International Journal of Science and Mathematics Education
https://doi.org/10.1007/s10763-022-10295-1

Effects of Data Nuggets on Student Interest in STEM Careers, Self-efficacy in Data Tasks, and Ability to Construct Scientific Explanations

Elizabeth H. Schultheis[1,2] · Melissa K. Kjelvik[1] · Jeffrey Snowden[3] · Louise Mead[1] · Molly A. M. Stuhlsatz[3]

Received: 19 February 2021 / Accepted: 23 May 2022

Abstract

This paper describes a randomized and controlled efficacy study conducted in high school biology classrooms across the USA. In this study, teachers implemented the use of Data Nuggets, activities designed to bring real research and data into the classroom. These materials can be embedded within the existing instructional modality of any given curriculum, thus infusing these curricula with science stories and associated datasets. Our design had teachers incorporate Data Nuggets into one of their class sections, while teaching a second class section in a business-as-usual manner. Although students in both conditions improved similarly in quantitative reasoning over the course of the study semester, we saw several key differences for students taught using the intervention as compared to those taught using only standard instruction. Students in classrooms that utilized Data Nuggets spent more time engaged in the practices of science and had greater improvement in their ability to construct scientific explanations. In addition, students using the intervention activities showed increases in both their self-efficacy in data-related tasks and their interest in STEM careers. Finally, the effects of teacher implementation on student outcomes when using Data Nuggets were assessed.

Keywords Authentic data · Career motivation · Claim-evidence-reasoning · Data literacy · Quantitative reasoning · Science storytelling · Self-efficacy

✉ Elizabeth H. Schultheis
eschultheis@gmail.com

[1] Michigan State University, East Lansing, MI, USA

[2] Kellogg Biological Station LTER, Hickory Corners, MI, USA

[3] BSCS Science Learning, Colorado Springs, CO, USA

Published online: 27 June 2022

Introduction

Quantitative abilities are necessary to engage in both science and society, particularly skills related to quantitative literacy, rationality, and the ability to apply scientific and mathematical thinking in everyday endeavors (Capraro et al., 2014; Mayes et al., 2014; National Research Council [NRC], 2014). Therefore, a goal of STEM education should be to produce graduates who are adept in these areas (Capraro et al., 2014; Wise, 2020). Data Nuggets are classroom activities designed to engage students in the work of scientists with the goal of improving student quantitative abilities as well as their interest, motivation, self-efficacy, and engagement in STEM.

Originally inspired by conversations between teachers and scientists, Data Nuggets, found at https://datanuggets.org, bring authentic data from cutting-edge research into the classroom (Schultheis & Kjelvik, 2015). Each activity is written by the scientist behind the research and data, allowing the activities to include first-hand expertise. When completing an activity, students engage in the work of scientists by reading scientific texts, creating and interpreting graphs, doing basic statistics, asking and answering scientific questions, and constructing explanations. Because of their simplicity and flexibility, Data Nuggets can be used repeatedly throughout a single semester, as well as across K-12 and undergraduate levels.

In the present study, we aimed to determine the effects of Data Nuggets on the students who use them. We conducted a controlled and randomized efficacy study in high school biology classrooms across the USA. In our design, teachers used Data Nuggets activities in one section of their classes, while teaching their other section in a business-as-usual manner as a comparison. At the start and end of the study semester, we measured student quantitative reasoning (QR) abilities, interest, motivation, and self-efficacy. In addition, we measured engagement at the end of the semester. During the semester, we observed implementation of resources in all classrooms.

Theoretical Framework

Engaging in Science Practices Constructivism is the foundation of Data Nuggets development. In the USA, we see the roots of constructivism reflected in today's science standards, including the Framework for K-12 Science Education (National Research Council [NRC], 2012) and the Next Generation Science Standards (National Research Council [NRC], 2013). These standards detail eight science and engineering practices. Our materials focus on four in particular: (#4) Analyzing and interpreting data, (#5) Using mathematics and computational thinking, (#6) Constructing explanations, and (#7) Engaging in argument from evidence. Practices (#4) and (#5) cover students' abilities to summarize and analyze data using statistics, graph creation, and the interpretation of meaning based on numerical information. In Data Nuggets, students work with real datasets to conduct basic statistics, such as calculating averages and ranges in data, and visualize data by graphing (Fig. 1). Practices (#6) and (#7) cover students' abilities to use data to answer scientific

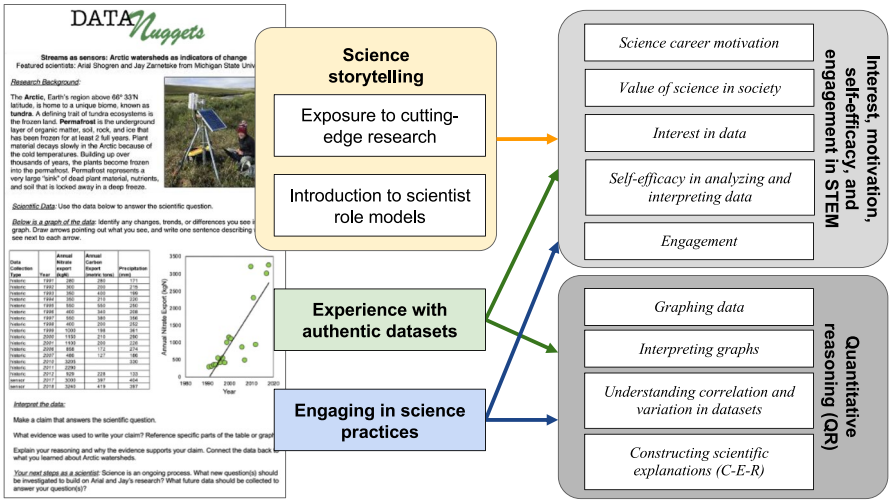

Fig. 1 Conceptual framework for the intervention's impact on students. Example activity (column 1) is labeled with the features (column 2) we hypothesize will impact student outcomes (arrows and column 3)

questions, support claims, and draw conclusions. In Data Nuggets, students read scientific text, and then use their graphs to answer scientific questions posed by the scientist. This requires the use of data as evidence to answer the question and construct explanations and arguments (Fig. 1). When using multiple applications of this supplement, students get repeated exposures, experiences, and engagements in these practices. The following sections clarify the connections between the sections of Data Nuggets activities and specific scientific practices.

Science Storytelling In addition to engaging students in the practices of science, a main design principle behind Data Nuggets is the use of science storytelling in the Research Background (Fig. 1). Science stories capture the cause-and-effect explanations of science and place them within a narrative structure (Collins, 2021; Hoffmann, 2014; Wilson, 2002). The use of stories in science is a powerful strategy to make content more interesting, easier to understand, and more memorable (Britton, 1983; Graesser et al., 1994; Willingham, 2004). Stories capture our interest because they appeal to our desire for causality and goals (Gentner, 1976; Wilson, 2002). The stories within our materials are told through writing and images. They center on (a) exposure to cutting-edge research and (b) introduction to scientist role models.

Exposure to cutting-edge research. Science is more than a body of knowledge. It is also an ongoing endeavor to understand the world. However, traditional laboratory activities often require students to follow a set of procedures that lead to a predetermined outcome (Gould et al., 2014; Schultheis & Kjelvik, 2015). This may lead to the misconception that science is about learning a set of facts, and engaging in science is simply demonstrating known concepts. Therefore, the exploration of the unknown and developing a set of skills and tools to make these discoveries is

equally important as content knowledge in science education (Hammett & Dorsey, 2020; NRC, 2012). Data Nuggets create opportunities for students to develop their conceptual understandings of the world through science. By working with data from cutting-edge research, students take on the role of active scientists and explore questions that have not yet been answered (Schultheis & Kjelvik, 2015, 2020). A short Research Background is provided that details the history and development of a scientific area of inquiry and the motivation behind the research. It goes on to detail the particular study's contribution to the field (Fig. 1).

