Authentic science experiences with STEM datasets: post-secondary results and potential gender influences

Andria C. Schwortz and Andrea C. Burrows

aSecondary Education, University of Wyoming, Laramie, WY, USA; bPhysics & Astronomy, University of Wyoming / Natural Sciences, Quinsigamond Community College, Worcester, MA, USA

ABSTRACT

Background: Dataset skills are used in STEM fields from astronomy to zoology. Few fields explicitly teach students the skills to analyze datasets, and yet the increasing push for authentic science implies these skills should be taught.

Purpose: The overarching motivation of this work is to understand authentic science learning of STEM dataset skills within an astronomy context. Specifically, when participants work with a 200-entry Google Sheets dataset of astronomical data, what are they learning, how are they learning it, and who is doing the learning?

Sample: The authors studied a total of 82 post-secondary participants, including a matched set of 54 pre/post-test (34 males, 18 females), 26 video recorded (22 males, 2 females), and 3 interviewed (2 males, 1 female) participants.

Design and methods: In this mixed-methods study, participants explored a three-phase dataset activity and were given an eight-question multiple-choice pre/post-test covering skills of analyzing datasets and astronomy content, with the cognitive load of questions spanning from recognition of terms through synthesizing multiple ideas. Pre/post-test scores were compared and ANOVA performed for subsamples by gender. Select examples of qualitative data are shown, including written answers to questions, video recordings, and interviews.

Results: This project expands existing literature on authentic science experiences into the domain of dataset education in astronomy. Participants exhibited learning in both recall and synthesis questions. Females exhibited lower levels of learning than males which could be connected to gender influence. Conversations of both males and females included gendered topics.

Conclusions: Implications of the study include a stronger dataset focus in post-secondary STEM education, and the need for further investigation into how instructors can ameliorate the challenges faced by female post-secondary students.

KEYWORDS

Dataset; post-secondary student learning; STEM education; authentic science experiences; gender

CONTACT Andria C. Schwortz aschwort@uwyo.edu Physics & Astronomy, University of Wyoming/Natural Sciences, Quinsigamond Community College, 670 W Boylston St, Box 224, Worcester, MA, 01606, USA

Supplemental data for this article can be accessed here.

© 2020 Informa UK Limited, trading as Taylor & Francis Group
Introduction and literature review

Processing large datasets is ubiquitous in science, mathematics, engineering, and technology (STEM) careers, and yet there is little formal teaching of necessary dataset skills at any level of STEM education. Not only are practicing scientists expected to know how to work with large amounts of data, but all people in modern societies face the use of datasets in their daily lives and careers (Johri and Olds 2014). Evidence-based education reform is crucial to improving education (e.g. Slavin 2020), and a focus on datasets could present pathways for post-secondary student access to software applications essential to science (NAS 2018a).

This work addresses the literature gap that while science education researchers understand conceptual learning, knowledge about effective instruction and acquisition of scientific skills lags behind (Scott, Asoko, and Leach 2007). While several models describe how learners acquire the knowledge and skills to transition from novices to experts, efforts need to be made to put these models into practice in teaching (Bransford, Brown, and Cocking 2000). The following literature review scans the topics of data education as a form of authentic science, the role of astronomy in STEM education, and the challenges faced by female STEM students.

Authentic science

Students need to experience realistic versions of science so they can make informed decisions about areas of study and careers (Spuck 2014). These experiences not only introduce students to the tools used by scientists but also can increase motivation in the sciences (Hellgren and Lindberg 2017). Authentic science is recapped by Burrows et al. (2016) as real-world science experiences that work to/towards a solution summarize information, use technology, analyze data, use findings for conclusions, develop questions, procedures, and methods, communicate the work, collaborate with others, and make results accessible to others.

In the United States (USA), where this study takes place, teachers of primary (ages 6–12 years) and secondary (ages 13–18 years) students are explicitly expected to utilize authentic experiences in both math classes (e.g. National Governors Association Center for Best Practices – Common Core State Standards, CCSS 2010) and science classes (e.g. Next Generation Science Standards – NGSS, NGSS Lead States 2013). The USA’s NGSS were based upon other countries’ successes (Achieve 2010a, 2010b). Additionally, much effort has been exerted into what other countries are implementing well in computer science, since today’s science requires processing of data using computational skills (Hubwieser et al. 2015). The NGSS highlights authentic science experiences, and this same trend is seen in informal science education as well, through summer camps or nature-based programs for children (e.g. Burrows et al. 2018). However, studies of authentic science experiences in primary and secondary school are limited in number, and post-secondary studies are not found. For example, Slavin et al. (2014) identified no papers published from 1990–2012 on authentic science experiences in primary or secondary education, and Cheung et al. (2017) identified only a single such paper. That paper, Harris et al. (2015), focuses on middle school rather than post-secondary, physical science and Earth science content rather than astronomy, and does not appear to address dataset learning.
The need for authentic science activities does not stop at the secondary level but continues through the post-secondary level (age 18+, also called collegiate or university). Research shows that post-secondary students benefit from domain-specific authentic science activities in many STEM and social science fields (e.g. environmental science: Carey and Gougis 2017; physics: Wilcox and Lewandowski 2016). Several papers describe ways of putting domain-specific authentic science into practice (e.g. astronomy: Kobulnicky and Dale 2016; psychology: Brothen 1984).

**Authentic science requires dataset skills**

In the modern age, advances in technology and data storage have led to the situation where researchers often find themselves with more data than they can analyze by hand. The data collected can span a large range of time and space, as well as many forms or modes of data (Cheruvelil and Soranno 2018). While relevant literature generally does not re-articulate nor define the term ‘dataset’ (e.g. Abello, Pardalos, and Resende 2002; Brunner et al. 2002; Leskovec, Rajaraman, and Ullman 2011), here the authors of this paper use ‘dataset’ to refer to any collection of data, often arranged in a spreadsheet. The use of computers allows not only for faster analysis of large datasets but also for the discovery of trends and patterns that may not be otherwise apparent. Data mining is this process of sifting through large amounts of data, finding patterns, making predictions, and explaining underlying properties of the dataset (Whitten, Frank, and Hall 2011).

Within the realm of astronomy, for example, automated surveys allow for the collection of large amounts of data, more than can be analyzed object-by-object (Brunner et al. 2002; National Research Council 2010). Astronomers therefore must analyze these datasets en masse by programming their own data analysis tools (e.g. Schneider et al. 2007; Shen et al. 2011). Assisting in this endeavor, any dataset created in the USA with government funding becomes public, allowing astronomers to address questions that the original researchers may not have considered.

