written for a subset of the science modules. Overall, the results provide detailed and valuable insight into curriculum functionality as well as evidence of significant increases in student learning in some modules. Findings from these data sources were used by curriculum developers to inform later iterations of the modules.*

Keywords: *K-12, STEM, evaluation*

The purpose of this paper is to describe a multifaceted evaluation of a set of one-week science and mathematics modules designed for use in middle schools. These modules provide students with instruction on, and experience working with, data-related STEM practices grouped into three clusters entitled Data Visualization, Experimental Design, and Data-Driven Decision Making. The practices in these clusters were drawn from the Next Generation Science Standards (NGSS) (NGSS Lead States, 2013) and the Standards of Mathematical Practice (SMP) (National Governors Association Center for Best Practices, 2010). The modules are situated within the context of grade-level specific science and math disciplinary content. This evaluation was carried out by a team of education researchers and utilized multiple data sources. The core of the achievement of curriculum objectives is driven in part by the evaluation process used during the development of the curriculum. This study addresses how different sources of data can be used to evaluate whether or not the curriculum reached its objectives. The goal is to provide an evaluation process for curriculum development, which is a continuous iterative process carried out until the objectives are reached.

[1] Corresponding author: Sunni Newton, Georgia Institute of Technology, Center for Education Integrating Science, Mathematics, and Computing (CEISMC), 817 W. Peachtree Street, NW Suite 300 Atlanta, GA 30308, Email: sunni.newton@ceismc.gatech.edu

Newton, S., Alemdar, M., Koval, J., Gale, J., Grossman, S., Wind, S., Ryan, M., & Usselman, M. (2019). Evaluation Approach: Practice-Focused Middle School Science Module. *Journal of Research in STEM Education*, *5*(2), 170-188.

ISSN:2149-8504 (online)

The evaluation of the math and science modules was a complex research effort due to a number of factors. First, there are three science modules for each middle school grade level, for a total of nine distinct instructional science modules across grades 6, 7 and 8. These are accompanied by a parallel set of nine math modules; for this paper, we are focusing solely on the evaluation of the science modules. In order to be comprehensive, our evaluation needed to cover not only the STEM practices common to the modules, but also the logistical details and disciplinary core ideas specific to each module and its associated activities. Therefore, it was crucial that researchers developing the evaluation have a deep understanding of the content and flow of each module. Second, there were multiple aspects of the module implementation that needed to be evaluated: how the curriculum functioned in the classroom, teachers' experiences with implementing the modules, students' experiences participating in the modules, and the nature and extent of student learning that occurred during module implementation. Third, the logistical considerations associated with carrying out observations were nontrivial, as the module implementations occurred in over twenty classrooms in four middle schools, and data collection schedules required flexibility to accommodate changes due to school events.

This paper outlines the efforts undertaken to evaluate this complex curricular innovation while attending to the factors described above. Further, we provide an alternative and novel way of evaluating the quality of this complex curriculum by utilizing evaluation data from a variety of sources. A review of the literature yielded no comparable studies in which a variety of data sources were combined in this manner to inform iterative curriculum development. As such, we feel this description of our evaluation process represents a meaningful contribution to the literature. The goal of the paper is to detail our evaluation efforts, including its multiple components and how each was selected, designed, and executed. We will also discuss the manner in which the various evaluation data sources were compiled and used to inform later iterations of the curriculum design.

Project Overview

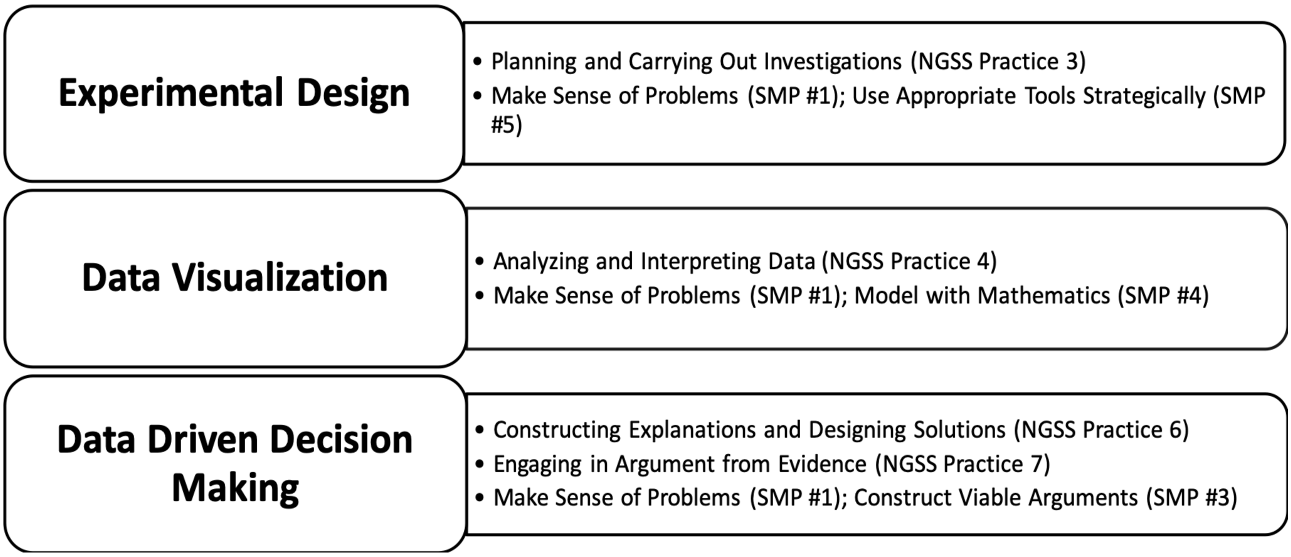
This paper describes an evaluation effort embedded within a larger project, titled "Advanced Manufacturing and Prototyping Integrated to Unlock Potential" (AMP-IT-UP). This project was a 5-year National Science Foundation (NSF) Math and Science Partnership (MSP) between a higher education institute and an urban-fringe school district within the same state. The goals of the partnership were to promote STEM integration in middle school engineering, math, and science classes, increase STEM relevance by emphasizing advanced manufacturing and data-driven problem solving, and ultimately to increase student engagement and academic achievement in math, science and engineering. A major component of AMP-IT-UP was the development, implementation, and evaluation of math and science modules that were implemented in core math and science classes.

In this paper, we describe a methods-focused approach and outline the various components of the multi-faceted evaluation effort undertaken to investigate the effectiveness of the modules. First, we provide an overview of the modules, briefly describing their development and highlighting their role within the overarching project. Next, we discuss our goals in evaluating this complex curricular component, the instruments and methods involved in the evaluation, and how the choice of evaluation methods and instruments helped us to achieve these goals. We further discuss how the evaluation data were utilized for curriculum development. We conclude with a brief discussion of a small subset of sample results, presented not in service of drawing summative conclusions or providing a comprehensive overview of the analyses and results, but rather to offer some insight as to how the evaluation results were synthesized and used within the iterative curriculum development process employed in this project.

Module Description

Each module focuses on a particular cluster of STEM practices and on grade-level appropriate disciplinary content, and presents student teams with an inquiry-driven problem or challenge. The three STEM practice clusters, labeled Experimental Design, Data Visualization, and Data-Driven Decision Making, emphasize, respectively, the collection of data, the representation of data, and the use of data to inform decision making. Standards included in the modules, for both STEM practices and disciplinary core ideas, were drawn from the Next Generation Science Standards (NGSS Lead States, 2013) and the Standards of Mathematical Practice (National Governors Association Center for Best Practices, 2010). The STEM practices in each cluster are listed in Table 1. The curriculum designers intended the STEM practice clusters to unify overlapping skills across these national science and math standards, allowing for an investigation of integrated STEM learning.