Introduction to scientist role models. Scientists highlighted in Data Nuggets come from a variety of backgrounds, representing diversity in gender, race, age, career stage, and other elements of identity. The Research Background section follows their journeys as they formulate their ideas and endeavor to answer scientific questions (Fig. 1). Including scientist role models in instruction has been shown to help students develop their STEM identities, particularly when a student perceives that they share similarities with a particular role model (Collins, 2021; Estrada et al., 2011). Highlighting role models in education has also been shown to increase interest in science (Schinske et al., 2016), improve grades and intellectual growth (Schinske et al., 2016), and support decisions to pursue STEM careers (Gibson, 2004). With this intervention, students have direct experiences with activities written by real scientists and have the opportunity to follow up with these scientists to ask questions and learn more (Fig. 1).

Experience with Authentic Datasets Each Data Nuggets activity centers around an authentic dataset, which students use to construct graphs and construct scientific explanations (Fig. 1). We argue there are unique learning opportunities that arise from the use of these authentic datasets in the classroom (Kjelvik & Schultheis, 2019; Schultheis & Kjelvik, 2020). Real datasets from scientific research are often messy, have high variability and outliers, and might be missing values, all of which are factors that create learning opportunities (Wilkerson et al., 2021). For example, variability, or the extent to which data points differ from one another and the average (Bargagliotti et al., 2020), can be used to highlight key features of a study system or experimental design (Gould et al., 2014; Kjelvik & Schultheis, 2019). With these activities, variability can be explored as students create graphical displays of data to represent variation around a mean or trend. Exploration of variability can also inform their scientific explanations as they grapple with whether the dataset provides support for a hypothesis. As another example, a missing data point can be used to stimulate class discussion about the realities of research, such as weather conditions preventing data collection, or a sample being lost. The use of messy data has the added benefit of showing students that their own datasets are often of similar quality to those collected by scientists (Gould et al., 2014), a fact that may increase student confidence in their own abilities.

Literature Review

Here, we discuss the importance of the STEM constructs that are foundational to the Data Nuggets program and central measures in our efficacy study.

Student Quantitative Reasoning Abilities Quantitative reasoning (QR) captures a student's ability to think mathematically and understand numerical information (Mayes et al., 2014). QR skills include graphing, interpreting graphs, understanding correlation and variation in datasets, and constructing scientific explanations and arguments (Fig. 1). QR also includes the ability to apply mathematical principles in real-world scenarios using higher-order critical thinking and logic (Mayes et al., 2014). In order for students to fully develop these skills during their education, they need to repeatedly practice and apply them in a variety of contexts (Holmes et al., 2015; Neumann et al., 2013). Students also need to see mathematics as relevant to their lives; learning mathematics in context reinforces its importance and may keep students more engaged (Šorgo et al., 2010).

Constructing scientific explanations is a key component of QR and requires students to identify data as evidence and justify why evidence supports any given claim, using scientific concepts and understanding (Toulmin, 1958; McNeill & Krajcik, 2008b; NRC, 2013; Jin et al., 2020). This process relies on students' ability to comprehend and incorporate information from scientific texts, including data, into their own understanding of the scientific phenomenon. Pedagogical practices such as scaffolding (Osborne et al., 2004) and faded scaffolding (McNeill et al., 2006) have been shown to be beneficial when helping students learn to construct scientific explanations. In Data Nuggets, scaffolding is incorporated surrounding both the graphing task and constructing explanation tasks. For constructing explanations, we scaffold the task using the claim-evidence-reasoning (C-E-R) framework (McNeill & Krajcik, 2008b). Furthermore, teachers can choose to use a faded scaffolding approach (McNeill et al., 2006), which (1) starts with a tool that breaks scientific explanations down into basic sub-components (lowest level), (2) uses the C-E-R prompts within the activity (middle level), or (3) removes the prompts and has students construct explanations with no guidance on structure (highest level). For graphing, teachers can choose to use three different levels: (1) graph provided for students (lowest level), (2) axes and scales provided but students must add the data to the graph (middle level), and (3) blank graph paper where students must fully create the graphs on their own (highest level).

Student Interest, Motivation, Self-Efficacy, and Engagement in STEM Effective QR also involves beneficial habits of mind and emotional responses toward quantitative information. This includes the learner's interest in data and course material (Aikens & Dolan, 2014), as well as their disposition to engage with quantitative information (Karaali et al., 2016). These affective constructs are important indicators of success and persistence in STEM. Avenues to improve QR according to the factors outlined below are strategically incorporated in this intervention. Discussion of this incorporation is in the "Methods" section of this paper.

Engagement. Students in both the USA and internationally are not fully engaged in science and mathematics classrooms (Hiebert & Stigler, 2000; Shernoff et al., 2000; Yair, 2000), meaning the traditional ways of teaching these

subjects are not holding their attention (Ahlfeldt et al., 2005). Steps should be taken to actively engage students in the classroom, as more engaged students tend to have increased understanding of content and increased motivation to pursue STEM careers (Franz-Odendaal et al., 2016). This can be achieved by providing more opportunities for students to contribute to class discussions and ask questions. The use of cutting-edge research engenders these opportunities. Explorations of the unknown allow for classrooms in which teachers are no longer the sole distributors of knowledge, thus putting students on a more equal footing where everyone is exploring something new together.

Interest in data. When students enjoy classroom materials and understand the importance of what they are learning, they are more likely to be more interested in that material (Linnenbrink-Garcia et al., 2010). Therefore, student interest in data is likely shaped by the ways in which they are introduced in the classroom. For example, if data are used to teach a mathematical concept devoid of context, students may see activities as busy work and lose interest. However, if data are presented in context and as the means to answer important questions that have relevance to students' lives, interest may increase.

Value of science in society. Students must be shown the value of classroom activities in their STEM education, within their own communities, and throughout the world at large in order for them to see scientific work as interesting, important, and meaningful (Hammett & Dorsey, 2020; Schultheis & Kjelvik, 2020). Value falls within interest (Hulleman & Harackiewicz, 2009; Linnenbrink-Garcia et al., 2010) and may come from personal utility, when students see a set of skills as useful for their everyday lives (Hulleman & Harackiewicz, 2009), or from field utility, if they see the work of scientists as generally important (Linnenbrink-Garcia et al., 2010). In order to build student interest, classroom materials should emphasize that science is an ongoing and active enterprise that can help all members of society address questions, including questions important to students' own personal and, in the future, professional lives.

Science career motivation. In order to foster the next generation of scientists and STEM professionals, we need to help students build science identities. This includes identifying themselves as individuals who know about and can contribute to science (Center for the Advancement of Informal Science Education, 2018). Formation of a science identity has been shown to be a strong indicator of persistence in STEM, including motivation to pursue these fields throughout their education and careers (Chemers et al., 2011; Simpson & Bouhafa, 2020). Sharing a diversity of science stories gives students the opportunity to see themselves in these narratives, increasing motivation to pursue a career in science (Gladstone & Cimpian, 2021; Murcia et al., 2020). Students may also be more confident in their abilities if they are able to observe the behaviors of scientists who are conducting research and collecting data.

Self-efficacy in analyzing and interpreting data. Student self-efficacy when conducting STEM tasks is a measure of their perceived abilities and confidence in taking on the tasks and functions of a scientist (Chemers et al., 2011). Self-efficacy has been shown to predict a variety of behavioral outcomes, including perseverance and persistence in STEM fields (Bandura & Locke, 2003; Estrada et al., 2011). Increased self-efficacy can also lead to increased engagement, interest, satisfaction,

and expectancy for success (Lent et al., 2018). Reading stories in which scientists overcome challenges during research and data collection may help students build their self-efficacy when faced with similar hurdles.

The Purpose of This Study

To investigate the impacts of Data Nuggets on students, the research team conducted a randomized and controlled efficacy study in high school biology classrooms. We asked:

- **Research Question 1:** *To what extent do students in classrooms using Data Nuggets show improved quantitative reasoning (QR) compared to students in business-as-usual comparison classrooms?*
- **Research Question 2:** *To what extent do students in classrooms using Data Nuggets have higher interest, motivation, self-efficacy, and engagement in STEM compared to students in business-as-usual classrooms?*
- **Research Question 3:** *Does the quality of teacher implementation of the intervention moderate student outcomes?*

Methods

We used a cluster randomized design in which sections of high school biology classes were randomly assigned to learn biology using either the business-as-usual traditional curriculum (comparison), or the intervention, in which Data Nuggets activities were integrated into the traditional curriculum (treatment). While our data were collected at the student level, the results are reported at the section level, with students nested within the context of their teacher and section.