Because using datasets is inevitable in STEM settings, dataset skills are crucial to undergraduate students in STEM fields from astronomy to zoology. Even at the primary and secondary level, US students are asked to demonstrate skills consistent with dataset use (CCSS, 2010; NGSS Lead States 2013). However, education researchers have yet to determine at what educational level dataset skill instruction occurs in STEM fields, and how this varies across different STEM disciplines.

**Astronomy education**

The authors of this study utilized astronomy as the content area to study the acquisition of dataset skills for a number of reasons, but mainly since data is often made public after a proprietary period. This practice allows everyone with an internet connection access to large datasets. Astronomy is a field that serves to capture the interest of the public, with even non-scientists being able to participate in research through amateur groups and citizen science projects (e.g. Bailey 2011; Hecker et al. 2018; Raddick et al. 2010; Schwamb et al. 2013). In the USA, post-secondary students are often required to complete courses outside their major or program of study. The basic math level required for introductory astronomy makes the course more accessible than physics for these non-majors (Bailey 2011).
A large fraction of astronomy education research is focused on individual concepts (Bailey 2011; Lelliott and Rollnick 2010; Türk and Kalkan 2018). Curriculum development remains another key aspect of the field; however, the use of authentic STEM experiences is rarely discussed in this context (Bailey 2011). When authentic STEM experiences are discussed, it is generally in the context of using computers to perform small-scale authentic scientific inquiries (e.g. Danaia, McKinnon, and Fitzgerald 2017). However, the use of computers to manipulate large datasets may allow students to learn different skills than those developed by working with individual problems (see section on dataset education).

Important to this study, research astronomy is increasingly moving away from a realm of studying individual images, spectra, or time-based information, and towards studying large data sets derived from computerized analysis of those raw data products (Brunner et al. 2002; National Research Council 2010). Data skills and techniques needed by future generations of astronomers include classification (grouping similar objects), identification of outliers, and data visualization (Brunner et al. 2002; National Research Council 2010). The astronomy context of this study is discussed further in Table 1.

**Dataset education in STEM as authentic science**

Scientific research in many STEM fields is in the midst of transitioning from the scale of addressing individual problems or situations, to manipulating large sets of data, due to technological advances in sensing equipment, automated data collection, and data storage (Brunner et al. 2002; Langen et al. 2014; National Research Council 2010; Resnick, Kastens, and Shipley 2018). Because of this revolution in scientific research, science education practice is increasingly including data education, and science education research is working to determine effective pedagogies (Anderson et al. 2007). Students of all levels demonstrate difficulty performing data analysis ranging from straightforward tasks like graphing (Berg and Boote 2017; Gültepe 2016; Jackson, Edwards, and Berger 1993), to more nuanced analysis of data (Wallace et al. 2000).

Using large datasets as part of authentic science assists students in experiencing actual science, and satisfies a sense of curiosity (Langen et al. 2014). Students’ sense of

| Table 1. Astronomy context of the study. |
|-----------------------------------------|
| Quasars are a type of galaxy where the central supermassive black hole (SMBH) is actively accreting dust and gas from a surrounding disk. Quasar glow brightly due to friction of material circling the SMBH, and thus quasars can be seen from extreme distances. Radio light from quasars can yield information about the presence or absence of jets, and previous research on quasars has found that approximately 10% of quasars possess radio jets (Shen et al. 2011). The redshift of light is used to determine three properties: the speed of objects away from the observer (due to the expansion of the Universe and the Doppler effect), the distance to the object, and also how long ago the light was emitted from the source (also called lookback time). The quasars in the source sample have a mean distance from Earth of approximately 10.5 billion light years, expressed as redshift. The distribution of quasars is important to understanding the properties of the Universe, as how tightly clustered they are has implications for their gravity and evolution over time. The location of each quasar (or any astronomical object) is recorded in spherical coordinates: two angular coordinates on the sky (right ascension or RA, and declination or Dec), plus distance. The dataset used in this study is a subset of the Sloan Digital Sky Survey Data (SDSS) Release 5 (Schneider et al. 2007) catalog, plus data from the Very Large Array (VLA) Faint Images of the Radio Sky at Twenty-cm (FIRST) radio survey catalog. While the full catalog contains more than 30,000 quasars with more than 100 descriptors, the subset used with the participants contained 200 quasars (cases) in rows, and five descriptors in columns: quasar “name” (an identification number containing the quasar’s coordinates), two angular coordinates, distance (given as redshift or z), and FIRST radio magnitude. |
self-efficacy in using authentic technological tools is improved through authentic STEM dataset activities (Carey and Gougis 2017). The use of authentic, professionally collected, datasets teaches students additional skills beyond when they collecting their own datasets (Resnick, Kastens, and Shipley 2018). Further work is required in three aspects of dataset learning: the development of additional exercises and ‘best practices’ (Langen et al. 2014), the generalization of previous work (Carey and Gougis 2017), and determining behaviors and strategies associated with expertise (Resnick, Kastens, and Shipley 2018).

The use of datasets in a classroom setting is a form of authentic science, distinct from educational technology, computational thinking, and programming skills. To investigate the impact of an authentic STEM experience using astronomy datasets, this study’s authors designed an investigation where a STEM dataset was manipulated by participant groups to test their dataset knowledge. The goal was to gain insight into how post-secondary STEM learners progress in the acquisition of dataset expertise, while working in the content area of astronomy.

**Gender in STEM learning**

Since females’ and males’ results are displayed in the findings, a description of female learning in STEM can put some of those findings into context. The fact that females participate less in STEM as they advance in age, called the ‘leaky pipeline,’ has been well-studied (e.g. Hill, Corbett, and St. Rose 2010; Microsoft 2017). Causes for this are complex, with contributing factors including stereotype threat, low-self assessment, peer pressure, unconscious bias, lack of female role models, reduced access to preparatory courses or skills training, and stereotypes of gendered professions (Hill, Corbett, and St. Rose 2010; Microsoft 2017; Nissen and Shemwell 2016).

The large number of studies addressing gender in STEM learning (Brickhouse, Lowery, and Schultz 2000; Nieminen, Savinainen, and Viiri 2013; Nissen and Shemwell 2016; Nyhof-Young 2000; Nyström 2007) is indicative of the fact that no easy solution exists. When working with computers, some studies have shown tantalizing clues that could lead towards better practices. For example, in the field of computer education, group dynamics differ depending on the gender composition of the student groups (Busch 1996). In mixed-gender groups, males have been observed to control computers for a larger percentage of the time than females (Day et al. 2016).

