Contemporary Educational Psychology 59 (2019) 101802

Contents lists available at ScienceDirect

Contemporary Educational Psychology

journal homepage: www.elsevier.com/locate/cedpsych

Scientific sensemaking supports science content learning across disciplines and instructional contexts

Matthew A. Cannady[a,*], Paulette Vincent-Ruz[b], Joo Man Chung[c], Christian D. Schunn[b]

[a] Lawrence Hall of Science, University of California, Berkeley, United States
[b] Learning Research and Development Center, University of Pittsburgh, United States
[c] Georgetown University, United States

ARTICLE INFO

Keywords:
Scientific sensemaking
Science learning
Middle school
Scientific practices

ABSTRACT

Science consists of a body of knowledge and a set of processes by which the knowledge is produced. Although these have traditionally been treated separately in science instruction, there has been a shift to an integration of knowledge and processes, or set of practices, in how science should be taught and assessed. We explore whether a general overall mastery of the processes drives learning in new science content areas and if this overall mastery can be improved through engaged science learning. Through a review of literature, the paper conceptualizes this general process mastery as scientific sensemaking, defines the sub-dimensions, and presents a new measure of the construct centered in scenarios of general interest to young adolescents. Using a dataset involving over 2500 6th and 8th grade students, the paper shows that scientific sensemaking scores can predict content learning gains and that this relationship is consistent across student characteristics, content of instruction, and classroom environment. Further, students who are behaviorally and cognitively engaged during science classroom activities show greater growth in scientific sensemaking, showing a reciprocal relationship between sensemaking ability and effective science instruction. Findings from this work support early instruction on sensemaking activities to better position students to learn new scientific content.

1. Introduction

Science consists of both a body of knowledge and a set of processes by which the knowledge is produced. Historically, these two aspects were assessed separately (e.g., Porter, 2002) and taught relatively separately (e.g., with an introduction section on skills or via isolated projects or labs). The last decade has been marked by a substantial shift to an integrated view of both how science should be taught and how science learning should be assessed. Now, consensus reports (e.g., NRC, 2007, 2012; OECD 2017) assert that the scientific processes that have historically been used to generate knowledge within a scientific field should also be used to learn science content in a goal directed way (e.g., by designing, conducting, and interpreting experiments, or by arguing from existing sources). In this new perspective, with strong social cognitive theoretical foundations (Bandura, 1986), the term "scientific practices" or "scientific literacy competencies" is used to carefully frame this notion of science processes as a set of core skills that supports the learning of science in a way that is connected to the historical processes used by scientists for scientific knowledge building but is using these processes in a meaningful way rather than just memorizing how scientists applied them (Berland et al., 2015). Further, new science standards (e.g., Next Generation Science Standards, NGSS; PISA scientific literacy competencies) strongly claim that science practices must be demonstrated in use with scientific content and that scientific content must be demonstrated through use with scientific practices.

While the central point about the importance of practices and content integration is well supported by existing data (for a summary, see NRC, 2012), embedded within these new conceptions of teaching and learning science are some open questions that bear further investigation. These questions have important implications for both assessment and instruction, and two of these questions are critically examined here. The first open question is about the generality and transferability of practices across situations or content. Can students who have developed their capacity to engage in science practices then apply those science practices across other domains (e.g., biology, chemistry) in order to support learning science content in those new and otherwise separate domains? If practices are very tightly bound to science content given how they are taught and learned (e.g. Brown, Collins, & Duguid, 1989; Sadler, 2009), students may struggle with using these practices in new content areas. Concretely, if a student has

* Corresponding author.
E-mail addresses: mcannady@berkeley.edu (M.A. Cannady), pvincentruz@pitt.edu (P. Vincent-Ruz), schunn@pitt.edu (C.D. Schunn).

https://doi.org/10.1016/j.cedpsych.2019.101802

Available online 17 September 2019

come to be able to construct and analyze graphs related to core biology ideas, will they be able to more readily be able to construct and analyze graphs as part of learning chemistry ideas?

The second, but related, open question has to do with the coherence of practices. If science consists of independent practices (e.g., developing and using models vs. planning and carrying out investigations) that each must be taught, do students acquire them in independent ways? If science practices work together in overall cycles of inquiry, then students who master some practices will be better positioned to master other practices, and there will be an overall mastery level of science practices which can be assessed with a single measure (albeit one that involves meaningful science content).