Research Participants and Experimental Procedure We recruited our study population from Michigan, Colorado, Illinois, and California, in the USA. All high school biology teachers in these states who expected to teach at least two sections of general high school biology in the following school year were invited to apply. From an initial pool of applicants, we selected 25 teachers to participate. Three teachers left the study before randomization occurred.

Each participating teacher taught at least two sections of the same course. One section for each teacher received the treatment and the other continued learning using the standard comparison.We randomly assigned sections to each condition. While random assignment of classrooms (by teacher) increased the likelihood that teacher effects would be distributed evenly across treatment groups, variations in classroom context were impossible to control. To account for this, we tested the baseline equivalence of the treatment groups using pretest measures,

to determine if there were significant differences between the treatment and comparison classrooms prior to the start of the study.

Assignment resulted in 46 classrooms (23 treatment and 23 comparison) nested within 22 teachers (one teacher had 4 classes in the study, 2 treatment and 2 comparison). One teacher completed the study, but pretest data was lost in the mail and so was not included in the analysis because we were unable to control for the pretest. Thus, 21 teachers and 44 classrooms were included in the statistical analyses for a total of 934 students participating in the study (treatment = 470, comparison = 464). Student demographics can be found in Supplemental Materials.

Study teacher professional development (PD). Following selection, in the summer of 2017, participant teachers took part in a 2-day PD workshop. The goal of this PD was to familiarize study teachers with the materials and study requirements. Workshop topics included QR in biology, the process of science, visualizing data using graphs, constructing scientific explanations, and asking questions. Study teachers were provided with opportunities to fully engage with project personnel and were given printed and online resources that they could refer to throughout the study. In addition, mentor teachers familiar with the development and use of the program were available to answer questions.

Because our design was susceptible to internal validity threats, such as contamination or treatment diffusion (Shadish et al., 2002), an important element of the PD was educating teachers about the research. Teachers were provided time to review and align study materials with their existing curricula and plan for implementation of both the treatment and comparison conditions. To mimic the traditional usage of each teacher's program materials, we provided teachers flexibility in selecting activities out of 26 available options. Teachers also determined when in the semester they would implement the activities. In addition, we created and shared a set of potential activities for use in the comparison class. While teachers were instructed to continue with "business as usual" in their comparison sections, these resources provided options to balance instructional time between the two sections, if needed.

Study timeline. We conducted our study in the fall of 2017. At the start of the semester, students completed a series of baseline assessments, an interest, motivation, and self-efficacy survey, and a demographic survey. Following baseline data collection, teachers implemented the intervention activities in their treatment classrooms, and taught "business as usual" in their comparison classrooms. In the treatment, teachers were required to use eight Data Nuggets activities throughout the semester. Students in the comparison group did not use intervention activities in that semester. To keep both instruction courses running at the same pace, teachers used the aforementioned alternative activities in their comparison classrooms to parallel the time spent on the intervention. We allowed teachers complete autonomy to choose these parallel lessons. We created a database where teachers could compile and share resources to reduce workload. At the end of the fall semester, students in both treatment and comparison classes completed the assessments again, in addition to a self-report engagement survey. Once the study was complete, teachers had the opportunity to implement the intervention activities in comparison classrooms in the following spring semester.

Outcome Measures Our outcome measures fell into two categories: (1) student quantitative reasoning in biology, including graphing data, interpreting graphs, understanding correlation and variation in datasets, and constructing scientific explanations; and (2) student affective outcomes, including (a) interest in data, (b) the value of science to society, (c) science career motivation, (d) self-efficacy in analyzing and interpreting data, and (e) engagement in biology class (Fig. 1). Psychometric properties of each measure can be found in the Supplemental Materials.

Quantitative reasoning (QR). To measure QR in the context of biology, we used a vignette-based assessment developed by our research team at BSCS (Stuhlsatz et al., 2020). This instrument captures student understanding of correlation, variation, and time series data. The instrument requires students to analyze and interpret data, use mathematical thinking, and consider competing arguments to explain a scientific phenomenon. Students were randomly assigned to complete one of the three vignettes: (1) correlation between water temperature and oxygen levels in lakes, (2) variation between and within populations of bees and flowers, and (3) a time series evaluation of the prevalence of lung disease over time. The final component of the QR assessment included a graphing and scientific explanation task. This task asked students to read a short fictitious scenario that included a generated data table. Students used the data to create a graph and write a scientific explanation for the phenomenon in the scenario (see Supplemental Material for sample items).

The instrument went through iterative development, including pilot-testing with a sample of 109 high school biology students. During the pilot, we conducted cognitive interviews with six biology students to inform revisions. We received feedback on the content validity of the assessment from two current high school biology teachers and one curriculum developer. The graphing and explanation task section of the instrument was assessed by three scorers using an iterative training process, led by the instrument development lead. The training sample of student responses was randomly selected from the study sample. During the iterative scoring process, raters reached agreement between 81 and 100% on the training sample, with Kappa values ranging from 0.60 and 0.90. Kappa values are a statistic that measure inter-rater reliability while taking into account that agreement between raters may occur by chance (McHugh, 2012).

During pre- and post-testing, each student completed only one randomly assigned vignette per administration. While we wanted to include items that covered all three QR content areas (covariation, variation between and within populations, and time series), we knew that the testing burden could be very high and not feasible for teachers and students. In order to produce just one score, we included a collection of items common across all three forms and used the Winsteps Rasch measurement computer program (Linacre, 2021) to link the three assessment forms into one measure using common item equating (Bond & Fox, 2015; Boone & Scantlebury, 2006). This created a score for each student on the equated measure, resulting in one QR score per student. This resulted in psychometrics for the overall form from the sample of students in the study (Rasch Person Reliability = 0.81, see Supplemental Material). For exploratory analyses (Schochet, 2008), we created a subscale from the explanation task. This subscale was used to investigate the "constructing

explanations" outcome. The three polytomous items in the explanation task produced a Cronbach's alpha of 0.58.

Interest, motivation, self-efficacy, and engagement in STEM. We employed a series of short, affective scales to assess student interest, motivation, self-efficacy, and engagement in both treatment and comparison conditions. Before using these measures in the study, we piloted each to determine validity. We also factor-analyzed and removed or edited poorly performing items. Prior to analysis, each outcome measure was converted to a Rasch person measure in Winsteps (Linacre, 2021). Responses from our student population were used to produce the psychometric properties. Students responded to the "interest in data," "value of science to society," and "science career motivation" subscales using a 6-point scale that included the options "strongly disagree," "disagree," "somewhat disagree," "somewhat agree," "agree," and "strongly agree."

Engagement. The engagement measure, which we only collected once at the end of the study, asked students to self-report how often they did things like contributing to class discussions or asking questions during class on a 5-point scale that included the options "not at all," "a few times this semester," "a few times a month," "a few times a week," and "nearly every day." These items originated from the prior work of Ahlfeldt et al. (2005) and were adapted for the context of the current study.

Interest in data. The interest in data instrument captures student situational interest (Linnenbrink-Garcia et al., 2010), specifically when working with data in the context of science class. This scale included seven items like, "I like collecting data to answer scientific questions" and "Interpreting data makes it easier to evaluate someone's claims."

Value of science to society. The value of science to society measure originated from the Modified Attitudes Towards Science Inventory (Weinburgh & Steele, 2000). Weinburgh and Steele (2000) reported Cronbach's alpha of 0.62 with a sample of 5th grade students. This subscale included five items like, "Most people should study some science" and "Science is of great importance to a country's development."

Science career motivation. We measured career motivation using the Science Motivation Questionnaire II (SMQ-II) (Glynn et al., 2011). This included six items like, "I will use science problem-solving skills as part of my job" and "Learning science will help me get a good job." In the original validation study with college students, the career component of the SMQ-II produced a Cronbach's alpha of 0.92.