Although females do not constitute a majority of participants in this study (see participant sample), they are the majority of college students at 57% (Snyder and Dillow 2011). In the field of astronomy specifically, attrition issues disproportionately affect female students from primary through post-secondary levels (Bergstrom and Sadler 2016). In order to generalize results from this study, it is important to examine the results of the female participants in relation to dataset use. Thus, the authors argue that authentic science, which is shown to work with students, can utilize computer science and datasets to open doors of STEM opportunities, but these opportunities could be limited if gender issues affect learning or access to those datasets.
Research questions

The purpose of this study was to investigate post-secondary students learning to manipulate datasets within astronomy, as an expansion of existing literature into authentic STEM learning. The following three-part question was developed to investigate their learning.

How does a three-phase, 1.5-h astronomy dataset intensive activity impact participants’ short-term learning scores by:

(1) Skills and content focused questions? (‘What are they learning?’)
(2) Recall and synthesis leveled questions? (‘How are they learning?’)
(3) By gender? (‘Who is learning?’)

Theoretical framework

This article uses the framework that learning is social and constructivist in nature (Eymur and Geban 2017; Kutnick et al. 2017; Vygotsky 1978), and that authentic STEM experiences lead to improved interest and learning in STEM. These ideals informed the design of the three-phase learning activity: participants in this study worked in groups, the activity used actual data from the Sloan Digital Sky Survey (see Electronic Supplemental Materials, ESM, and Table 1), the activity guided post-secondary students to create their own data interpretation, and the students were supported through steps of data analysis similar to those taken by astronomers.

For the quantitative analysis, the authors view the data through the numerical results of the participants’ work, while the qualitative data provide preliminary clues to how the participants constructed their knowledge.

Materials and methods

This was a mixed-methods study, examining participant responses to eight multiple-choice questions on pre/post-tests on astronomy content and dataset skills, and searching for supporting data among multiple qualitative sources. On the pre/post-tests, participants performed the three tasks of data manipulation and processing, data analysis, and data visualization (Hampton et al. 2017; Weintrop et al. 2015).

Participant sample

A total of 82 individuals participated in this study from 2014 through 2015. Participants were post-secondary students in introductory astronomy courses for non-majors at a large research university and a junior college in separate states (USA). Fifty-four participants have matched pre/post-tests, nine groups were audio/video recorded during the activity, and three participants were interviewed 1 week to 3 months after the activity (see Table 2).
**Experimental procedure**

The authors worked with subsets of the participants for 1.5 hours on two dates over the course of 2 weeks, with data collected from different groups of participants taking place from June 2014 through February 2015. On the first contact date, participants took the pre-test (see ESM), approximately a half-hour in time. On the second contact date, participants completed a three-phase activity (see ESM) and the post-test (same as the pre-test), which took an average of an hour.

Quantitative data reported in this study were obtained from eight multiple-choice questions on both astronomy content and data skills used as a pre/post-test. This instrument was reviewed by four individuals in astronomy and physics (a post-doctoral researcher, a lab technician, a post-graduate student, and an advanced post-secondary student), and the questions modified based on their feedback. As cognitive load is known to be an important aspect of student learning (Tekkumru-Kisa et al. 2019), the pre/post-test was organized by level of cognitive load to explore students’ levels of understanding and provide scaffolding to their learning (e.g. Lutsky 1986). The term ‘recall’ is used throughout this paper for low cognitive load questions involving the tasks of remembering and understanding, and ‘synthesis’ for evaluating and creating (Krathwohl 2010). For example, if a student showed evidence of having correctly memorized the distance to a specific astronomical object, this would be considered a high level of recall. On the other hand, if the student used logic to correctly reason about whether a certain type of astronomical object was likely to be found near the Earth or farther away, this would be a high level of synthesis.

Qualitative data used in this study were obtained from audio/video recordings of groups of participants, individual written responses on the activity handout, and one-on-one interviews conducted within 3 months after the completion of the activities. See Table 3 and the ESM for further details.
Table 3. Activity summary.

| Task Name         | Activity Description                                                                 | Example Participant Tasks                                                                 | Data Used in this Study                                                                 |
|-------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------|
| Pre-Test and Post-Test (same instrument given at the beginning and end of the session) | Eight multiple choice questions:  
  - Three recall  
    - remembering and understanding, e.g. definitions of terms  
  - Two application or analysis  
    - Application or analysis, e.g. using information in data tables  
  - Three synthesis  
    - evaluation and creation, e.g. using evidence to support conclusions  
    - three skills, four content, & one combo questions |  
  - Answer multiple-choice questions  
  - Answer open response questions with words or sketches  
  - Correct # of multiple-choice questions  
  - Correct # of recall vs. synthesis  
  - Correct # of skills vs. content questions |  

Introduction  
- Instructor-led  
- 10–15 minutes of background content: quasars and redshift  
- Presentation style similar to typical introductory astronomy lab.

Participant Activity  
- Self-selected groups of 3–4 using one computer.  
- Dataset composed of 200 rows of data in Google Sheets (Table 1).  
- Each activity phase was completed (in order) before another phase was given out.  
- Open-ended prompts to glean understanding.  
- Given increasing levels of instruction in the three-phase activity.  
- Scaffolding was intended to approximate the steps taken by expert astronomers.  
- Discussion  
  - Use sketches  
  - Count spreadsheet entries  
  - Highlight spreadsheet columns  
  - Sort spreadsheet columns  
  - Graph data  
  - Change axis titles  
  - Analyze graphs  

Phase I  
- Open-ended questions asking the participants to characterize the data, and speculate on how it could be analyzed.  
- Discussion  
  - Count spreadsheet entries  

Phase II  
- Slightly more specific questions, such as how the data could be graphed, and what it would look like.  
- Discussion  
  - Use sketches  
  - Control keyboard  

Phase III  
- Precise directions to create a scatter plot of the quasars’ positions in the sky and histograms of the quasars’ redshifts and radio magnitudes.  
- Detailed questions asking students to determine specific properties of the dataset from the graphs.  
- Discussion  
  - Highlight spreadsheet columns  
  - Graph data  
  - Change axis titles  
  - Analyze graphs  

Demographics Survey  
- Demographics information requested including race, nation of origin, gender, level of education, and field of study.  
- Answer questions  
  - Demographics and majors provided

Interviews  
- One-on-one interviews, conducted 1 week to 3 months after the activity. All participants were invited to interviews, three individuals agreed, and participated in interviews.  
- Discuss experience with spreadsheets  
  - Recordings and transcripts  
  - Identified themes