Taking on both of these open questions, we pursue the hypothesis that there is a general overall mastery level of core scientific practices that supports science learning in new science content areas. If true, science instruction (in and out of school) should be organized around developing these core practices early in instruction (to accelerate later learning). Further, assessments will be needed that measure progress on these core practices while still involving meaningful content, but in more general ways to avoid content knowledge barriers in demonstrating capacity with scientific practices. We take up the broader benefits of such an assessment in the general discussion.

In particular, this paper argues that: (1) this general mastery level in science practices can be conceptualized as scientific sensemaking (SSM; although with an additional *meta*-level understanding of science as well); (2) scientific sensemaking can be effectively and efficiently measured using scenarios that invoke shared, intuitive content understandings of the natural world rather than embedded in complex, counterintuitive science content that requires extensive instruction; (3) students tend to vary coherently along this overall sensemaking dimension; (4) overall sensemaking levels are a strong predictor of new science learning; and (5) this overall scientific sensemaking dimension can improve with effective science instruction. The first point is established through a literature review. Then, after discussion of design considerations, a new measure is presented, and its psychometric properties are tested related to the second and third points. The fourth point is tested in a large-scale study of students learning diverse science content across middle school grades with diverse curricula and teaching approaches and specifically addresses our first research question: do variations in initial SSM abilities predict science content learning gains across diverse learning contexts and content? The fifth point is tested using the same dataset including measures of student engagement in classroom learning to answer our second research question: can SSM abilities be improved through high cognitive engagement in science instruction? In the final section of the paper, we discuss the implications of our findings related to scientific sensemaking as a cognitive resource, the modifiability of scientific sensemaking and compare the sub-constructs of our scientific sensemaking measure with the practices of NGSS.

1.1. Science as a sensemaking process

Science learning is not a matter of amassing bits of information from simple and concrete to complex and abstract, or simple logical revision to theories; rather, science learning involves active, semantically-rich processes (Driver, Asoko, Leach, Scott, & Mortimer, 1994; Lehrer & Schauble, 2006; Michael, 2006). To engage in these active processes, one must have the capacity to make sense of scientific phenomenon and content (Bathgate, Crowell, Schunn, Cannady, & Dorph, 2015; Bell, 2004; Zimmerman, 2007). Sensemaking involves seeking coherence and meaning through construction and reconstruction of explanations (Kapon, 2017) across multiple representations and knowledge (Danielak, Gupta, & Elby, 2014). While it can occur as a part of a collective sensemaking process across individuals (Zimmerman, Reeve, & Bell, 2010), in this paper we consider the processes that occur within individuals since these processes play a role in both individual and group sensemaking. Recent conceptualizations of sensemaking aligns with many of the practices inherent to argumentation (Kapon, 2017; Ford, 2012), recognizing that sensemaking is critical to determining the best possible explanation. *Scientific* sensemaking occurs when the criteria used to determine the best possible explanation is in accordance with canonical scientific explanations, rather than everyday sensemaking practices (Warren, Ballenger, Ogonowski, Rosebery, & Hudicourt-Barnes, 2001).

In other words, scientific sensemaking requires cognitive engagement with science-related content as an activity of constructing explanations across representations, using methods generally aligned with the practices of science. Some prior research suggests that students with greater reasoning skills have a stronger commitment to scientific views and demonstrate greater knowledge gains from instruction than students with lower reasoning skills (Gervais, 2015; Lawson & Weser, 1990). In addition to directly improving science content learning, such sensemaking is also thought to increase interest, thereby encouraging learners to spend time on further related activities (Songer, Ben Kelcey, & Gotwals, 2009; Zimmerman, 2007).

Here we expand the sub-constructs of sensemaking that contribute to science learning. Although they work together in a coherent sensemaking process, attending to the individual sub-constructs enables the inquiry into the circumstances in which sensemaking is especially useful for content learning as well as driving educational improvement processes. It is an open question regarding which sub-constructs are especially important. For purposes of theory testing, the sub-constructs that we focus on include asking good questions, seeking mechanistic explanations for natural and physical phenomena, engaging in argumentation about scientific ideas, interpreting data tables, and designing investigations (Apedoe & Ford, 2009; Lehrer, Schauble, & Petrosino, 2001), although we acknowledge there are likely more sub-constructs of importance. In addition to science-practice-based sub-constructs, we also believe *meta*-level understanding of the nature of science is important for sensemaking to take place. We briefly review each sub-construct, arguing that each of these are teachable and should play important, complementary roles in science learning.