Table 1.
Module practices and associated standards.



Experimental Design	• Planning and Carrying Out Investigations (NGSS Practice 3) • Make Sense of Problems (SMP #1); Use Appropriate Tools Strategically (SMP #5)
Data Visualization	• Analyzing and Interpreting Data (NGSS Practice 4) • Make Sense of Problems (SMP #1); Model with Mathematics (SMP #4)
Data Driven Decision Making	• Constructing Explanations and Designing Solutions (NGSS Practice 6) • Engaging in Argument from Evidence (NGSS Practice 7) • Make Sense of Problems (SMP #1); Construct Viable Arguments (SMP #3)

The modules also each support development of student understanding within disciplinary core ideas taught in middle school science (i.e., earth science, life science and physical science) and math (i.e., statistics, geometry, and algebra). It should be noted that these are one-week modules, and as such, are not in of themselves sufficient to promote proficiency with or mastery of these disciplinary core ideas. There are 18 modules across the three grade levels, and each requires approximately one week of instruction. A student receiving instruction on the full complement of modules would complete six modules per grade level, three in math and three in science. The full set of modules is presented in Table 2.

Table 2.
AMP-IT-UP Science and Math Modules

	Science Modules		
	Experimental Design	Data Visualization	Decision Making
6th	Lava	Earthquake	Winter Weather
7th	Oil Spill	Deep Sea Ecosystems	Coral Reef
8th	Marine Snow	Helmet	Skate Park
	Math Modules		
	Experimental Design	Data Visualization	Decision Making
6th	Packaging	Whale	Automated Packaging
7th	Board Game Piece	Crab Aquarium	Manufacturing Quality Control
8th	Clean Energy	Hot Shots	Power Finance

The three, one-week science modules developed for each grade level were in part informed by the approach used in Kolodner et al.,'s (2003) Launcher Unit, in which the key aim is providing instruction on critical skills rather than domain-specific content. The research of Kolodner et al., demonstrates the impact of launcher units: as students engage in project-based contextualized learning through skill-focused units, they have repeated opportunities to practice with and reflect upon the skills they are building, while simultaneously seeing the connection of these skills to science concepts. These units model and establish terminology, rituals, and norms that are later repeated in more concepts-focused units, and research suggests that launcher units promote transfer of skills and development of science conceptual knowledge (Kolodner et al., 1998; Kolodner et al., 2003). Because the one-week duration of modules limited the extent to which they would be able to fully present and cover substantial science and math content, we saw an opportunity to instead use the modules in a launcher unit-like fashion to develop understanding of critical and unifying skills across the three STEM practices.

Science modules were designed with the intention that teachers would implement them either early in the school year, to introduce students to practices that would be iterated upon and developed throughout the school year, or as an introductory experience before in depth instruction on relevant disciplinary content. The math modules were designed to be implemented during the teaching of the specific domains of content that they covered. Actual timing of module implementation varied across teachers based on their scheduling preferences and other logistical constraints.

Module implementation. The modules were rolled out gradually over the course of the project, with all 18 modules being implemented during Years 4 and 5. Teachers received professional development on the modules primarily at a summer institute held each year at a location within the school district. For this paper, we will focus on the nine science modules. This is largely due to their being in a more advanced stage of development during our focal year of curriculum implementation as compared to the math modules. This is because the science modules were fully developed and piloted earlier in the project than most of the math modules, and had been iterated upon more times than were the math modules prior to the focal school year.

The 2016-2017 school year represents the first year in the project that the full set of modules was available for implementation. During the 2016-2017 school year, 20 regular middle school science teachers and five special education science teachers implemented one or more of the science modules. During the 2017-2018 school year, 21 regular middle school science teachers and five special education science teachers implemented one or more of the science modules. In the early years of the project, participation in AMP-IT-UP was voluntary. However, by Year 5 (2017-2018), all middle school math and science teachers in the district were instructed to use the materials in their classes. Because of the large turnover of teachers in the district every year, there were some newly hired teachers who had not received instruction on the materials. We attempted to address this issue by providing professional development on the modules after the school year had started for late hires. Additionally, teachers who had prior experience with the modules often provided guidance to new teachers who may have missed the summer professional development. Despite these efforts, each year there were some teachers who did not implement the modules due to having missed the summer professional development.

Module instructional materials. The project provided teachers with all materials required to teach a module: a student content booklet (called the Student Edition), a set of student worksheets (separate from the booklet), two teacher curriculum guide booklets (called the Teacher Editions; one of these is a facilitation guide for module implementation, and the other is a preparation guide for pre-facilitation prep), videos, links to required computer simulations, supplemental materials, and supplies and/or manipulatives needed for the module implementation. The Student Edition includes the student text, instructions for activities, and guiding questions for discussion. Student worksheets are separate from the booklets and include graphing exercises, space for students to write procedures and record the data they collect, graphic organizers, and space for communicating evidence-based decisions or recommendations.

Two Teacher Editions were developed to aid teachers in implementing the modules. The first is an Annotated Teacher Edition, which is a copy of the Student Edition with notes for the teacher added throughout. Teachers use this Teacher Edition during instruction to follow along with students as they progress through the Student Edition. The Annotated Teacher Edition also provides "in the moment" tips for instruction and questions to ask students.

The second Teacher Edition is a preparatory guide that was developed halfway through the project based on the needs of our teachers; it was designed to align with the format of their lesson plans. The preparatory guide Teacher Edition contains the content teachers would need to aid them in lesson planning (e.g., curriculum standards) and to help them prepare for implementation. The preparatory guide also includes detailed instructions for setting up manipulatives and computer simulations for modules that utilize these activities, as well as expectations for student and teacher activities.

Each preparatory guide includes an overview of how the module maps onto the BSCS Science Learning 5E Instructional Model (Bybee, 2015), which was a pre-existing district-level requirement for teachers' lesson plans. Constructivist theory as it relates to learning is at the root of the 5E model, a sequence which consists of the following five phases: Engage, Explore, Explain, Elaborate, and Evaluate. Each phase has specific functions through which teachers and learners progress in support of achieving educational objectives (Bybee, 2006). Mapping the module curriculum onto the 5E Instructional Model provided teachers with an aid for implementation and an overview of expectations for student and teacher activities. Furthermore, the use of this model presented the module curriculum within the context of a framework with which these teachers were already familiar, and aided in their lesson planning by mapping the module curriculum onto the required organizational structure for their lesson plans.

Module examples. This section presents a brief overview of three sample modules, including one from each grade level and one from each of the STEM practice clusters. Please see Appendix A for information on accessing all modules and associated materials via our project website.

In the 6th grade Experimental Design module, titled "Molten Madness: Lava Challenge," students are challenged to help determine the speed of lava flow on land in order to develop evacuation plans in the event of a volcano eruption. This instructional approach of framing the module within the context of an overarching challenge students are tasked with solving is repeated across all modules. This module emphasizes writing procedures, which fits within the focal STEM practice cluster of Experimental Design. Students are asked to write their own procedure to measure the "lava" flow in their models, which use a plastic plate and dish soap to model the flow of lava down a mountain. Please see Appendix B, part 1, for a screen shot of the page of the student edition in which students are given instructions for writing their initial procedure.