Analysis: quantitative data

Statistical measures included one-way ANOVA to determine the significance of any improvement from pre-test to post-test, pre/post-test means and standard deviation,
Table 4. Pre/post-test scores, standard deviations, gains, and effect size by gender, and by question type.

| N = 54 | Pre-/Post-Test Scores | Gains/Effect Size |
|--------|-----------------------|-------------------|
|        | Male                  | Female            | Total             | Male | Female | Total |
| Overall| 57 ± 19/78 ± 19*      | 58 ± 16/70 ± 22   | 58 ± 18/75 ± 20* | 0.467/1.12 | 0.084/0.62 | 0.285/0.86 |
| Skills | 63 ± 20/81 ± 25**     | 68 ± 22/69 ± 25   | 66 ± 21/*         | 0.515/0.79 | 0.088/0.06 | 0.326/0.44 |
| Content| 45 ± 24/73 ± 21**     | 48 ± 21/64 ± 28** | 47 ± 23/70 ± 24** | 0.438/1.26 | 0.163/0.70 | 0.321/0.99 |
| Recall | 57 ± 28/78 ± 29*      | 54 ± 26/81 ± 23*  | 57 ± 27/78 ± 28* | 0.412/0.76 | 0.481/1.16 | 0.407/0.80 |
| Synthesis | 33 ± 18/46 ± 22* | 31 ± 21/43 ± 25   | 33 ± 19/45 ± 23* | 0.324/0.64 | 0.250/0.49 | 0.287/0.56 |

*p < 0.01, **p < 0.05.
matched normalized gains (Hake 1998), and Cohen’s d as effect size (e.g. Burrows et al. 2016; Cohen 1977; Sullivan and Feinn 2012). It is worth noting that while Cohen’s d was originally intended to compare two independent groups, the same equation is accepted to describe the change of one group between pre- and post-tests (e.g. Nissen et al. 2018; Sullivan and Feinn 2012), so the authors chose to adopt the nomenclature of calling this formula ‘Cohen’s d’ and ‘effect size’, despite the different context of its usage. Further effect size explanation is provided by Lakens (2013). With these parameters in mind, the participants’ performance on questions, by level (recall vs. synthesis), was compared to their performance by skills or content. Results were compared for male and female participants.

Analysis: qualitative data

A preliminary coding for themes was performed on all qualitative data (Creswell 2013). For the video recorded groups, a count was made of the number of males and females in each group, whether male or female participants controlled the computer’s mouse and keyboard during the activity, and general functionality of the group.

Results

Through this dataset activity, participants improved their pre/post-test scores on an eight-question multiple-choice astronomy content and dataset skills test, as shown in Table 4. Pre-test means increased from 58% to a post-test mean of 75%, with \( p < 0.01 \) and a sample size of \( N = 54 \), with a gain of 0.285 and effect size of 0.86, indicating short-term learning did occur. The following information is organized by research question, with supporting qualitative data presented to triangulate the results.

Learning of skills vs. content

Data pointed towards simultaneous improvement in both dataset skills and astronomy content, with more improvement on content than on skills questions. Participants
showed improvement in both content and skills questions, with content gains of 0.326 or effect size of 0.44, and skills gains of 0.321 or effect size of 0.99 (see Table 4).

Table 5 is a cross-tabulation, showing how many participants improved their score, stayed the same, or did worse. Twenty-seven participants (50%) improved in both categories of skills and content, while 13 participants (24%) did worse on one or both of skills and content.

**Evidence of dataset skills**

The participants scored 66% on the pre-test skills questions and 76% on the post-test ($p < 0.05$). In the first phase of the activity, many students commented that few quasars are detected in radio light. Milo wrote the following, ‘Well obviously there are a lot of stars – Quasar’s here mostly – what stands out is the fact that almost none of them sent back brightness from radio light.’ Another student, Michelle, was able to recognize the arbitrary nature of the original data organization, ‘It is also apparent that the data is organized by the name of the quasar and not by any of its’ specific information.’ And in a one-on-one interview, Jessica reported previous experience with budgeting via spreadsheet programs.

Other participants’ responses indicated poor understanding of dataset skills. In response to how students might draw a graph of the data, Alyssa responded, ‘Below is a/Example of how I would organize the data into a chart,’ followed by a replication of the columns given in the dataset. Alyssa did not distinguish between visual representations of data such as graphs, with tabulated representations such as charts or spreadsheets.

**Evidence of astronomy content knowledge**

Participants scored 47% on the content questions on the pre-test, and 76% on the post-test ($p < 0.05$). Participants exhibited a lower level of content understanding than dataset skills – Michelle’s earlier comment did not show understanding that the quasar names were derived from the quasar’s coordinates on the sky. Alyssa also conflated radio light with sound (a common misconception), speculating that quasars were not detected in radio ‘because they are so far away you cannot hear them.’

Sabeena showed mixed levels of understanding between dataset skills and astronomy content. When prompted to discuss how she might visualize the data, Sabeena responded, ‘If I drew this data as a graph I would put the Quasar numbers as the x-axis and the redshift as the y-axis … in a scatter plot.’ This response showed knowledge of xy scatter plots, and thus is evidence of at least rudimentary dataset skills. However, the graph she suggested contained no scientific meaning, exposing a gap in scientific understanding.

**Learning of recall vs. synthesis**

Data pointed towards simultaneous improvement in both recall and synthesis level questions (see Table 4). Participants scored higher on recall questions on both pre- and post-tests (57% and 78%, respectively), than on the synthesis questions (33% and 45%, respectively). They exhibited more improvement on recall questions (gains of 0.407 and effect size of 0.80) than on synthesis (gains of 0.287 and effect size of 0.56). Table 5 shows that while the largest category of participants (17 individuals or 32%) exhibited
improvement on both recall and synthesis questions, 15 individuals or 28% did worse on one or both categories of recall or synthesis questions.

Evidence of recall-level understanding
Participants discussed recall-level ideas during the activity, such as definition of terms. When one member of a group possessed more recall knowledge, that individual was more likely to define the term for the other group members. Other groups were more likely to debate the meaning of terms and come to a consensus.

The following response from Milo on Phase II is a good example of recall-level understanding: ‘I would sort the data by Redshift’s the objects that are closer to earth then the object furthest.’ Milo demonstrated that he understood the definition of redshift as distance, but did not make any further connections.

Evidence of synthesis level understanding
During Jessica’s interview (mentioned earlier), she discussed the power of using spreadsheets to analyze ordinal data and speculated about the distinction between ordinal quantitative data and qualitative data: an evaluation-level concept. Further, Diego wrote the following in reply to a prompt on visualizing data:

With the scatter plot you can connect the dots and also see how far one dot is from another to figure out the distance of the certain quasars. It could also tell you the location in RA and DA [sic: Dec] showing you what area quasars tend to appear the sky. With a histogram chart you could also find detailed information if these quasars are moving and there pattern throughout history.