Self-efficacy in analyzing and interpreting data. Our self-efficacy subscale captures the extent to which students are confident in their abilities to analyze and interpret data. We adapted this measure from the work of Bourdeau and Arnold (2009) and Ahlfeldt et al. (2005). Students responded to a 5-point rating scale to rate their level of confidence in doing things like, "Use data as evidence to support a claim I made" and "Work with complicated data sets that may have unclear patterns." The scale included the options "not at all confident," "not very confident," "somewhat confident," "confident," and "very confident."

Teacher implementation. Within randomized studies, student outcomes are influenced both by the intervention and how teachers adhere to the intended intervention (Stains & Vickrey, 2017). Therefore, we used a multidimensional approach

to measure the fidelity of implementation by different teachers (Nelson et al., 2012). Using implementation logs, we collected data on how instructors adhered to the dosage of eight intervention activities over the course of the semester, and how long they spent implementing each activity. We included questions asking (i) whether the materials were implemented in the context of a larger unit, (ii) how individual activities were used (i.e. small-group work, homework, whole-class lesson), (iii) how much time students spent engaged with each activity, and (iv) the depth of conversations surrounding QR and science practices.

Measuring use of scientific practices in treatment and comparison classroom. The two developers of Data Nuggets traveled to the study locations in California, Colorado, Michigan, and Illinois to conduct live observations at least two times per teacher during the implementation period. During the semester-long study, the team conducted a total of 106 observations. Of the 106 observations, 13% (14 observations) were conducted by both observers in order to calculate interrater reliability. Observations occurred in both treatment and comparison classrooms on the same day. Prior to conducting observations, observers completed iterative training to achieve inter-rater reliability. Rater agreement for overlapping observations during the study ranged from Kappa = 0.65 to 1.00, with an average of 0.84. During the observations, we measured the time that classes spent engaged in the eight practices of science, as defined by the NGSS (NRC, 2013). This measure was adapted from the Collaboratives for Excellence in Teacher Preparation Core Classroom Observation Protocol (Lawrenz et al., 2007). For each 5-min segment of the class period, an individual observer recorded whether students engaged in a practice of science by marking if a specific aspect of that particular practice occurred. We then calculated the proportion of time engaged in each practice during observations and then averaged the score of two observations within each treatment and comparison classroom. In the treatment classrooms, we also measured compliance and implementation quality. For an example of the observation protocol record sheet and the coding manual, see Supplemental Materials.

Measuring fidelity to the intervention. Finally, the observers rated the overall quality of Data Nuggets implementation in each treatment classroom. Observers scored each teacher independently and then came to consensus to generate the score. Implementation quality was measured on a 1 (low-level implementation) to 5 (high-level implementation) scale. We used several factors to determine the level of teacher implementation, including how well each teacher (i) made use of the discussion topics and teaching opportunities presented in the Data Nuggets teacher guides, (ii) made attempts to understand student thinking, (iii) engaged students in in-depth discussions of the content, (iv) addressed student misconceptions, and (v) made connections between the intervention activities and students' experience or made connections to the course content. In addition, we assessed the teacher's overall attitude toward the intervention activities.

Data analysis. We used generalized estimating equations (GEEs) (Zeger et al., 1988) to answer our research questions. GEEs are similar to hierarchical linear modeling (HLM) which is used when observations are nested within groups (Rabe-Hesketh & Skrondal, 2008). Compared to hierarchical linear models (HLM), GEEs better address mis-specification residual correlations when using robust standard

errors (Homish et al., 2010). All statistical models were conducted, and associated figures were created using STATA version 15 (StataCorp, 2017).

To what extent do students in classrooms using Data Nuggets show improved quantitative reasoning compared to students in business-as-usual comparison classrooms? We conducted a main effects test of the treatment variable for our primary outcome of QR and then exploratory tests for each of the affective measures controlling for baseline (pretest) in each statistical analysis. To test our main effects model (research question 1), we regressed the person measure for the outcome variable (QR) onto the associated baseline measure (PRE), treatment condition (TRT), which takes on a value of 1 for classrooms randomized to the treatment and a value of 0 for comparison classrooms. Teacher (TEACH) is a fixed-effect variable that controls for all time-invariant differences between teachers, and CORR represents the correlation between individuals within each classroom.

$$\text{OUTCOME} = \beta_0 + \beta_1(\text{PRE}) + \beta_2(\text{TRT}) + \beta_3(\text{TEACH1}) + \ldots \beta_{22}(\text{TEACH20}) + \text{CORR} + \text{Error}$$

To what extent do students in classrooms using Data Nuggets have higher interest, motivation, self-efficacy, and engagement in STEM compared to students in business-as-usual classrooms? We used the same statistical model for each of the exploratory analyses, but we changed the OUTCOME variable (Constructing Explanations, Interest in Data, Value of Science to Society, Science Career Motivation, Self-Efficacy in Analyzing and Interpreting Data) and the respective baseline measure (PRE). That is, each time we conducted an exploratory analysis, we used the baseline measure of the outcome as a covariate in our model. The only analysis that did not include a baseline measure was the Engagement measure, which was a post-treatment only measure of engagement in biology class.

Does the quality of teacher implementation of the intervention moderate student outcomes? To investigate this final exploratory question of whether the quality of implementation (QUALITY) impacted student outcomes within the treatment group, we again used the GEE model to regress QR in biology onto the quality of implementation (categorical variable ranging from 1 = low to 5 = high), while controlling for baseline QR. CORR represents the correlation between individuals within each classroom.

$$\begin{aligned}\text{QR Outcome} = {} & \beta_0 + \beta_1(\text{QRPRE}) + \beta_2(\text{TRT}) + \beta_3(\text{QUALITY1}) + \beta_3(\text{QUALITY2}) \\ & + \beta_4(\text{QUALITY3}) + \beta_5(\text{QUALITY4}) + \text{CORR} + \text{Error}\end{aligned}$$

Results

Baseline Equivalence To evaluate if there were significant differences between the treatment and comparison groups on the outcome measures prior to the start of the study, we calculated baseline equivalence effect sizes (Table 1). Equivalence of groups at baseline contributes to our confidence that findings of the study are not the result of a pre-study difference between the groups. We used the number of students in each condition, the associated mean score, and standard deviations from

Table 1 Mean scores by treatment on pre- and post-outcome measures

Scaled person measure outcome in logits	Timepoint	Comparison mean (SD)	Treatment mean (SD)	Effect size (d)	95% Confidence interval of the effect size	
					Lower	Upper
Quantitative reasoning	Pre	50.43 (6.88)	50.46 (7.47)	0.00	-0.12	0.13
	Post	52.37 (6.73)	52.04 (7.68)	-0.04	-0.18	0.09
Constructing explanations (not scaled)	Pre	1.91 (0.75)	1.97 (0.84)	0.08	-0.05	0.21
	Post	1.97 (0.74)	2.15 (0.93)	0.22	0.07	0.36
Interest in data	Pre	52.23 (11.87)	52.34 (10.75)	-0.03	-0.17	0.11
	Post	52.42 (11.82)	52.88 (11.34)	0.05	-0.10	0.20
Value of science in society	Pre	58.37 (14.49)	57.57 (14.34)	-0.05	-0.19	0.08
	Post	58.67 (13.63)	57.92 (14.94)	0.04	-0.11	0.19
Science career motivation	Pre	52.93 (19.92)	51.90 (19.87)	-0.05	-0.18	0.08
	Post	53.65 (17.88)	54.50 (20.10)	0.06	-0.08	0.21
Self-efficacy in analyzing and interpreting data	Pre	53.77 (14.67)	54.10 (14.80)	0.02	-0.12	0.16
	Post	57.68 (15.17)	59.54 (17.40)	0.11	-0.03	0.26
Engagement	Post	49.64 (13.58)	49.77 (15.96)	0.01	-0.15	0.14

All person measures scaled 0–100 with the exception of Constructing Explanations

the two groups (treatment and comparison) on each outcome measure (Table 1). We found no significant differences and very small effect sizes between the treatment and comparison groups at the start of the intervention on the QR measure (Cohen's $d = 0.00$). Similarly, we investigated the baseline equivalence for each of the exploratory outcome measures. These effect sizes ranged from −0.05 (value and career), 0.02 (self-efficacy), and −0.03 for interest in data. Students in the treatment group scored slightly higher at baseline on the sub-score for the explanation task ($d = 0.08$) than the comparison group (Table 1).