Generally, students write poor procedures at first; when students follow these low-quality procedures, in which important details are frequently ignored or not properly controlled, the resulting data varies widely across students. This is made clear to students when the data generated from their initial procedures are displayed on a histogram, providing a visual of the widely varying data. Over the course of the module, they re-write their procedures, improving them as they learn to control variables (e.g., the amount of soap, how they record elapsed time). As they redo the experiment, the results become more consistent across students. This module also provides instruction on the practice of modeling. Many students and teachers have the misconception that in order to address this challenge, they would need to create a model of a volcano, complete with lava flowing from its center. This module demonstrates that a simple model of the flow of liquid across a surface is sufficient.

The 7th grade Data Visualization module, titled "Deep Sea Ecosystems", is set within the context of research on coral reef health after the Deep Horizons oil spill in the Gulf of Mexico. In this module, students analyze a set of photos of deep-sea coral taken over time by researchers studying the after-effects of the oil spill. The students create a rubric to evaluate the health of coral, and then use a variety of visual representations of their data to present and emphasize different aspects of their coral sample's health. The multiple visual representation options presented in this module, including the use of temporal and pictorial data, align with its focus on data visualization. Please see Appendix B, part 2, for a screen shot of the page of the student edition in which students are given the challenge description for this module.

The 8th grade Data-Driven Decision-Making module, titled "Skate Park", tasks students with analyzing crash helmet strength through the use of a simulation, and then selecting the best helmet for a skateboard rider. During the activity students evaluate the cost and level of protection that different helmets provide, as well as the budget and risk level of the different riders. This focus on compiling and synthesizing various sources of information to make and defend a decision that is intentionally messy aligns with this module's focus on data-driven decision making. Please see Appendix B, part 3, for a screen shot of the page of the student edition in which students are asked to share and discuss with the class the helmet data they have collected.

Methodology

The curriculum development and research approaches adopted for this project were informed by the design-based implementation research (DBIR) framework (Fishman, Penuel, Allen, Cheng, & Sabelli, 2013), which holds as a central tenant "a commitment to iterative, collaborative design" (Fishman et al., 2013, p. 142). The interplay between the data collected on curriculum implementation, the reporting of this data to the curriculum team, and the use of this data to inform subsequent iterations of the curriculum exemplifies DBIR through its emphasis on both collaboration and iteration, with the results of a collaboratively designed evaluation effort directly informing multiple iterations of the curriculum.

Evaluation of a complex intervention such as the large, multi-faceted, multi-year curriculum development and implementation work undertaken in this project is a difficult endeavor; the evaluation was designed to satisfy several goals. The primary goal of the evaluation approach was to evaluate the overall extent to which these modules reached their objectives; more specifically, the goals were as follows:

1. to collect data on the logistics and challenges involved in teachers' module implementations;
2. to assess student learning of the content and practices covered in the modules;
3. to provide general formative feedback to the curriculum developers to inform revisions to future versions of the module as part of an iterative cycle of research on, and further development of, the module curricula.

This evaluation sought to provide information on various aspects of the module implementations, as well as insights into the perceptions of, and impacts on, both teachers and students as a result of their experiences with the modules. To this end, data were collected via numerous instruments from various stakeholder groups to allow for synthesis and comparison of findings across multiple data sources. It is our hope that by utilizing four distinct data sources which vary across method (quantitative, e.g., surveys and assessments; and qualitative, e.g., classroom observations and open-ended online posts), people (students completed assessments, teachers completed surveys and online posts, and both students and teachers were involved in classroom observations), and time (these data were collected at various time points across the focal school year), we will enrich our study and provide robust evaluation data on which to base curriculum development decisions (Flick, 2009).

Student learning of the STEM practices and disciplinary content presented in the modules was assessed via a single source, the pre and post assessments. The student assessments were heavily focused on STEM practices to allow us to directly measure student gains in knowledge and application of the STEM practices over the course of the module implementation. We captured teachers' inclusion of the STEM practices within the three clusters, as well as additional elements, in their module implementations via both their self-reports of what they did and did not do during the implementation in the online surveys and co-lab posts, as well as our own observations of whether and how the practices were taught during our classroom observations. This utilization of data across multiple sources allowed for a complete and multi-faceted picture of module implementations. Furthermore, the use of various data sources allows for increased confidence in our findings.

Student Pre-Post Assessments

To measure student learning over the course of the module, multiple choice (MC) assessments were developed and administered in a pre-post format (i.e., the same assessment was given to students both prior to and immediately after the module). Although researchers and practitioners have frequently pointed out the limitations of MC items (e.g., Klassen, 2006), well-designed MC-based assessments can provide valuable insights into student understanding without the need for time-consuming scoring procedures (Haladyna, 1999; Sadler, 1998).

All MC items were written by members of the research team and reviewed and edited by members of the curriculum team to ensure an appropriate match to the module curriculum practices and content. MC assessments were developed for the Data Visualization and Experimental Design modules only; it was determined that the focus on written content and defending one's decision using data inherent in the Data-Driven Decision Making modules did not align well with a MC assessment. As part of planned future work, student learning in Data-Driven Decision Making modules will be assessed through analysis of student work products.

Initially, MC assessment items focused solely on the clusters of STEM practices and did not address the module-specific content. However, in the early years of using these assessments, teachers expressed concerns that the assessments did not measure what they were teaching in the modules, which we interpreted as a face validity (Nevo, 1985) issue. As a result, module-specific MC items were written for some modules. Some of the general STEM practice-focused items were retained for these module-specific assessments, and items specific to the STEM practices within the context of the specific module were added. For example, this was the case for Deep Sea Ecosystems (7th grade Data Visualization module) and Helmet (8th grade Data Visualization module), which had four module-specific items and three module-specific items, respectively, added to the existing Data Visualization assessment items. Two sample MC items from the 7th grade Data Visualization module, one general practice-focused item (item 1) and one module-specific item (item 2) are presented in Appendix C.

The practice-focused items are intended to assess students' ability on a given cluster of STEM practices irrespective of the specific module content. As such, these items were written to apply generally to the STEM practice cluster to which they correspond, and can be used across module assessments for all modules on the given STEM practice. The Data Visualization STEM practice-focused item provided in Appendix C (item 1) asks students to use data on cafeteria food sales by day of the week to answer a question about meal planning for the cafeteria staff, specifically helping the staff select the food with the lowest sales to be eliminated from the cafeteria offerings. Students must be able to interpret what the data in the table represent by reading the row and column headers in the table, recognize that they will need to add the sales of a given food over the five days of the week, and identify which food has the lowest sales for the week. Per the NGSS Practice description in which this STEM practice is rooted (NGSS Practice 4: Analyzing and Interpreting Data), students in grades 6-8 should be able to "construct, analyze, and/or interpret graphical displays of data and/or large data sets to identify linear and nonlinear relationships" and "analyze and interpret data to provide evidence for phenomena" (NGSS Lead States, 2013, p. 9). This assessment item asks students to demonstrate their ability to interpret a data table to determine the rank ordering of the various foods in terms of weekly sales, and to use this data to provide evidence to advise cafeteria staff on which menu item to eliminate.

The module-specific assessment item provided in Appendix C (item 2) relates directly to the data visualization content students are instructed on and asked to apply as they work through the module challenge. In this challenge, students construct a rubric to evaluate the health of coral within a marine ecosystem affected by an oil spill. They use the rubric they construct to evaluate and compare data from different time periods (temporal data) and from different geographic areas (spatial data). This module content directly ties to the NGSS practice component stating that students should be able to "use graphical displays (e.g., maps, charts, graphs, and/or tables) of large data sets to identify temporal and spatial relationships" (NGSS Lead States, 2013, p. 9). The assessment item specifically asks students to consider what a rubric is and what situations it would and would not be useful within. This content about what a rubric is and what it can be used for is taken directly from the module, and is appropriate for this module assessment only given its close ties to the module content, as opposed to the more general STEM practices. Please see Figures 1 and 2 for a visual representation of these two sample assessment items and their underlying links to the STEM practices and corresponding NGSS practice content.