This response shows a synthesis level of understanding, where Diego makes the connection that coordinates on the sky can be used to determine the distance between quasars, a measurement relevant to determining whether quasars are caused by high-density environments (Shen et al. 2011). Diego also understood that a histogram of redshift can track not only the distance and speed of quasars but also their history due to lookback time (see Table 1).

Steve showed evidence of beginning to develop a synthesis level of understanding. When discussing the histogram of redshift, he wrote, ‘Maybe we don’t see so many in the higher range is because is the redshift is weaker the further away a quasar gets and were just not seeing quasars that may actually be there? I can’t fathom why there would be less quasars closer to us.’ Although inexpertly worded, Steve’s response indicates that he was attempting to use logic to reason through information about where astronomical objects are located, rather than using rote memorization. Steve reasoned that because a larger value of redshift

Table 6. ANOVA significance levels of gender on pre/post-test improvement overall, by skills/content questions, and by recall/synthesis question level.

|                  | Significance (p) |
|------------------|------------------|
| Overall          | 0.004**          |
| Skills           | 0.012*           |
| Content          | 0.074            |
| Recall           | 0.051            |
| Synthesis        | 0.796            |

*p < 0.05; **p < 0.01.
corresponds to a farther distance, the radio light could be too faint to detect. However, Steve did not understand how geometry affects the number of quasars observed at closer distances.

**Learning by gender**

ANOVA significance levels of gender on improvement from pre- to post-test are shown in Table 6. Gender revealed a statistically significant effect on the improvement of overall pre/post-test scores (p = 0.004), and of the dataset skills questions (p = 0.012).

Male participants had a large overall effect size and gains of 1.12 and 0.467, respectively, while those for females were 0.62 and 0.084 (see Table 4), indicating a higher level of short-term learning from male participants than female. In fact, T-test indicated female participants’ overall scores did not statistically significantly improve (p > 0.05).

**Evidence of gendered behaviors during activity**

Of the nine groups video recorded during the activity, participants self-selected into three mixed-gender groups, and six all-male groups. In the mixed-gender groups, males tended to control the computer for the majority of the time (with females only touching the mouse occasionally).

Conversations in mixed-gender groups were dominated by the males and included gendered comments, such as females discussing stereotypically gendered behaviors. Felicity, a female student, was in a group with two males. She referred to taking notes as ‘a female thing,’ implying the stereotype that females are better at secretarial tasks and worse at scientific ones. Contrary to Felicity’s self-assessment, she performed diverse actions consistent with authentic scientific research: recording data and ideas, following directions, keeping the group focused on their goals, and correcting partners’ errors.

One extreme example of gendered comments took place in a group of three males, Luke, Jack, and Shawn, who repeatedly made crude jokes. When Luke read aloud the prompt of ‘What would you do if your boss gave you this data set?’, Jack replied with obscene suggestions about this hypothetical boss, and innuendos about his wife. When Luke tried to steer the conversation back to work, Shawn supported Jack in his joking.

**Discussion**

Researchers realize that authentic STEM experiences are necessary for students to learn the skills of science, to develop a realistic picture of science as a field they may or may not wish to enter, and to motivate students in their STEM courses (Hellgren and Lindberg 2017; Spuck 2014). As research science increasingly moves towards the use of large datasets, educators must move towards teaching dataset skills during classes. Since astronomy often serves as a gateway to other sciences both for students and for the public (Bailey 2011), astronomy datasets could be vital for classroom and general use.

This section first discusses the results about the overarching theme of authentic science experiences, the research questions, and astronomy content vs. dataset skills. Additional discussion follows regarding the revealed gender issues, the limitations, and the greater implications of this study.
Research questions

The astronomy dataset activity demonstrates the feasibility of using the activity as a form of authentic STEM experience. As stated in the literature review, astronomy content can be used as a tool to teach dataset skills to students in an engaging and interesting manner (Bailey 2011). The post-secondary students utilized technology with a real dataset to create questions, procedures, methods, analyze patterns, and potentially propose solutions. They discussed their findings with their group and explained their reasoning and potential models. Thus, the dataset astronomy activity can be classified as an authentic science experience, albeit a short one.

Participants self-identified as having little prior knowledge in STEM datasets, and less experience with astronomy content. The data (such as pre-test scores) reflect the lack of experience, and post-secondary students most certainly enter the classroom with varying levels of prior knowledge. Both content and skills were gained by the participants, with many learning both simultaneously, providing a proof of concept that a single authentic STEM activity can address both aspects of STEM learning at least in the short term. The qualitative data confirmed that the level of knowledge about dataset skills and astronomy content was not constant among participants, but showed improvement in understanding of both content and dataset skills.

Participants exhibited learning at both the recall and synthesis levels on their pre/post-tests, with some individuals learning both simultaneously. Qualitative data showed that some participants exhibited traits of recall-level understanding, while others demonstrated a synthesis level. Importantly, and linking back to previously mentioned work, materials that contain multiple levels of content can provide differentiated opportunities and scaffolding that benefit students at all levels and abilities (Lutsky 1986; Westwood 2001).

The issue of learning by gender

The performances of the male and female participants were drastically different and this captured the authors’ attention. The ANOVA revealed that the lesser learning of the female participants was statistically significant. This finding cannot be attributed to prior knowledge as the females’ pre-test scores were comparable to that of the male participants. While the numerical data of this study do not allow the authors to speculate as to causes of difference by gender, the dynamics of female and male participants were particularly revealing and are discussed in the following section.

In this study, when female students worked in mixed-gender groups, they were rarely observed to control the computer mouse or keyboard, in agreement with the findings of Day et al. (2016). This study’s lack of female computer use in mixed-gender groups could be a factor in the females’ lower post-test scores and especially their dataset skills learning. The concept of (literally) ‘hands-on’ might be the difference in the learning between the groups.

Gendered insults from peers and self-deprecation may have been important factors, as these are classic examples of issues females face in STEM (Hill, Corbett, and St. Rose 2010; NAS, 2018b). For example, ‘behaviors that convey hostility, objectification, exclusion, or second-class status about members of one gender’ are forms of gendered harassment (NAS, 2018b). The inappropriate conversation of the male group fits this description and could have negatively impacted any female students who overheard the conversation.
But it is important to note that male students can be negatively impacted as well, as Luke’s learning was interrupted by continually having to steer his group away from gendered discussions and back to work.