Research Question 1 *To what extent do students in classrooms using Data Nuggets show improved quantitative reasoning compared to students in business-as-usual comparison classrooms?* We predicted that students using the intervention activities would experience improved QR abilities including graphing, graph interpretation, and ability to construct explanations using the claim-evidence-reasoning framework (C-E-R). The main effect of treatment on QR in the context of biology after controlling for the pretest was not significant (Table 1 and Fig. 2a; $B = -0.22$, SE = 0.31, $p = 0.48$, Cohen's d effect size = −0.04). However, in an exploratory analysis, we examined performance on the various components of the assessment. For the component that required students to develop a scientific explanation, we found a significant difference between treatment and comparison classrooms favoring the treatment (Table 1 and Fig. 2b; $B = 0.15$, SE = 0.06, $p = 0.01$, Cohen's d effect size = 0.22).

Research Question 2 *To what extent do students in classrooms using Data Nuggets have higher interest, motivation, self-efficacy, and engagement in STEM compared to students in business-as-usual classrooms?* We predicted that students using the intervention activities would show improvement in our affective constructs. We found no significant difference between treatment and comparison students' interest in data (Table 1 and Fig. 2c; $B = 0.63$, SE = 0.48, $p = 0.20$, Cohen's d effect size = 0.05) or in their appraisals of the value of science (Table 1 and Fig. 2d; $B = 0.29$, SE = 0.52, $p = 0.58$, Cohen's d effect size = 0.04). However, students in the treatment condition did report significantly higher self-efficacy regarding their ability to analyze and interpret data compared to students in the comparison condition (Table 1 and Fig. 2e; $B = 1.60$, SE = 0.67, $p = 0.02$, Cohen's d effect size = 0.11). Relative to the comparison condition, treatment students also reported significantly greater motivation to pursue a science career (Table 1 and Fig. 2f; $B = 1.71$, SE = 0.70, $p = 0.02$, Cohen's d effect size = 0.06). Finally, there was no difference in student engagement between the treatment and comparison classrooms (Table 1; $B = -0.15$, SE = 0.51, $p = 0.76$, Cohen's d effect size = 0.01).

Research Question 3 *Does the quality of teacher implementation of the intervention moderate student outcomes?* Finally, we investigated the extent to which level of implementation may have impacted student outcomes (Fig. 3). We detected a significant difference ($t = 9.89$, $p = 0.001$) in the extent of time spent on the practices of science, with treatment classrooms (M = 1.39, SD = 0.56) engaged in the practices of science to a greater extent than comparison classrooms (M = 0.46, SD = 0.40).

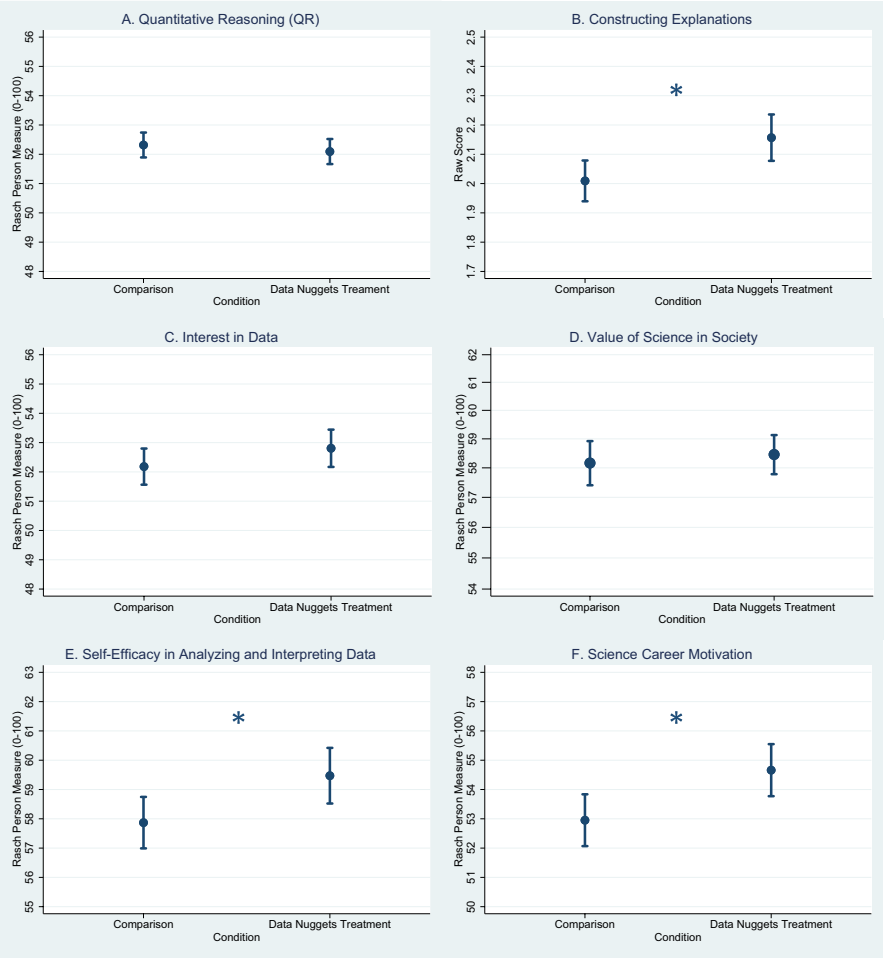

Fig. 2 **a**–**f** Main effect of treatment on student **a** quantitative reasoning (QR), **b** constructing explanations, **c** interest in data, **d** value of science in society, **e** self-efficacy in analyzing and interpreting data, and **f** science career motivation, controlling for pretest. Means marked with * are significantly different ($p \leq 0.05$)

We also explored whether students within the treatment condition have higher scores on the QR assessment when teachers implemented with higher fidelity to the intervention. Comparing the five levels of the implementation quality measure, after controlling for pretest (baseline) score, we found that students in the highest-level implementation classrooms (level 5) outperformed those in the lowest-level classrooms (level 1) on the overall QR assessment ($\beta = 4.02$, $p = 0.03$; Fig. 3).

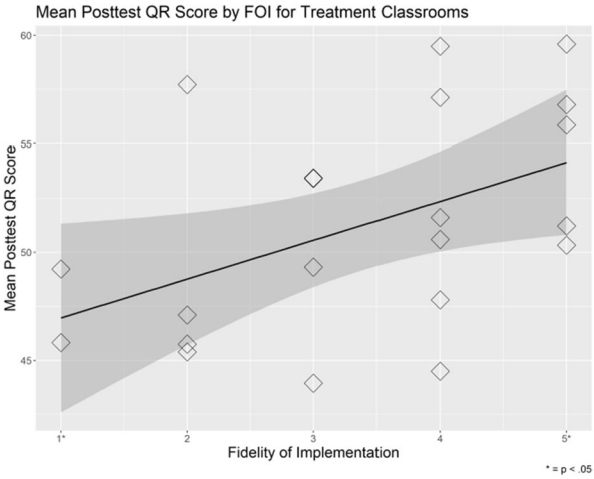

Fig. 3 Fidelity of Implementation to the treatment as a predictor of student post-test score on the quantitative reasoning (QR) measure. Each diamond represents one treatment classroom (note that a darker diamond denotes 2 teachers at the same location). The solid line is the regression line and the shaded area around the line is the 95% confidence interval

Discussion

Quantitative Reasoning Over the course of the study, students across both conditions improved in their QR abilities (Table 1). These results indicate no difference in overall QR attributable to the intervention. However, we did detect several meaningful differences on components within QR, specifically in students' ability to construct scientific explanations.

Data Nuggets improve student abilities to construct scientific explanations. We found that students using Data Nuggets were better able to construct scientific explanations using the claim-evidence-reasoning framework (C-E-R) (McNeill & Krajcik, 2008b; McNeill et al., 2006; Fig. 2b). Students were more likely to select the correct claim in response to a scientific question, as well as use data as evidence to support their claims and use appropriate reasoning to back up their claims. This suggests that the activities helped students with the challenging tasks of extracting meaning from datasets and visual information to construct scientific explanations.