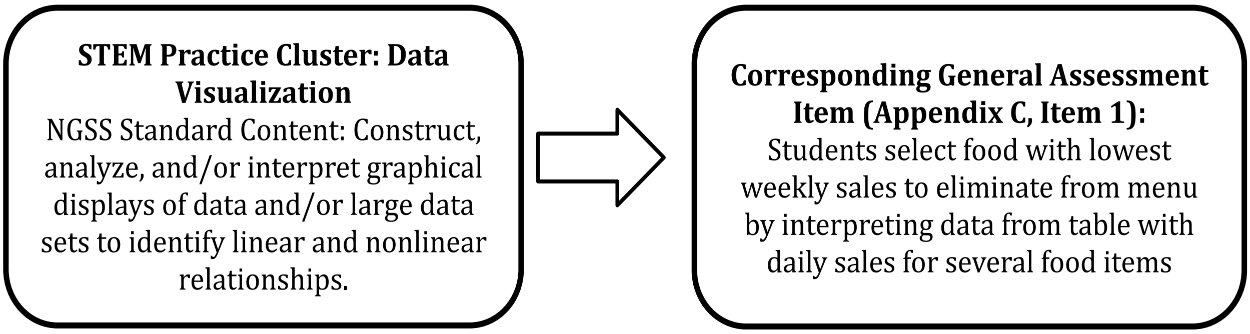

STEM Practice Cluster: Data Visualization
NGSS Standard Content: Construct, analyze, and/or interpret graphical displays of data and/or large data sets to identify linear and nonlinear relationships.
Corresponding General Assessment Item (Appendix C, Item 1):
Students select food with lowest weekly sales to eliminate from menu by interpreting data from table with daily sales for several food items

Figure 1. STEM practice and NGSS standard associated with sample assessment item 1 (Appendix C)

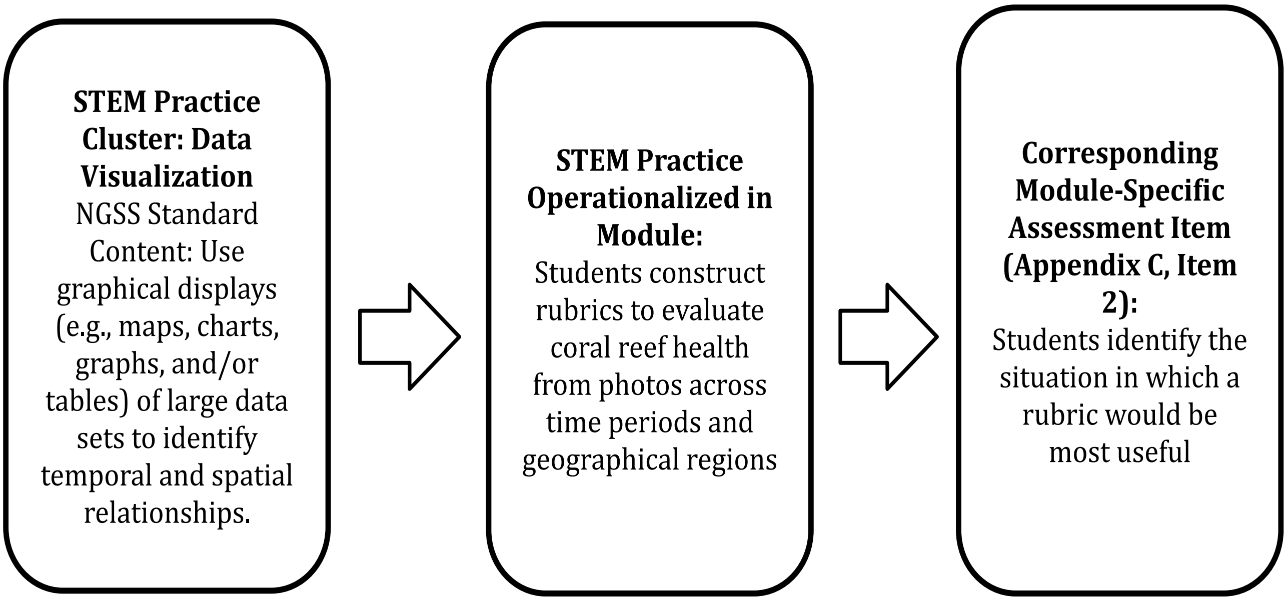

STEM Practice Cluster: Data Visualization
NGSS Standard Content: Use graphical displays (e.g., maps, charts, graphs, and/or tables) of large data sets to identify temporal and spatial relationships.
STEM Practice Operationalized in Module:
Students construct rubrics to evaluate coral reef health from photos across time periods and geographical regions
Corresponding Module-Specific Assessment Item (Appendix C, Item 2):
Students identify the situation in which a rubric would be most useful

Figure 2. STEM practice and NGSS standard associated with sample assessment item 2 (Appendix C)

These assessments take approximately one 50-minute class period or less to administer and range from 4 to 10 items in length; the items vary in cognitive complexity and required skills/computations, which is why the assessments vary widely in total number of items. For the purposes of this paper, 2016-2017 school year student data is used. The six science module assessments were each administered to between 278 and 501 students across the four middle schools. Between four and eight teachers administered each of the module assessments in their classes.

Teacher Enactment Surveys

Enactment surveys were developed for each module based on the corresponding teacher curriculum guide (Teacher Edition), which provides teacher-specific instructions as well as the general layout and flow of the module. These surveys are designed to elicit teacher accounts of specific components and activities of the module, including yes/no items asking whether teachers completed various activities, and open-ended items for teachers to provide feedback and additional details about how they completed various activities. The survey also includes questions about module logistics such as start and end date and duration of each section within the module, as well as questions about student engagement and any adaptations teachers made throughout the module. Sample enactment survey items from one section of the 6th grade Experimental Design module are shown in Appendix D. Enactment surveys were administered online via Survey Monkey and took roughly 20 minutes to complete. Links to each module-specific survey were e-mailed out by project staff when the teacher completed a module. Teachers who had not completed the surveys were prompted to do so.

Classroom Observations

Classroom observations were included in the research design in order to allow researchers to view curriculum implementation as it unfolded in the field. Teachers provided their perceptions of their module implementations via surveys; classroom observations provided another source of information on curriculum implementation and enabled researchers to study implementations in more detail and contrast what was directly observed with what teachers reported. In their text on participant observation, Dewalt & Dewalt (2002) describe observation as follows: “the researcher explicitly and self-consciously attending to the events and people in the context they are studying” (p. 68). Observations are considered by some researchers to be the most rigorous means of studying implementation (Ruiz-Primo, 2006).