Asking Good Questions is a central feature of scientific inquiry as well as a major component in scientific discourse (Cuccio Schirripa & Steiner, 2000). Chin and Osborne (2008) note that students' questions can direct student learning and drive knowledge construction, foster discussion and debate, help students to self-evaluate and monitor their understanding, and increase their motivation and interest in a topic. Further, specific instruction focused on improving student's ability to ask investigable questions can improve this ability and lead to higher learning gains (Cuccio-Schirripa & Steiner, 2000; Allison & Shrigley, 1986).

Designing Investigations is the ability to design a process which isolates a phenomenon to be examined. This includes a control of variables strategy (CVS): attempting to determine if a variable is causal by allowing that variable to vary, while holding other potentially causal variables constant. It extends beyond CVS by including the identification of appropriate counterfactuals and allowing for investigations that explore questions about how phenomena operate. This way of thinking demonstrates a guided search that is important for concept formation and problem solving (Klahr & Dunbar, 1988; Zhou et al., 2016). Interventions focused on active manipulation of materials (virtual or physical) can increase students' ability to design investigations without confounds and thereby improve causal inference (Triona & Klahr, 2010).

Interpreting Data Tables and Graphs refers to the ability to a) analyze and interpret data presented in a table or graph accurately and with intention and b) extract relevant information from a data table or graph in order to answer a research question. These skills are important in evaluating and communicating science (Erduran & Jiménez-Aleixandre, 2007; Sampson & Clark, 2006). Students who are unable to draw evidence from data tables or graphs are limited in their ability to make

inferences about a phenomenon (Sandoval & Millwood, 2005). Student understanding of science content can be improved when students learn to display data in visual forms and when existing graphical representation contain enough cues for students to make sense of them (Stewart, Cipolla, & Best, 2013).

Seeking mechanistic reasoning involves the search for an explanation of a cause and effect relationship (Koslowski, 1996; Russ, Scherr, Hammer, & Mikeska, 2008; Schauble, 1996). Mechanistic or causal reasoning has been found to be crucial for understanding science phenomena (Carey, 2000) and is valued within scientific argumentation (Russ et al., 2008). Through context specific instruction, students are able to increase their ability to engage in mechanistic reasoning (Duncan & Tseng, 2010; Hmelo-Silver, 2004).

Engaging in argumentation about science ideas is an important tool for science learning in the classroom (Bell, 2004; Osborne, Erduran, & Simon, 2004) and a core component of authentic scientific thinking (Bathgate et al., 2015; Toulmin & Rieke, 1984). Engaging in argumentative activities in classroom science has been shown to develop scientific thinking (Koslowski, 1996; Kuhn, 1992; Sadler & Zeidler, 2005) and promote conceptual change (Andriessen, 2006). Further, targeted instruction with particular attention to a learning progression can improve student's ability to engage in scientific argumentation (Berland & McNeill, 2010). Given the breadth of argumentation in science and a limited space to measure the construct in our assessment, we focused on students' ability to connect claims to evidence, including prioritizing among forms of justification to use as evidence, and refuting alternative claims. These specific elements were chosen as they are fairly consistent across the frameworks used to assess student argumentation (Sampson & Clark, 2008) and are associated with coherence of the explanation.

Understanding the changing nature of science (e.g., that theory is grounded in evidence and is revised with new methods and evidence) is a critical *meta*-understanding component of science learning, especially science learning based in the practices of science. The National Science Teachers Association (NSTA) argues that "all those involved with science teaching and learning should have a common, accurate view of the nature of science" (NSTA, 2000). A weak understanding of the nature of science has been linked to students misunderstanding of scientific practices (Osborne, Collins, Ratcliffe, Millar, & Duschl, 2003; AAAS, 1994). Instruction on the nature of science, when grounded in learner's experience in science research projects, enhances students understanding of how science works (Duschl & Grandy, 2012).

1.2. Background on the development of an SSM assessment

To test our theoretical claims about the coherence of scientific sensemaking and the function of SSM in science learning, we needed to develop a functional SSM assessment. Developing such a measure involved several challenges. In this section, we describe the challenges and our strategies for addressing these challenges.