We designed Data Nuggets to give students repeated practice in constructing explanations, with faded scaffolding to break apart the components of the task (McNeill et al., 2006), likely contributing to these results. Data Nuggets also provide ample context around a dataset, connecting students to the subject area as well as the motivations and rationale for data collection. Research has found that the ability to interpret graphs is contextually dependent, and an understanding of the content leads to greater abilities in this area (Roth & Bowen, 2001). Therefore, scaffolding and the context provided in the activities may be important in helping students to create appropriate visual displays of data and then subsequently use data as evidence.

Even with improvements in C-E-R seen in our study population, fewer than 25% of all students were able to identify the correct claim at the end of our study. This may have been due to a limitation of the QR instrument. We asked students to

make claims about statistical significance, including interpretation of error bars, a statistical competency more advanced than what most Data Nuggets activities and high school biology instruction requires. In addition, we asked students to interpret fictional vignettes and fake datasets, which contrasts to the authentic research and data in Data Nuggets. These differences may have influenced students, particularly in treatment classrooms, where the authenticity of science stories had been emphasized.

On the post-test, 65% of students identified at least some evidence to support their claim, regardless of whether the claim they chose was correct. However, very few students in either the treatment or the comparison successfully identified appropriate reasoning (6.6% in comparison, 9.4% in treatment). This is consistent with prior findings that students struggle most with the reasoning component of C-E-R, in which they are expected to discuss relationships and connect claims to both evidence and scientific concepts (McNeill & Krajcik, 2008b; Osborne et al., 2016).

Interest, Motivation, Self-Efficacy, and Engagement in STEM Data Nuggets increased student science career motivation. Students who used the intervention activities demonstrated increased interest in pursuing scientific careers compared to students in the business-as-usual classes (Table 1, Fig. 2f). We hypothesize this effect is due to the presence of science stories and scientist role models within each activity.

Each Data Nuggets activity is written by the scientist behind the data. Through images and storytelling, scientists share the motivation behind their work, how their ideas originated, and any challenges they faced along the way. Storytelling has been shown to be a powerful tool in science education (Collins, 2021; Estrada et al., 2011; Gibson, 2004; Schinske et al., 2016). Developers provide extensive revisions and encourage Data Nuggets authors to use metaphors and storytelling techniques throughout their writing, and to take opportunities to connect with student experiences. The activities include engaging photos of scientists in the field or laboratory. In addition, they incorporate humanizing elements, such as hobbies and personal stories about why a scientist decided to pursue STEM as a career.

Typical science classroom instruction does not represent the diversity found within STEM (Becker & Nilsson, 2021; Damschen et al., 2005). In contrast, students using Data Nuggets are introduced to a broader representation of scientist role models. These scientist stories and accompanying images may be providing students with a new perspective on who scientists are and what the jobs of scientists look like (Gladstone & Cimpian, 2021). This is consistent with teacher feedback. Teachers involved in the study shared that Data Nuggets changed their students' perspectives about who might be a scientist, and that after the study, students felt more connected with the idea of being a scientist. Furthermore, the addition of activities written by younger scientists offers a fresh perspective on academic training and early career excitement that students may not be exposed to otherwise. One teacher shared that their students do not often get to see scientists who are women, and the fact that so many young female scientists were represented in the program's materials was something that stood out to them.

Data Nuggets increase student self-efficacy in analyzing and interpreting data. We found that students who used Data Nuggets showed greater gains in self-efficacy, specifically for tasks related to the interpretation, analysis, and use of data in the context of science (Table 1, Fig. 2e). Self-efficacy is an important construct associated with science achievement (Bandura, 1997; Britner & Pajares, 2006). We would expect that the more often a student performs a task, such as engaging in science practices (NRC, 2012, 2013), the more their confidence would increase in their ability to successfully complete that task in the future. Because many activities were used across the course of the semester, it could be that repeated practice using a familiar format helped students feel more confident in their abilities, thus increasing their self-efficacy.

Science stories are useful when preparing data-rich materials for use in the classroom (Giamellaro et al., 2020). However, in traditional biology instruction, the steps of the scientific process are often removed. We hypothesize that it is exposure to these components in our activities that drives student confidence when engaging in science themselves, showing that the natural world is fascinating, often unpredictable, and that scientists are also continually learning how the world works (Gould et al., 2014; Kjelvik & Schultheis, 2019). At the conclusion of each activity, students are challenged to think like scientists and ask questions to build on the information gathered from a single study. Through this experience, students see that surprising results do not indicate that something has gone wrong in an investigation, but are instead part of the scientific enterprise and the journey to discover the unknown.

Teacher Implementation Influenced Student Outcomes We found that the quality of implementation varied across the teachers in our study, and that this in turn affected student outcomes. Strong implementers embraced the activities as teaching opportunities, highlighting exciting themes that came up within the research, guiding students to engage in active reading, addressing student misconceptions as they arose, and teaching QR concepts throughout. Strong implementers engaged their students in science practices and exploration, both of which have been demonstrated to help students better learn material and perform better on tests when compared to lecture and memorization (Freeman et al., 2014). Weak implementers used the intervention activities as worksheets that students completed alone, rather than as a teaching tool connected to instruction.

Differences in implementation explain some of the variation that we observed across treatment classrooms. Teachers with the highest-level implementation tended to have students that made the greatest gains in QR across the course of the study (Fig. 3). Students in classrooms with the lowest-level implementation had significantly lower QR scores than those in classrooms with the highest-level implementation ($\beta = 4.02$, $p = 0.03$). This is consistent with findings that implementation matters in active learning and engaging in science practices (McNeill & Krajcik, 2008a; Settles, 2009). A recent review found active learning helped narrow achievement gaps for students in underrepresented groups, but found

large variability across classrooms and that it only impacted students in classrooms when teachers implemented at the highest levels (Theobald et al., 2020).

When students engage in science practices, teachers are essential in helping make sense of this form of instruction, establishing norms (Driver et al., 1994) and structuring and guiding students' experiences (American Association for the Advancement of Science [AAAS], 1993; National Research Council [NRC], 1996). Teachers should serve as models of the behaviors of a scientist, for example, demonstrating how to work through complex scientific data (Crawford, 2000; Giamellaro et al., 2020). However, research shows teachers may struggle when assisting students with the practices of science (Marx et al., 1997), and few studies have looked at teacher practice surrounding the implementation of the practices of science in the classroom (McNeill & Krajcik, 2008a). Although we used our teacher PD workshop at the beginning of the study to familiarize our study teachers to Data Nuggets, the two days we had with teachers were likely insufficient to both introduce our study design and influence how teachers used the materials with their students. This resulted in our study reflecting a wide variety of implementation styles, and not eliciting results based on only the strongest implementation. This result did meet our desires for the study; we intended to capture what typical Data Nuggets instruction looks like, and most of our users do not attend our PD sessions and find materials online or through word-of-mouth.

As an implication for instruction, Data Nuggets should be used as a tool to increase the use of science practices and bring authentic data into science classrooms, but not as a worksheet for students to complete without any surrounding instruction. Within each activity there are plentiful learning opportunities, unique and genuine to the story of an individual research project. However, these only become apparent to students through intentional instruction.

Conclusion

Data Nuggets use authentic data and cutting-edge research to give students repeated practice in quantitative reasoning (QR) in the context of science. The purpose of these activities is to be embedded within typical instruction, supplementing existing programs by giving teachers the means to infuse real data and research throughout the curriculum. While they require relatively small commitments from teachers, they have been shown to positively influence students in several ways. The scientists who write these activities serve as role models for students, sharing their stories and demonstrating the practices of science. In turn, students spend more time engaging in the work of scientists, and come to see science as an active endeavor and something useful in their everyday lives, increasing their interest in STEM careers and confidence in their own ability to work with data. Because Data Nuggets can be used several times throughout the semester and across grade levels, they allow for repeated practice and scaffolding in important QR skills, resulting in an increase in student abilities to construct scientific explanations.