Classroom observations, conducted by members of the research staff, were guided by protocols created directly from the enactment surveys (discussed above). Protocols were utilized in order to guide the observer to focus on recording specific details that would be most useful for informing curriculum iteration, given that “all observation is partial” (Agar, 1996; cited in Dewalt & Dewalt, 2002, p. 76). Furthermore, the protocols were used to help promote consistency across observers given that a team of several researchers carried out the observations. These protocols were comprised primarily of checklists for specific activities (e.g., reading text passages, showing videos, guiding class discussions, running simulations, completing worksheets, etc.) that the observer could indicate Y/N as to whether the activity occurred during the observation. Observers also recorded the start and end times for each section as well as notes on how teachers guided students through reading text and class discussions. Additional notes related to modifications of the module curriculum, challenges encountered by teachers during the module implementation, and any other relevant occurrences were also included in the classroom observation protocol. These protocols were completed during each classroom observation. Nearly all science modules were formally observed at least once during the time period between Spring, 2016 and Spring, 2017. This time period was targeted because by this point in the project, most modules had at least their first iteration finalized and ready for full implementation. Four researchers carried out this series of observations.

Co-Lab Posts

Teachers were asked to respond to a series of open-ended prompts about their experiences with the module. These responses were done in the context of an online forum hosted on a Google site called the “co-lab”. Co-lab posts provided teachers an opportunity to provide more holistic feedback on the module implementation, focusing on details and module components of their own choosing, as compared to the more scaffolded and directed feedback elicited in the enactment surveys. The co-lab posts complemented the enactment surveys and provided researchers with a means of contrasting teacher data across multiple sources.

The prompts to which teachers responded in the co-lab posts were written by the curriculum developers and were largely specific to either the module itself or the STEM practice it addressed. A prompt asking about obstacles and barriers to implementation was included for each module. General topics on which these prompts focused across modules include student misconceptions around and understanding of the STEM practices and specific adaptations that teachers made to the module. The supportive, collaborative, and informal nature of the co-lab setting was intended to provide a space where teachers would feel comfortable sharing both their positive and negative experiences with the modules. Sample prompts are provided in Appendix E.

Another important feature of the co-lab posts is that they served as a mechanism for teachers to share their experiences with, and feedback on, the module with each other. Teachers are distributed across four middle schools within the county, and they rarely have a chance to meet with teachers outside of their school, other than at the summer institute. The co-lab posts allowed teachers to learn about the experiences teachers outside of their schools had with the modules, and to gain insights about possible adaptations, additional materials and resources, etc.

Results

Over a period of several academic years, various types of data on the module implementations, teacher and student perceptions of the modules, and student learning were collected, analyzed, and shared with the curriculum developers; curriculum developers then used these findings to inform subsequent iterations of the modules.

While this paper is intended to focus on the overall evaluation approach and methodology, we present illustrative results from the module evaluation here to demonstrate the outcomes of the various components of the evaluation effort and how the results were compiled and shared to promote iterative curriculum development.

Student Assessments

Analyses using models based on Rasch measurement theory (Rasch, 1960) were used to test both the psychometric properties of module assessments and also to examine the degree to which there was evidence of improvements in student achievement on these assessments over the time period from before instruction to after instruction (approximately one-week duration in most cases). The results of this validation effort indicated that overall, the items functioned properly, in that their functioning aligned with the expectations of the Rasch models.

Pre-post comparisons were available for the six science modules with assessments implemented during Fall, 2016 (the three Data-Driven Decision-Making science modules did not have assessments due to their focus on a practice that does not lend itself well to a brief multiple-choice assessment). Across all students who participated in the module administration, positive achievement gains were evident for all six science modules implemented in Fall, 2016. These gains range in size from 0.09 to 0.55 logits; of these, the gains were statistically significant at the level of $p < .05$ for three of the modules.

However, we note the limitations of this pre-post design in which a control group was not used. Test effects, in which improvements arise from prior exposure to the test, could be present in our data because the same version of the test was used in the pre and post administration (Marsden & Torgerson, 2012). Issues of maturity, in which students improve simply by virtue of getting older, and history, whereby improvements are driven by simultaneous experiences external to the intervention, should be minimized by the short duration (one week in most cases) between the pre and post administrations (Marsden & Torgerson, 2012). Additionally, we did not measure more distal knowledge retention through the use of a post test administered at some further time point from the intervention's conclusion. A more robust design would include the use of a control group and/or a longer time horizon post-test; these are design elements we will consider including in future work.

Classroom Observations, Enactment Surveys, and Co-Lab Posts

Classroom observations. Classroom observations allowed researchers to directly observe module implementation taking place in the classroom. In many cases, observers noted smooth and well-functioning implementations, successful utilization of teacher additions, effective modifications to the curriculum, and high levels of student engagement. For example, the 7th grade Data Visualization module asks students to use rubrics to evaluate the health of organisms within deep sea ecosystems. The curriculum initially included an example related to criteria for safe driving in an effort to familiarize students with the notion of using a rubric. The observer described the limitations observed when the original rubric activity was used, noting that the teacher supplemented the

ISSN:2149-8504 (online)

module with an additional rubric activity (for evaluating the quality of hamburgers) that seemed more relevant to students:

> Students are not old enough to drive, and needed more rubric examples. Maybe a more relatable example?....[the teacher] had to provide a lot of additional scaffolding for the rubric, and how to use rubrics. She created the highly relatable McDonald's rubric…

In some cases, researchers observed logistical difficulties related to problematic manipulatives, student and/or teacher confusion driven by insufficient or unclear instructions, and activities with components targeting a difficulty level that exceeded many students' ability levels. One example of this occurred during the classroom observation of the 7th grade Data-Driven Decision-Making science module, which focuses on weighing the economic benefits of fishing and tourism against their potentially negative effects on the health of coral reefs. An activity in this module uses plastic counters to enable students to run a population simulation. During this simulation, the observer noted:

> Students in some cases needed extra counters for the 'reproduction' events at the start of each year – after a while, [the teacher] realized this and told students she had extra counters for this situation. But some students had already proceeded through a few years of the simulation without replacing the counters as specified because they did not have them in their discard counter. We may want to consider revising instructions to make clear to teachers and students that they may need to get additional counters during the simulation.

In this case, observation notes pointed to a somewhat minor component of the instructions that was nonetheless critical to students' successful progress through the simulation, and suggested that this part of the instructions needed to be clarified and/or emphasized given that it was initially overlooked by both students and the teacher. The observation notes also indicated additional problems occurring during this activity that resulted from the teacher having set up and distributed the manipulatives incorrectly prior to the students beginning the activity. This finding suggested a need to clarify instructions in the Teacher Edition and perhaps emphasize the importance of correct set-up for this particular module during professional development sessions.

Enactment surveys and co-lab posts. Multiple teachers often noted the issues apparent in our observation data in their enactment surveys and/or co-lab posts, allowing researchers to synthesize data across these sources and pinpoint problematic areas within the module curriculum. Echoing the observer note above about replacing counters during the population simulation, the same teacher wrote in her co-lab post:

> I also adjusted students visiting the 'Pet Store'. Some of my students were not making the connection to come pick up extra fish. If I had to choose any part to be the most difficult, it would be the discard and replace the population.

Another teacher provided similar feedback regarding general student confusion during the simulation activity, noting:

> The students were so confused during the Sorting/Counting section. I had to explain the process to the students numerous times before they actually began to grasp what should be done or how many organisms would be removed.

Receiving similar feedback from multiple teachers in the co-lab posts served the dual purposes of helping the research and curriculum teams hone in on points of confusion across multiple classrooms and also providing teachers with the supportive feeling of not being alone in their challenges with the modules.