The gendered self-deprecation of Felicity fits the known pattern of females in STEM possessing a lower self-assessment of their abilities than do males (Hill, Corbett, and St. Rose 2010; Microsoft 2017; Nissen and Shemwell 2016). Felicity’s self-assessment of her abilities as being merely secretarial ‘convey[ed] . . . second-class status about members of one gender’ (NAS, 2018b), which is another example of gendered harassment. Not only did Felicity’s self-assessment reflect inner challenges to her learning but expressing this aloud could have negative influences on other female students, and reinforced gender stereotypes for male students in her group.

Only 17% of the female post-secondary students in this study were exposed to women role models in the highest position of authority in their astronomy courses. The presence of visible female mentors in STEM is another factor known to improve female students’ outlook of STEM (Microsoft 2017), and the lack of such role models may have been a factor in these female participants’ struggles in this study. Investigations into female dataset use within authentic science are worthwhile future endeavors for researchers.

**Limitations**

This study highlighted only short-term participant learning and retention in content and skills acquisition. Due to test fatigue, further validation of the multiple-choice pre/post-test questions was unavailable via free-response questions. Participants were self-selected from among post-secondary students taking introductory astronomy at the university and college level.

Although the total number of participants in this study was 82, a reasonably large sample size for a mixed-methods study, it is small for a dataset study. It is unlikely that disparity by gender is due exclusively to small number statistics (N = 13 for females with matched pre/post-tests), as ANOVA confirmed these results to be statistically significant. In addition, the qualitative data does confirm the quantitative results, and both agree with previous results on gender in STEM learning.

While the qualitative data are rich, the authors cannot know the purpose behind participant’s language choices. Either use of language, whether profanity, innuendos, and gender-based discussion, or the lack of such discussion, may have been natural language choices, or may have been influenced by participants knowing that they were being recorded.

**Implications**

The findings of this study imply the need for further dataset education as authentic STEM experiences at the post-secondary level as well as further research into this skill set. If STEM fields require the use and analysis of datasets, then it is essential to educate various primary, secondary, and post-secondary populations in these skills along with the content (CCSS, 2010; Microsoft 2017; NGSS Lead States 2013). The need for handling big STEM datasets is ubiquitous, and thus it is important to investigate the process by which individuals transition from novices to experts in this domain (Bransford, Brown, and Cocking 2000).
Attention to gender issues must be prevalent during the dataset teaching. Females are disproportionately leaving STEM fields, including astronomy, and they are not being well served in the classes they take (Hill, Corbett, and St. Rose 2010; Microsoft 2017; Nissen and Shemwell 2016). To better serve female STEM students of all levels, researchers need to find effective pedagogies and approaches to STEM education that do not disadvantage students based upon their gender. Many steps that educators can take to help female students will also help male students, such as focusing students away from gendered behaviors and towards the assignment at hand.

Because the gender-based behavior observed among these post-secondary students was so drastic at times, a deeper analysis of the rich qualitative data from these participants and additional ones may provide more insight into these interactions. It would also be intriguing to see if the gender disparity (in both pre/post-test scores, and discussion topics) were replicated among the educators who serve as role models and examples to students (Scantlebury and Baker 2013). Regardless of the cause and age level of students, it is incumbent on instructors to address gender-based behavior as part of classroom management as one step to level the playing field for female students (NAS, 2018b).

Instructors of any gender can take many approaches to help female students overcome the barriers they face both inside and outside the classroom. Assigning single-gender groups have long been known to support the learning of female students using computers (e.g. Busch 1996; Lee 1993), but has yet to be widely implemented. When learners work in mixed-gender groups, instructors can mandate rotation of computer control so all individuals can learn computer skills. Instructors should encourage professional language in the classroom, taking special care to stop jokes and insults (Hill, Corbett, and St. Rose 2010). Self-derogatory talk by females should be discouraged as well, as this too contributes to a hostile environment (NAS, 2018b; Nissen and Shemwell 2016). Role models and mentors should be provided for female students and students from other underrepresented groups to help these individuals see that they can overcome the stereotypes against them (Microsoft 2017; NAS, 2018b). And using authentic STEM experiences has been shown to increase the interest and improve retention of female students (Hill, Corbett, and St. Rose 2010).

For the public to learn either science content, or dataset skills, they need exposure to STEM datasets and training materials. Astronomy dataset education is important for STEM educators, including pre-service and in-service teachers, and there are available, accurate datasets ready to use with these populations. It is reasonable to suggest that if primary or secondary teachers are to adequately teach astronomy or related content, they need to understand it at a level better than their students. Thus, dataset professional development for teachers is desirable and could be integrated with other professional developments, or taught separately to facilitate dataset skill use. This study reinforces what is found in the literature that not only students but teachers of all levels, need access to and instruction about classroom activities (e.g. Hampton et al. 2017). Likewise, professional development opportunities are necessary for pre-service teachers, in-service teachers, and other science educators (Burrows 2015; Burrows et al. 2016).

As stated earlier, the authors argue that authentic science, which is shown to work with students, can utilize computer science and datasets to open doors of STEM opportunities, but these opportunities could be limited if gender issues affect learning or access to those datasets. Students in many disciplines – especially in STEM – could be exposed to large datasets and their content knowledge and skills would improve as evidenced in the
findings of this study. Dataset instruction should be included for students at all levels for both exposure and future success in STEM field careers. Public scholarship (at any level) to educate diverse democracies includes access to all STEM field components. To be competitive in STEM and open options to all students, education on dataset analysis must become a standard part of any STEM curriculum, most likely as an integrated piece of authentic STEM learning. Lastly, researchers must focus on creating a ‘safe space’ for everyone to learn about STEM datasets and their uses.

Acknowledgments

The authors would like to thank the participants of the study. Author 1 would like to thank the members of the University of Wyoming Department of Physics and Astronomy for their support during this study.