This study used Data Nuggets in their entirety for the intervention, and we can only now hypothesize which elements of the design drove the observed changes in students (see arrows in Fig. 1). Future research will explore the mechanisms behind

these results and determine which elements of the program materials are responsible for student gains. For example, based on these results we hypothesize that the scientist role models contribute to students' increased self-efficacy and interest in pursuing STEM careers. Our next step is to manipulate the degree to which students experience scientist role models in our activities and to increase representation in our materials even further.

Conflict of Interest

The authors declare no competing interests.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10763-022-10295-1.

Acknowledgements Thanks to Brian Donovan, Kjelvik and Schultheis, Christopher Wilson, May Lee, Alexa Warwick, Monica Weindling, Alex Duncan, Paul Strode, Audrey Mohan, Zoë Buck Bracey, Kristin Bass, and the participating teachers for their contributions to the research study. Thank you to the anonymous reviewers for their thoughtful comments. This material is based upon work supported by the National Science Foundation under DRK-12 grant numbers 1503211 and 1503005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additional funding from Kellogg Biological Station (KBS) Long-Term Ecological Research program (NSF DEB 1832042) and NSF IUSE 2012014. This is KBS Contribution #2294.

Funding This research was completed with funding from NSF DRK-12 1503211 and 1503005.

References

Ahlfeldt, S., Mehta, S., & Sellnow, T. (2005). Measurement and analysis of student engagement in university classes where varying levels of PBL methods of instruction are in use. *Higher Education Research & Development, 24*(1), 5–20. https://doi.org/10.1080/0729436052000318541

Aikens, M. L., & Dolan, E. L. (2014). Teaching quantitative biology: Goals, assessments, and resources. *Molecular Biology of the Cell, 25*(22), 3478–3481. https://doi.org/10.1091/mbc.e14-06-1045

American Association for the Advancement of Science [AAAS]. (1993). *Benchmarks for science literacy*. Oxford University Press.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Freeman.

Bandura, A., & Locke, E. A. (2003). Negative self-efficacy and goal effects revisited. *Journal of Applied Psychology, 88*(1), 87–99. https://doi.org/10.1037/0021-9010.88.1.87

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Pre-K–12 report II*. Retrieved January 15, 2021 from https://www.amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx

Becker, M. L., & Nilsson, M. R. (2021). College chemistry textbooks fail on gender representation. *Journal of Chemical Education, 98*(4), 1146–1151. https://doi.org/10.1021/acs.jchemed.0c01037

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human Sciences* (2nd ed.). Routledge.

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education, 90*(2), 253–269. https://doi.org/10.1002/sce.20106

Bourdeau, V. D., & Arnold, M. E. (2009). *The science process skills inventory*. 4-H Youth Development Education, Oregon State University.

Britner, S. L., & Pajares, F. (2006). Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching, 43*(5), 485–499. https://doi.org/10.1002/tea.20131

Britton, B. K. (1983). What makes stories interesting. *Behavioral and Brain Sciences, 6*(4), 596–597.

Capraro, M. M., Caparo, R. M., & Jones, M. (2014). Numeracy and algebra: A path to full participation in community and society? *Reading Psychology, 35*(5), 422–436. https://doi.org/10.1080/02702711.2012.739263

Center for the Advancement of Informal Science Education. (2018). Broadening participation task force: February 2018 update. Retrieved January 15, 2021 from http://www.informalscience.org/news-views/broadening-participation-task-force-february-2018-update

Chemers, M. M., Zurbriggen, E. L., Syed, M., Goza, B. K., & Bearman, S. (2011). The role of efficacy and identity in science career commitment among underrepresented minority students. *Journal of Social Issues, 67*(3), 469–491. https://doi.org/10.1111/j.1540-4560.2011.01710.x

Collins, S. N. (2021). The importance of storytelling in chemical education. *Nature Chemistry, 13*(1), 1–2. https://doi.org/10.1038/s41557-020-00617-7

Crawford, B. A. (2000). Embracing the essence of inquiry: New roles for science teachers. *Journal of Research in Science Teaching, 37*(9), 916–937. https://doi.org/10.1002/1098-2736(200011)37:9%3c916::AID-TEA4%3e3.0.CO;2-2

Damschen, E. I., Rosenfeld, K. M., Wyer, M., Murphy-Medely, D., Wentworth, T. R., & Haddad, N. M. (2005). Visibility matters: Increasing knowledge of women's contributions to ecology. *Frontiers in Ecology and the Environment, 3*(4), 212–219. https://doi.org/10.1890/1540-9295(2005)003[0212:VMIKOW]2.0.CO;2

Driver, R., Asoko, H., Leach, J., Scott, P., & Mortimer, E. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher, 23*(7), 5–12. https://doi.org/10.3102/0013189X023007005

Estrada, M., Woodcock, A., Hernandez, P. R., & Schultz, P. W. (2011). Toward a model of social influence that explains minority student integration into the scientific community. *Journal of Educational Psychology, 103*(1), 206–222. https://doi.org/10.1037/a0020743

Franz-Odendaal, T. A., Blotnicky, K., French, F., & Joy, P. (2016). Experiences and perceptions of STEM subjects, careers, and engagement in STEM activities among middle school students in the Maritime Provinces. *Canadian Journal of Science, Mathematics and Technology Education, 16*, 153–168. https://doi.org/10.1080/14926156.2016.1166291

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Gentner, D. R. (1976). The structure and recall of narrative prose. *Journal of Verbal Learning and Verbal Behavior, 15*, 411–418.

Giamellaro, M., O'Connell, K., & Knapp, M. (2020). Teachers as participant-narrators in authentic data stories. *International Journal of Science Education, 42*(3), 406–425. https://doi.org/10.1080/09500693.2020.1714093

Gibson, D. E. (2004). Role models in career development: New directions for theory and research. *Journal of Vocational Behavior, 65*(1), 134–156. https://doi.org/10.1016/S0001-8791(03)00051-4

Gladstone, J. R., & Cimpian, A. (2021). Which role models are effective for which students? A systematic review and four recommendations for maximizing the effectiveness of role models in STEM. *International Journal of STEM Education, 8*(1), 59. https://doi.org/10.1186/s40594-021-00315-x

Glynn, S. M., Brickman, P., Armstrong, N., & Taasoobshirazi, G. (2011). Science motivation questionnaire II: Validation with science majors and nonscience majors. *Journal of Research in Science Teaching, 48*(10), 1159–1176. https://doi.org/10.1002/tea.20442

Gould, R., Sunbury, S., & Dussault, M. (2014). In praise of messy data: Lessons from the search for alien worlds. *The Science Teacher, 81*(8), 31–36.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371–395. https://doi.org/10.1037/0033-295X.101.3.371

Hammett, A., & Dorsey, C. (2020). Messy data, real science: Exploring harmful algal blooms with real-world data. *The Science Teacher, 87*(8), 40–49.

Hiebert, J., & Stigler, J. W. (2000). A proposal for improving classroom teaching: Lessons from the TIMSS video study. *The Elementary School Journal, 101*(1), 3–20. https://doi.org/10.1086/499656

Hoffmann, R. (2014). The tensions of scientific storytelling. *American Scientist, 102*(4), 250–253. https://doi.org/10.1511/2014.109.250

Holmes, N. G., Wieman, C. E., & Bonn, D. A. (2015). Teaching critical thinking. *Proceedings of the National Academy of Sciences, 112*(36), 11199–11204. https://doi.org/10.1073/pnas.1505329112

Homish, G. G., Edwards, E. P., Eiden, R. D., & Leonard, K. E. (2010). Analyzing family data: A GEE approach for substance use researchers. *Addictive Behaviors, 35*(6), 558–563. https://doi.org/10.1016/j.addbeh.2010.01.002

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science, 326*(5958), 1410–1412. https://doi.org/10.1126/science.1177067

Jin, H., Yan, D., Mehl, C. E., Llort, K., & Cui, W. (2020). An empirically grounded framework that evaluates argument quality in scientific and social contexts. *International Journal of Science and Mathematics Education, 19*(4), 681–700. https://doi.org/10.1007/s10763-020-10075-9