Co-lab posts and enactment surveys also provided evidence of instances where the curriculum was working exactly as the curriculum writers intended. In the 6th grade Lava Challenge Experimental Design module, students write their own procedures to model how lava flows down a mountain. The intention of the module is that students' initial procedures are inconsistent and not sufficiently detailed, which results in a wide spread of data collected across student groups. As students refine and improve their procedures for subsequent trials, the data should become more uniform. This approach requires teachers to allow students to flounder a bit early on in the module; this is something teachers often shy away from, preferring instead to correct students' work immediately or prevent them from making mistakes in the first place. Evidence from the co-lab posts and enactment surveys shows that teachers embraced this component of the module:

> During the first trial, I did not give students input on their procedures. I wanted them to feel lost in a way. I wanted them to learn from their mistakes. If they had vague procedures, they had a difficult time doing the trials. They then learned the hard way that they need to write more specific procedures. – Enactment Survey Response
>
> We reviewed the procedures, but I let the students complete their own procedures…If they weren't detailed, they were lost when they were completing the experiment. This is memorable. – Co-Lab Post

Enactment surveys also provided specific timing details from multiple teachers for each module, allowing curriculum developers to tailor the module's activities as needed to reliably fit within one week of instructional time. The preparatory guide Teacher Edition also included pacing information for each section. Those pacing times came directly from what teachers reported in the enactment surveys, giving more credibility to the Teacher Editions.

The yes/no checklist allowed examination of which, if any, activities were being skipped by teachers. As an example, in the introductory section to the Deep Sea Ecosystems module, there were several suggestions for activities and resources for teachers to provide additional content beyond just the text in the Student Edition: a provided website with trivia questions, a provided website about the Deepwater Horizon oil spill, and an opportunity to supplement with ecology-related vocabulary and content of the teacher's choosing. Of the five teachers who completed this module in Fall, 2016, these suggested activities were completed by four, three, and three teachers, respectively. Each teacher skipped only one of these activities, suggesting that they considered the various options and selected those that worked best for them.

Module Reports

Formal module reports were created for each module administered during the 2016-2017 school year. All module evaluation products described above were compiled in these reports. The research team distributed these reports to the curriculum team, and then held a joint meeting in which the module evaluation results were presented and discussed with the curriculum developers. The contents of these reports were then taken into consideration during the revision process of modules for the 2017-2018 school year.

Discussion

There are numerous examples of instances in which module feedback from one of the data sources described above directly informed a revision to the subsequent version of the module curriculum and also helped curriculum developers and researchers evaluate the extent to which the curriculum reached its objectives. One such example occurred with the Deep Sea Ecosystems Challenge, which is the 7th grade life science Data Visualization module. During this module students learn how to analyze images and quantify data from these images. One activity involves students creating a rubric to evaluate two images. Originally, students analyzed images of what would be considered safe driving and unsafe driving. Teachers reported that students struggled with this rubric activity because they could not relate to driving. To combat this problem, one teacher replaced the driving photos with her own images of hamburgers instead. As a result of the finding that the driving related rubric did not resonate with students, the curriculum was revised to include the teacher-created sample rubric for what makes a good hamburger, and the safe driving sample was eliminated. The observation notes in this case provided the curriculum team with both evidence of a problem with a component of the curriculum as well as a solution one teacher had created to deal with the issue.

In another example, consistent issues related to confusion with instructions and manipulatives that were noted in the coral reef population simulation led curriculum writers to make changes to both the student and teacher materials. The curriculum team created separate student handouts with detailed procedures for the activity, enhanced both teacher guides with tips and suggestions for navigating the activity, and created a separate highly detailed instruction guide for preparing materials, to ensure proper setup of the manipulatives. For this module, the manipulatives themselves were also changed due to teacher feedback, from split peas to plastic counters that were more uniform in size and shape and were easier for students and teachers to work with.

The collection of data across these multiple sources allowed for the synthesis of findings regarding successful and problematic areas within the curriculum for each module. Sharing these findings in a concise and timely manner with the curriculum developers supported an iterative development cycle in which data directly informed the subsequent iteration of each module. This micro-level approach, focused on specific aspects of teacher and student experiences within the modules, provided insights that a more macro approach would not have allowed for. These insights guided the creation of versions of these modules in which elements leading to confusion and problems were clarified and improved upon. This multi-source data collection effort aligned well with the complexity inherent in the project, and directly supported the optimization of the curriculum, activities, and materials for each module.

Acknowledgement

We would like to acknowledge Emily Frobos, Anna Holcomb, and Olivia Shellman for their assistance with this project.

References

Bybee, R. W. (2015). *The BSCS 5E instructional model: Creating teachable moments*. Arlington, VA: NSTA Press.

Bybee, R. W., Taylor, J. A., Gardner, A., Van Scotter, P., Carlson Powell, J., Westbrook, A., & Landes, N. (2006). *The BSCS 5E instructional model: Origins and effectiveness*. Colorado Springs, CO: BSCS.

DeWalt, K. M. & DeWalt, B. R. (2002). *Participant observation: A guide for fieldworkers*. Walnut Creek, CA: AltaMira Press.

Fishman, B. J., Penuel, W. R., Allen, A., Cheng, B. H., & Sabelli, N. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. *National Society for the Study of Education*, *112*(2), 136–156.

Flick, U. (2009). *An Introduction to Qualitative Research (4th ed)*. Thousand Oaks, CA: Sage.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items (2nd ed.)*. Mahwah, NJ: Erlbaum.

Klassen, S. (2006), Contextual assessment in science education: Background, issues, and policy. *Science Education*, *90*(5), 820–851.

Kolodner, J.L., Camp, P., Crismond D., Fasse, B., Gray, J., Holbrook, J., Puntembakar, S., and Ryan, M. (2003). Problem-based learning meets case-based reasoning in the middle-school science classroom: Putting Learning-by-Design™ into practice. *Journal of the Learning Sciences*, *12*(4), 495–548.

Kolodner, J. L., Crismond, D., Gray, J., Holbrook, J., & Puntambekar, S. (1998). Learning by design from theory to practice. *Proceedings of the international conference of the learning sciences*, *98*, 16-22.

Marsden, E. & Torgerson, C. J. (2012). Single group, pre- and post-test research designs: Some methodological concerns. *Oxford Review of Education*, *38*(5), 583-616.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards, Standards of Mathematical Practice*. Washington, DC: National Governors Association Center for Best Practices.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement*, *22*(4), 287 – 293.

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

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).

Ruiz-Primo, M. A. (2006). *A multi-method and multi-source approach for studying fidelity of implementation*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Report: Los Angeles, CA.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, *35*(3), 265–296.