Funding

This work was partially supported by the Wyoming Department of Education under Grant #WY140202; by the US National Science Foundation Division of Undergraduate Education under Grant #1339853; and by the Space Telescope Science Institute under Grant #HST-EO-13237.001-A

ORCID

Andria C. Schwortz http://orcid.org/0000-0003-1211-7620
Andrea C. Burrows http://orcid.org/0000-0001-5925-3596

References

Abello, J., P. Pardalos, and M. Resende, Eds. 2002. {Handbook of Massive Data Sets}. 1st ed. Dordrecht: Kluwer Academic Publishers.
Achieve. 2010a. Connecting Science Standards with Assessment: A Snapshot of Three Countries’ Approaches - England, Hong Kong, and Canada (Ontario). USA: Achieve. https://www.achieve.org/publications/connecting-science-standards-assessment
Achieve. 2010b. International Science Benchmarking Report: Taking the Lead in Science Education: Forging Next-generation Science Standards. USA: Achieve. https://www.nextgenscience.org/international-benchmarking
Anderson, J. O., H. S. Lin, D. F. Treagust, S. P. Ross, and L. D. Yore. 2007. “Using Large-scale Assessment Datasets for Research in Science and Mathematics Education: Programme for International Student Assessment (PISA).” International Journal of Science and Mathematics Education 5 (4): 591–614. doi:10.1007/s10763-007-9090-y.
Bailey, J. M. 2011. Astronomy Education Research: Developmental History of the Field and Summary of the Literature. Las Vegas, NV: National Research Council Board on Science Education.
Berg, C., and S. Boote. 2017. “Format Effects of Empirically Derived Multiple-choice versus Free-response Instruments When Assessing Graphing Abilities.” International Journal of Science and Mathematics Education 15 (1): 19–38. doi:10.1007/s10763-015-9678-6.
Bergstrom, Z., and P. Sadler. 2016. “Evolution and Persistence of Students’ Astronomy Career Interests: A Gender Study.” Journal of Astronomy & Earth Sciences Education 3 (1): 77–92.
Bransford, J. D., A. L. Brown, and R. R. Cocking. 2000. How People Learn: Mind, Brain, Experience, and School. Washington, DC: National Academies Press.
Brickhouse, N. W., P. Lowery, and K. Schultz. 2000. “What Kind of Girl Does Science?: The Construction of School Science Identities.” Journal of Research in Science Teaching 37 (5): 441–458. doi:10.1002/(SICI)1098-2736(200005)37:5<441::AID-TEA4>3.0.CO;2-3.

Brothen, T. 1984. “Three Computer-assisted Laboratory Exercises for Introductory Psychology.” Teaching of Psychology 11 (2): 105–107. doi:10.1207/s15328023top1102_14.

Brunner, R. J., S. G. Djorgovski, T. A. Prince, and A. S. Szalay. 2002. “Massive Datasets in Astronomy.” In Handbook of Massive Data Sets, edited by J. Abello, P. M. Pardalos, and M. G. Resende, 931–979. 1st ed. Dordrecht: Kluwer Academic Publishers.

Burrows, A., M. Lockwood, M. Borowczak, E. Janak, and B. Barber. 2018. “Integrated Stem: Focus on Informal Education and Community Collaboration Through Engineering.” Education Sciences 8 (1): 1–4. doi:10.3390/educsci8010004.

Burrows, A. C. 2015. “Partnerships: A Systemic Study of Two Professional Developments with University Faculty and K-12 Teachers of Science, Technology, Engineering, and Mathematics.” Problems of Education in the 21st Century 65: 28–38.

Burrows, A. C., M. A. DiPompeo, A. D. Myers, R. C. Hickox, M. Borowczak, D. A. French, and A. C. Schwortz. 2016. “Authentic Science Experiences: Pre-Collegiate Science Educators Successes and Challenges During Professional Development.” Problems of Education in the 21st Century 70: 59–73.

Busch, T. 1996. “Gender, Group Composition, Cooperation, and Self-Efficacy in Computer Studies.” Journal of Educational Computing Research 15 (2): 125–135. doi:10.2190/KQJJ-RTW1-VVUY-BHLG.

Carey, C. C., and R. D. Gougis. 2017. “Simulation Modeling of Lakes in Undergraduate and Graduate Classrooms Increases Comprehension of Climate Change Concepts and Experience with Computational Tools.” Journal of Science Education and Technology 26 (1): 1–11. doi:10.1007/s10956-016-9644-2.

Cheruvelil, K., and P. Soranno. 2018. “Data-intensive Ecological Research Is Catalyzed by Open Science and Team Science.” BioScience 68 (10): 813–822. doi:10.1093/biosci/biy097.

Cheung, A., R. E. Slavin, E. Kim, and C. Lake. 2017. “Effective Secondary Science Programs: A Best-evidence Synthesis.” Journal of Research in Science Teaching 54 (1): 58–81. doi:10.1002/tea.21338.

Cohen, J. 1977. Statistical Power Analysis for the Behavioral Sciences. New York, NY: Academic Press.

Creswell, J. W. 2013. Qualitative Inquiry & Research Design: Choosing among Five Approaches. Los Angeles, CA: Sage Publications.

Danaia, L., D. H. McKinnon, and M. Fitzgerald. 2017. “Ideal Pictures and Actual Perspectives of Junior Secondary School Science: Comparisons Drawn from Australian Students in an Astronomy Education Programme.” Research in Science & Technological Education. 35 (4): 445–460. doi:10.1080/02635143.2017.1344959.

Day, J., J. B. Stang, N. G. Holmes, D. Kumar, and D. A. Bonn. 2016. “Gender Gaps and Gendered Action in a First-year Physics Laboratory.” Physical Review Physics Education Research 12 (2): 1–14. doi:10.1103/PhysRevPhysEducRes.12.020104.

Eymur, G., and Ö. Geban. 2017. “The Collaboration of Cooperative Learning and Conceptual Change: Enhancing the Students’ Understanding of Chemical Bonding Concepts.” International Journal of Science and Mathematics Education 15 (5): 853–871. doi:10.1007/s10763-016-9716-z.

Gültepe, N. 2016. “Reflections on High School Students’ Graphing Skills and Their Conceptual Understanding of Drawing Chemistry Graphs.” Kuram Ve Uygulama Egitim Bilimleri 16 (1): 53–81.

Hake, R. 1998. “Interactive-engagement versus Traditional Methods: A Six-thousand-student Survey of Mechanics Test Data for Introductory Physics Courses.” American Journal of Physics 66 (1): 64–74. doi:10.1119/1.18809.

Hampton, S. E., M. B. Jones, L. A. Wasser, M. P. Schildhauer, S. R. Supp, J. Brun, and R. R. Hernandez. 2017. “Skills and Knowledge for Data-intensive Environmental Research.” BioScience 67 (6): 546–557. doi:10.1093/biosci/bix025.