Karaali, G., Hernandez, E. H. V., & Taylor, J. A. (2016). What's in a name? A critical review of definitions of quantitative literacy, numeracy, and quantitative reasoning. *Numeracy: Advancing Education in Quantitative Literacy*, *9*(1), Article 2. https://doi.org/10.5038/1936-4660.9.1.2

Kjelvik, M. K., Schultheis, E. H., & Gardner, S. (2019). Getting Messy with Authentic Data: Exploring the Potential of Using Data from Scientific Research to Support Student Data Literacy. *CBE—Life Sciences Education, 18*(2), es2. https://doi.org/10.1187/cbe.18-02-0023

Lawrenz, F., Huffman, D., & Gravely, A. (2007). Impact of the collaboratives for excellence in teacher preparation program. *Journal of Research in Science Teaching, 44*(9), 1348–1369. https://doi.org/10.1002/tea.20207

Lent, R. W., Sheu, H. B., Miller, M. J., Cusick, M. E., Penn, L. T., & Truong, N. N. (2018). Predictors of science, technology, engineering, and mathematics choice options: A meta-analytic path analysis of the social-cognitive choice model by gender and race/ethnicity. *Journal of Counseling Psychology, 65*(1), 17–35. https://doi.org/10.1037/cou0000243

Linacre, J. M. (2021). *Winsteps® Rasch measurement computer program (version 5.1.1)*. Winsteps.com.

Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement, 70*(4), 647–671. https://doi.org/10.1177/0013164409355699

Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., & Soloway, E. (1997). Enacting project-based science. *The Elementary School Journal, 97*(4), 341–358. https://doi.org/10.1086/461781

Mayes, R., Forrester, J., Schuttlefield Christus, J., Peterson, F., & Walker, R. (2014). Quantitative reasoning learning progression: The matrix. *Numeracy: Advancing Education in Quantitative Literacy, 7*(2), Article 5. https://doi.org/10.5038/1936-4660.7.2.5

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

McNeill, K. L., & Krajcik, J. S. (2008a). Scientific explanations: Characterizing and evaluating the effects of teachers' instructional practices on student learning. *Journal of Research in Science Teaching, 45*(1), 53–78. https://doi.org/10.1002/tea.20201

McNeill, K. L., & Krajcik, J. S. (2008b). Assessing middle school students' content knowledge and reasoning through written scientific explanations. In J. Coffey, R. Douglas, & C. Stearns (Eds.), *Assessing science learning: Perspectives from research and practice* (pp. 101–116). NSTA Press.

McNeill, K. L., Lizotte, D. J., Krajcik, J. S., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *The Journal of the Learning Sciences, 15*(2), 153–191. https://doi.org/10.1207/s15327809jls1502_1

Murcia, K., Pepper, C., & Williams, J. (2020). Youth STEM career choices: What's influencing secondary students' decision making. *Issues in Educational Research, 30*(2), 593–611.

National Research Council [NRC]. (1996). *National science education standards*. National Academies Press. https://doi.org/10.17226/4962

National Research Council [NRC]. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press. https://doi.org/10.17226/13165

National Research Council [NRC]. (2013). *Next Generation Science Standards: For states, by states*. National Academies Press. https://doi.org/10.17226/18290

National Research Council [NRC]. (2014). *STEM integration in K-12 education: Status, prospects, and an agenda for research*. National Academies Press. https://doi.org/10.17226/18612

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The Journal of Behavioral Health Services & Research, 39*(4), 374–396. https://doi.org/10.1007/s11414-012-9295-x

Neumann, M. M., Hood, M., Ford, R. M., & Neumann, D. L. (2013). Letter and numeral identification: Their relationship with early literacy and numeracy skills. *European Early Childhood Education Research Journal, 21*(4), 489–501. https://doi.org/10.1080/1350293X.2013.845438

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994–1020. https://doi.org/10.1002/tea.20035

Osborne, J. F., Henderson, B., MacPherson, A., Szu, E., Wild, A., & Shi-Ying, Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching, 53*(6), 821–846. https://doi.org/10.1002/tea.21316

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. STATA press.

Roth, W.-M., & Bowen, G. (2001). Professionals read graphs: A semiotic analysis. *Journal for Research in Mathematics Education, 32*(2), 159–194. https://doi.org/10.2307/749672

Schinske, J. N., Perkins, H., Snyder, A., & Wyer, M. (2016). Scientist spotlight homework assignments shift students' stereotypes of scientists and enhance science identity in a diverse introductory science class. *CBE—Life Sciences Education, 15*(3), ar47. https://doi.org/10.1187/cbe.16-01-0002

Schochet, P. Z. (2008). *Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations* (NCEE 2008–4018). National Center for Education Evaluation and Regional Assistance.

Schultheis, E. H., & Kjelvik, M. K. (2015). Data Nuggets. *The American Biology Teacher, 77*(1), 19–29. https://doi.org/10.1525/abt.2015.77.1.4

Schultheis, E. H., & Kjelvik, M. K. (2020). Using Messy, Authentic Data to Promote Data Literacy & Reveal the Nature of Science. *The American Biology Teacher, 82*(7), 439–446. https://doi.org/10.1525/abt.2020.82.7.439

Settles, B. (2009). *Active learning literature survey*. Department of Computer Sciences, University of Wisconsin-Madison.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton.

Shernoff, D., Knauth, S., & Makris, E. (2000). The quality of classroom experiences. In M. Csikszentmihalyi & B. Schneider (Eds.), *Becoming adult: How teenagers prepare for the world of work* (pp. 141–164). Basic Books.

Simpson, A., & Bouhafa, Y. (2020). Youths' and adults' identity in STEM: A systematic literature review. *Journal for STEM Education Research, 3*(2), 1–28.

Šorgo, A., Verčkovnik, T., & Kocijančič, S. (2010). Information and communication technologies (ICT) in biology teaching in Slovenian secondary schools. *Eurasia Journal of Mathematics, Science and Technology Education, 6*(1), 37–46. https://doi.org/10.12973/ejmste/75225

Stains, M., & Vickrey, T. (2017). Fidelity of implementation: An overlooked yet critical construct to establish effectiveness of evidence-based instructional practices. *CBE—Life Sciences Education, 16*(1), rm1. https://doi.org/10.1187/cbe.16-03-0113

StataCorp. (2017). *Stata Statistical Software: Release 15*. StataCorp LLC.

Stuhlsatz, M., Snowden, J., & Donovan, B. (2020). *Quantitative reasoning in high school biology assessment* [Unpublished manuscript]. Colorado, Springs: BSCS Science Learning

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., Chambwe, N., Cintrón, D. L., Cooper, J. D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Iranon, N., Jones, L., II, Jordt, H., Keller, M., Lacey, M. E., Littlefield, C. E., ... Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences, 117*(12), 6476–6483. https://doi.org/10.1073/pnas.1916903117

Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.

Weinburgh, M. H., & Steele, D. (2000). The modified attitudes toward science inventory: Developing an instrument to be used with fifth grade urban students. *Journal of Women and Minorities in Science and Engineering*, *6*(1), 87–94. https://doi.org/10.1615/JWomenMinorScienEng.v6.i1.50

Wilkerson, M. H., Lanouette, K., & Shareff, R. L. (2021). Exploring variability during data preparation: A way to connect data, chance, and context when working with complex public datasets. *Mathematical Thinking and Learning,* 1–19. https://doi.org/10.1080/10986065.2021.1922838

Willingham, D. T. (2004). Ask the cognitive scientist: The privileged status of story. *American Educator.* Retrieved January 15, 2021 from https://www.aft.org/periodical/american-educator/summer-2004/ask-cognitive-scientist

Wilson, E. O. (2002). The power of story. *American Educator*, *26*(1), 8–11.

Wise, A. F. (2020). Educating data scientists and data literate citizens for a new generation of data. *Journal of the Learning Sciences, 29*(1), 165–181. https://doi.org/10.1080/10508406.2019.1705678

Yair, G. (2000). Educational battlefields in America: The tug-of-war over students' engagement with instruction. *Sociology of Education, 73*(4), 247–269. https://doi.org/10.2307/2673233

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics, 44*(4), 1049–1060. https://doi.org/10.2307/2531734