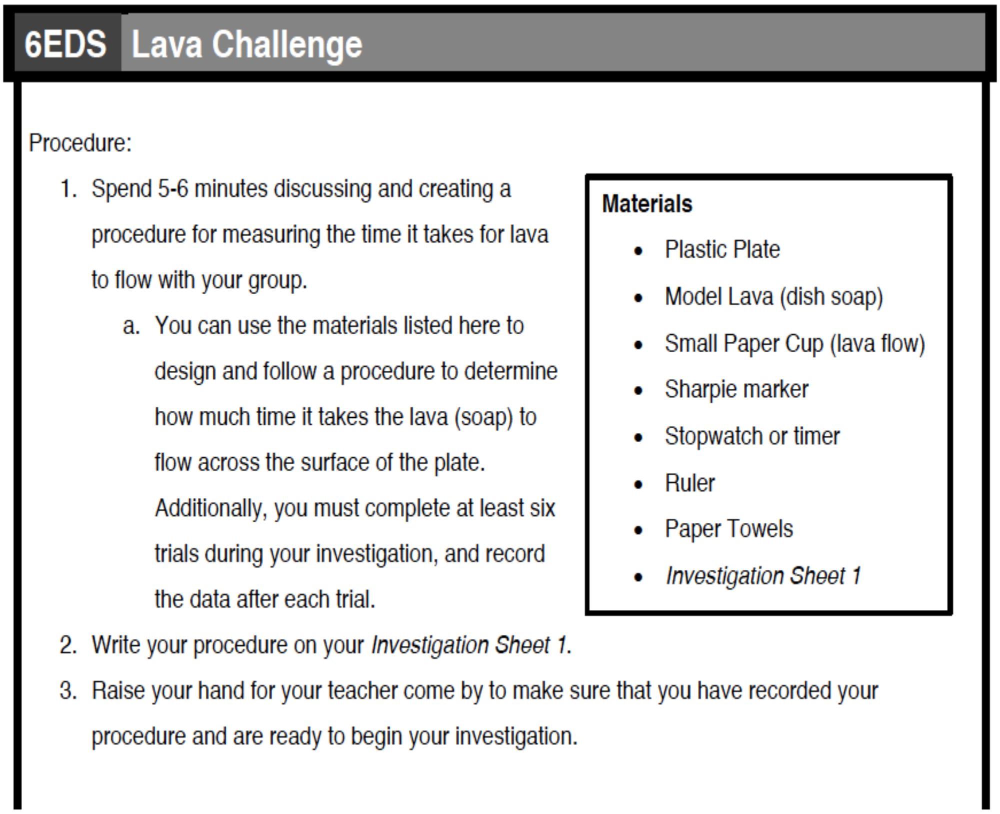
Appendix A

All modules and associated materials are available for free download via our project website. The website can be accessed using the following link: https://ampitup.gatech.edu

Appendix B

Sample Module Pages from the Student Editions

Part 1: Procedure Writing Instructions, 6th Grade Experimental Design Module



6EDS Lava Challenge

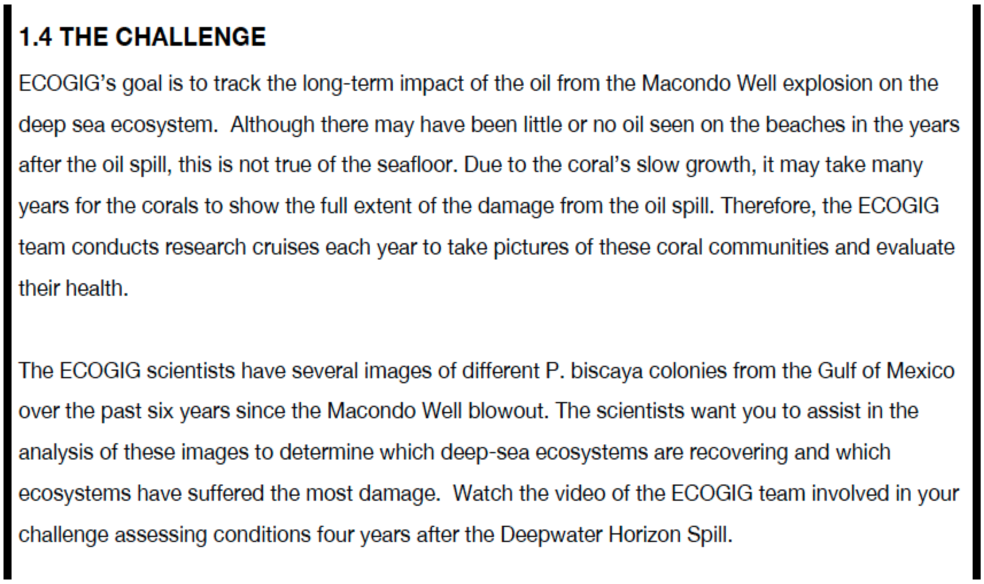
Procedure:

1. Spend 5-6 minutes discussing and creating a procedure for measuring the time it takes for lava to flow with your group.
    a. You can use the materials listed here to design and follow a procedure to determine how much time it takes the lava (soap) to flow across the surface of the plate. Additionally, you must complete at least six trials during your investigation, and record the data after each trial.
2. Write your procedure on your *Investigation Sheet 1*.
3. Raise your hand for your teacher come by to make sure that you have recorded your procedure and are ready to begin your investigation.

Materials

- Plastic Plate
- Model Lava (dish soap)
- Small Paper Cup (lava flow)
- Sharpie marker
- Stopwatch or timer
- Ruler
- Paper Towels
- *Investigation Sheet 1*

Part 2: Challenge, 7th Grade Data Visualization Module

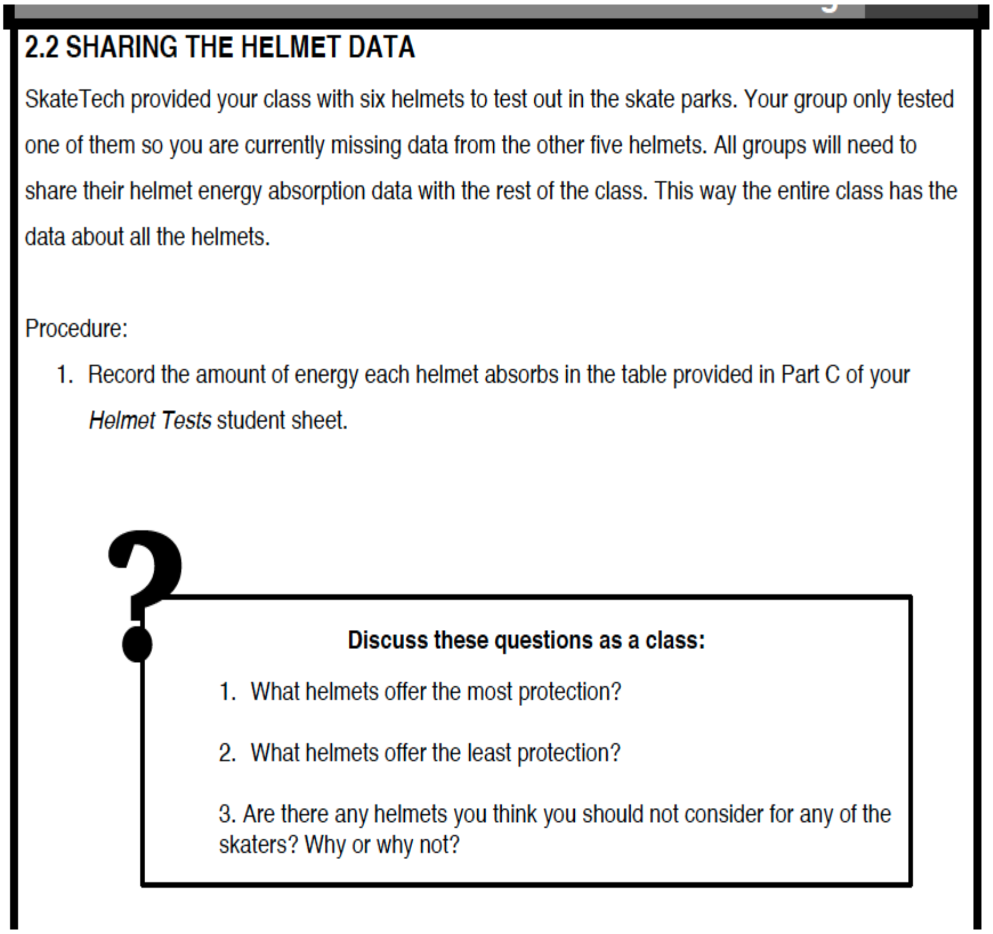


1.4 THE CHALLENGE

ECOGIG's goal is to track the long-term impact of the oil from the Macondo Well explosion on the deep sea ecosystem. Although there may have been little or no oil seen on the beaches in the years after the oil spill, this is not true of the seafloor. Due to the coral's slow growth, it may take many years for the corals to show the full extent of the damage from the oil spill. Therefore, the ECOGIG team conducts research cruises each year to take pictures of these coral communities and evaluate their health.