Harris, C. J., W. R. Penuel, C. M. D’Angelo, A. H. DeBarger, L. P. Gallagher, C. A. Kennedy, B. H. Cheng, and J. S. Krajcik. 2015. “Impact of Project-Based Curriculum Materials on Student Learning in
National Research Council. 2010. **New Worlds, New Horizons in Astronomy and Astrophysics.** Washington, D.C. http://sites.nationalacademies.org/bpa/bpa_049810

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

Nieminen, P., A. Savinainen, and J. Viiri. 2013. “Gender Differences in Learning of the Concept of Force, Representational Consistency, and Scientific Reasoning.” *International Journal of Science and Mathematics Education* 11 (5): 1137–1156. doi:10.1007/s10763-012-9363-y.

Nissen, J. M., and J. T. Shemwell. 2016. “Gender, Experience, and Self-efficacy in Introductory Physics.” *Physical Review Physics Education Research* 12 (2): 1–16. doi:10.1103/PhysRevPhysEducRes.12.020105.

Nissen, J. M., R. M. Talbot, A. N. Thompson, and B. Van Dusen. 2018. “Comparison of Normalized Gain and Cohen’s D for Analyzing Gains on Concept Inventories.” *Physical Review Physics Education Research* 14 (1): 010115. doi:10.1103/PhysRevPhysEducRes.14.010115.

Nyhof-Young, J. 2000. “The Political Is Personal: Reflections on Facilitating Action Research in Gender Issues in Science Education.” *Educational Action Research* 8 (3): 471–498. doi:10.1080/09650790000200134.

Nyström, E. 2007. “Exclusion in an Inclusive Action Research Project: Drawing on Student Perspectives of School Science to Identify Discourses of Exclusion.” *Educational Action Research* 15 (3): 417–440. doi:10.1080/09650790701549693.

Raddick, M. J., G. Bracey, P. L. Gay, C. J. Lintott, P. Murray, K. Schawinski, A. S. Szalay, and J. Vandenbergh. 2010. “Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers.” *Astronomy Education Review* 9 (1): 1–18. doi:10.3847/AER2009036.

Resnick, I., K. A. Kastens, and T. F. Shipley. 2018. “How Students Reason about Visualizations from Large Professionally Collected Data Sets: A Study of Students Approaching the Threshold of Data Proficiency.” *Journal of Geoscience Education* 66 (1): 55–76. doi:10.1080/10899995.2018.1411724.

Scantlebury, K., and D. Baker. 2013. “Gender Issues in Science Education Research: Remembering Where The Difference Lies.” In *Handbook of Research on Science Education*, edited by, 1. Mahwah, NJ: Lawrence Erlbaum Associates.

Schneider, D. P., P. B. Hall, G. T. Richards, M. A. Strauss, D. E. Vanden Berk, S. F. Anderson, and W. N. Brandt. 2007. “The Sloan Digital Sky Survey Quasar Catalog. IV. Fifth Data Release.” *The Astronomical Journal* 134 (1): 102–117. doi:10.1086/518474.

Schwamb, M. E., J. A. Orosz, J. A. Carter, W. F. Welsh, D. A. Fischer, G. Torres, and A. W. Howard. 2013. “Planet Hunters: A Transiting Circumbinary Planet in A Quadruple Star System.” *The Astrophysical Journal* 768 (2): 127–148. doi:10.1088/0004-637X/768/2/127.

Scott, S., H. Asoko, and J. Leach. 2007. “Student Conceptions and Conceptual Learning in Science.” In *Handbook of Research on Science Education*, edited by S. Abell and N. Lederman, 31–56. Vol. 1. Mahwah, NJ: Lawrence Erlbaum Associates.

Shen, Y., G. T. Richards, M. A. Strauss, P. B. Hall, D. P. Schneider, S. Snedden, D. Bizyaev, et al. 2011. “A Catalog of Quasar Properties from Sloan Digital Sky Survey Data Release 7.” *The Astrophysical Journal Supplement Series* 194 (2): 45. doi:10.1088/0067-0049/194/2/45.

Slavin, R. 2020. “How Evidence-based Reform Will Transform Research and Practice in Education.” *Educational Psychologist* 55 (1): 21–31.

Slavin, R. E., C. Lake, P. Hanley, and A. Thurston. 2014. “Experimental Evaluations of Elementary Science Programs: A Best-evidence Synthesis.” *Journal of Research in Science Teaching* 51 (7): 870–901. doi:10.1002/tea.21139.

Snyder, T., and S. Dillow. 2011. *Digest of Education Statistics*, 2010. Washington, DC: National Center for Education Statistics.

Spuck, T. 2014. “Putting the “Authenticity” into Science Learning.” In *Einstein Fellows: Best Practices in STEM Education*, edited by T. Spuck and J. Leigh, 118–157. New York, NY: Peter Lang.

Sullivan, G. M., and R. Feinn. 2012. “Using Effect Size – Or Why the P Value Is Not Enough.” *Journal of Graduate Medical Education* 4 (3): 279–282. doi:10.4300/JGME-D-12-00156.1.

Tekkumru-Kisah, M., C. Schunn, M. K. Stein, and B. Reynolds. 2019. “Change in Thinking Demands for Students across the Phases of a Science Task: An Exploratory Study.” *Research in Science Education* 49 (3): 859–883. doi:10.1007/s11165-017-9645-z.
Türk, C., and H. Kalkan. 2018. “Teaching Seasons with Hands-on Models: Model Transformation.” Research in Science & Technological Education 36 (3): 324–352.

Vygotsky, L. S. 1978. Mind in Society: The Development of Higher Mental Process. Cambridge, MA: Harvard College Press.

Wallace, R. M., J. Kupperman, J. Krajcik, and E. Soloway. 2000. “Science on the Web: Students Online in a Sixth-grade Classroom.” The Journal of the Learning Sciences 9 (1): 75–104. doi:10.1207/s15327809jls0901_5.

Weintrop, D., E. Beheshti, M. Horn, K. Orton, K. Jona, L. Trouille, and U. Wilensky. 2015. “Defining Computational Thinking for Mathematics and Science Classrooms.” Journal of Science Education and Technology 25 (1): 127–147. doi:10.1007/s10956-015-9581-5.

Westwood, P. 2001. “Differentiation’ as a Strategy for Inclusive Classroom Practice: Some Difficulties Identified.” Australian Journal of Learning Disabilities 6 (1): 5–11. doi:10.1080/19404150109546651.

Whitten, I. H., E. Frank, and M. A. Hall. 2011. Data Mining: Practical Machine Learning Tools and Techniques. 3rd ed. Burlington, MA: Morgan Kaufmann.

Wilcox, B. R., and H. J. Lewandowski. 2016. “Open-ended versus Guided Laboratory Activities: Impact on Students’ Beliefs about Experimental Physics.” Physical Review Physics Education Research 12 (2). doi:10.1103/PhysRevPhysEducRes.12.020132.