The ECOGIG scientists have several images of different P. biscaya colonies from the Gulf of Mexico over the past six years since the Macondo Well blowout. The scientists want you to assist in the analysis of these images to determine which deep-sea ecosystems are recovering and which ecosystems have suffered the most damage. Watch the video of the ECOGIG team involved in your challenge assessing conditions four years after the Deepwater Horizon Spill.

Part 3: Class Discussion, 8th Grade Data Driven Decision Making



2.2 SHARING THE HELMET DATA

SkateTech provided your class with six helmets to test out in the skate parks. Your group only tested one of them so you are currently missing data from the other five helmets. All groups will need to share their helmet energy absorption data with the rest of the class. This way the entire class has the data about all the helmets.

Procedure:

1. Record the amount of energy each helmet absorbs in the table provided in Part C of your *Helmet Tests* student sheet.

?

Discuss these questions as a class:

1. What helmets offer the most protection?
2. What helmets offer the least protection?
3. Are there any helmets you think you should not consider for any of the skaters? Why or why not?

Apendix C

Sample Assessment Items, 7th Grade Data Visualization Module

Use the following information to answer question 1.

You have collected data from your school's cafeteria on the quantities of certain types of food sold each day within a given week. Here are your data:

Day	Pizza (# slices)	Hamburgers	Caesar Salads
Monday	100	50	200
Tuesday	150	50	150
Wednesday	200	100	100
Thursday	50	200	50
Friday	200	150	100

1. The cafeteria is considering removing the food that was eaten the least for the week. Which food should they eliminate?

   A. Pizza

   B. Hamburgers

   C. Caesar Salads

   D. Pizza & Hamburgers (they have equal mode values)

2. In which of the following situations would a rubric be most helpful?

   A. Searching for an online resource

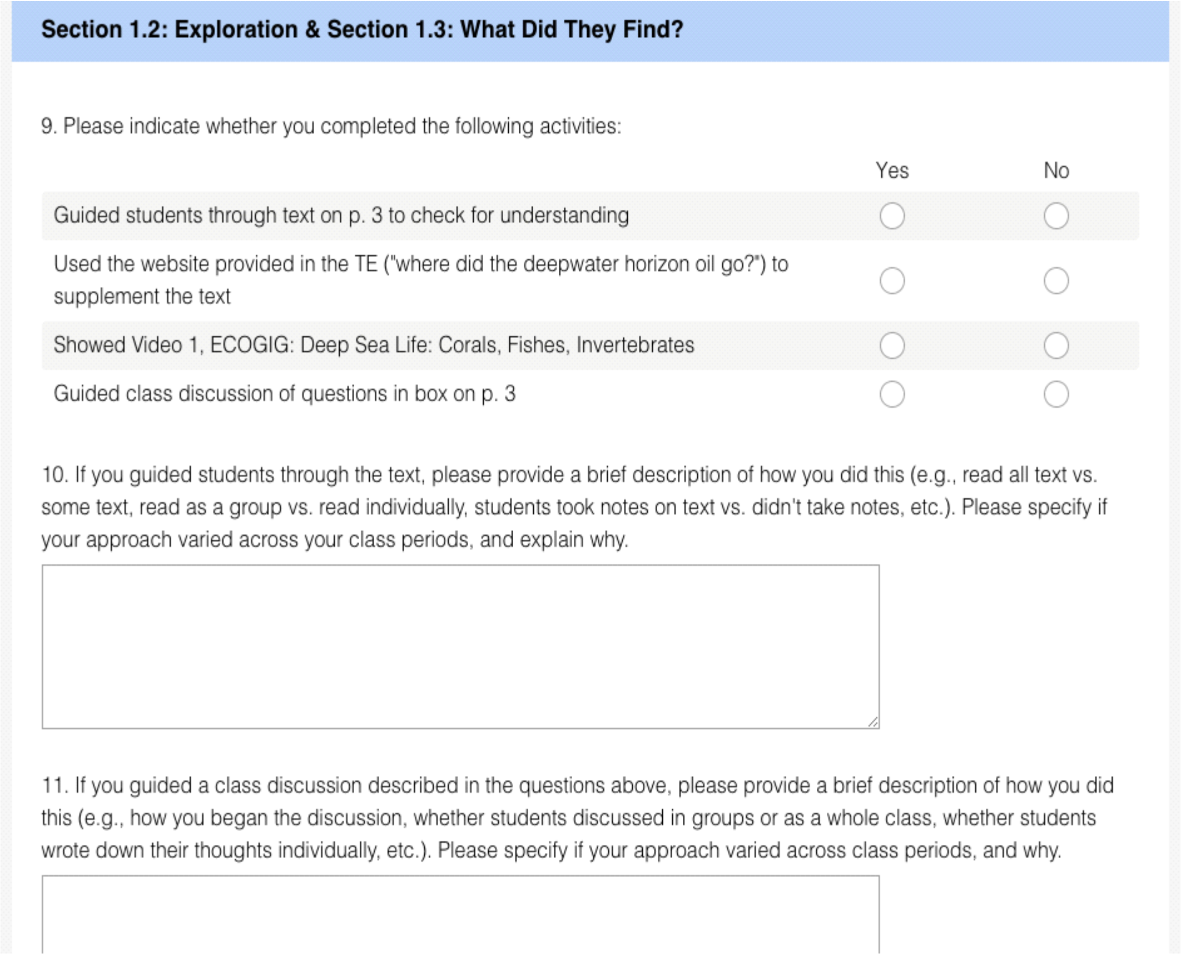
   B. Designing an experiment

   C. Evaluating your classmates' presentations

   D. Writing a report on the results of your research

Appendix D

Sample Enactment Survey Items, 6th Grade Experimental Design Module



Section 1.2: Exploration & Section 1.3: What Did They Find?

9. Please indicate whether you completed the following activities:

	Yes	No
Guided students through text on p. 3 to check for understanding	○	○
Used the website provided in the TE ("where did the deepwater horizon oil go?") to supplement the text	○	○
Showed Video 1, ECOGIG: Deep Sea Life: Corals, Fishes, Invertebrates	○	○
Guided class discussion of questions in box on p. 3	○	○

10. If you guided students through the text, please provide a brief description of how you did this (e.g., read all text vs. some text, read as a group vs. read individually, students took notes on text vs. didn't take notes, etc.). Please specify if your approach varied across your class periods, and explain why.

11. If you guided a class discussion described in the questions above, please provide a brief description of how you did this (e.g., how you began the discussion, whether students discussed in groups or as a whole class, whether students wrote down their thoughts individually, etc.). Please specify if your approach varied across class periods, and why.

ISSN:2149-8504 (online)

Appendix E
Sample Co-Lab Prompts

Prompts specific to the Lava module

- Were you able to get varying results after the first procedure?
- Did any of the students have trouble getting results?
- What did their models look like?
- Were you able to get consistent results after the second procedure?

Prompts used across modules

- What adaptations or modifications did you make to the module?
- Did you include any additional content in the module?
- Were there any challenges or obstacles to implementing this module?
- What recommendations do you have for improving the module?

ISSN:2149-8504 (online)