Content-integrated but not rare content-dependent. It is difficult to meaningfully engage in scientific sensemaking void of content (NRC, 2012). However, in measuring scientific sensemaking, our focus is on the ability to apply the practices themselves in some meaningful science situation. To address this challenge, we embedded our instrument within scenarios that leverage common, rather than rare, science content knowledge. We also structured our scenarios to include and support access to the necessary content knowledge within the assessment. This approach is also used in the PISA assessments of scientific literacy competencies (OECD, 2017).

Effort worthy. Engaging in scientific sensemaking requires effort, and there is often little incentive for students to put forth effort to perform on an assessment, and ability scores are confounded by motivational levels. In order to motivate students to put forth effort, the content we selected for the assessment scenarios were so-called "charismatic megafauna" (i.e., dolphins and monkeys), in which a widely-shared interest in the topic (at the tested ages) motivates some basic level of effort (Bathgate, Schunn, & Correnti, 2014).

Repeated measurement. In order to measure gains in scientific sensemaking, we needed to be able to measure the construct multiple times for the same individuals. Using the same instrument in a pre/post setting can invite repeated testing effects wherein students can remember correct answers. To address this challenge, we developed multiple versions of the instrument, changing the animal in the content (from dolphins to monkeys), but maintaining the structure of the assessment, the ways in which the sub-constructs were assessed, and the overall length.

1.3. Study goals

We present a large-scale study that was designed to vet the psychometric properties of an SSM instrument and test two core research questions using this instrument:

RQ1: Do variations in initial SSM abilities predict science content learning gains across diverse learning contexts and content?
RQ2: Can SSM abilities be improved through high cognitive engagement in science instruction?

The psychometric properties are presented to address the coherence of sensemaking in an overall construct. The first research question addresses the theoretical issues of transferability of practices across science disciplines and instructional approaches to support learning new content for all learners. The second research question addresses the practical viability of sensemaking as a target of science instruction, rather than a difficult-to-modify cognitive capacity like working memory (Kyllonen & Christal, 1990), intelligence (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002), or grit (Duckworth, Peterson, Matthews, & Kelly, 2007).

2. Methods

2.1. Study overview

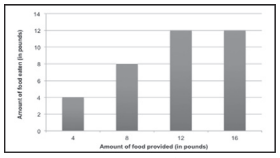
This work drew upon the Activated Learning Enables Success study of 2014 (ALES14; Dorph, Cannady, & Schunn, 2016, Bathgate & Schunn, 2017a, Bathgate & Schunn, 2017b, Vincent-Ruz & Schunn, 2017). This dataset, available to other researchers for analysis, includes over 2500 6th and 8th grade students drawn from schools of varying demographic makeup implementing a variety of biology or chemistry science curriculum units over the course of a school year. The research design, including recruitment and the measures used, were reviewed and approved by an institutional review board. The students completed an SSM assessment and an assessment of science content, carefully aligned to each classes' science instruction, pre and post the first instructional unit (approximately 4 months long). Cognitive-behavioral and affective engagement during science instruction was also measured in multiple science classes, and follow-up regression analyses examined the relationship of classroom engagement to pre-post changes in SSM levels. The two grade levels represent a cross-section of the middle school years, while the repeated measurements of SSM and science knowledge provide a view of growth within middle school years. Further, to characterize diversity in instructional approaches, measures of classroom instructional practice (e.g., student-centric vs. teacher-centric instruction, emphasis on hands-on material use vs. use of traditional materials) were collected through teacher logs and classroom observations. Regression analyses examined the relationship of initial SSM levels to amount of science content learning across grades, science content, and instructional approaches, controlling for key demographic variables. In addition to analyses of this larger dataset, we conducted analyses on additional data collected from a smaller subset of students on an open-response assessment to provide additional validation of

Table 1
Participants' mean and standard deviations of age along with gender and ethnicity percentages across locations and by grades.

			Age (Years)			Race/Ethnicity				
Location	Grade	n	M	SD	Female	White	Asian	African American	Hispanic/Latino	Other
Eastern US City	6th	773	11.5	0.7	51%	47%	5%	49%	8%	1%
	8th	737	13.4	0.7	52%	55%	7%	47%	11%	1%
Western US City	6th	954	11.3	0.6	52%	41%	14%	21%	28%	1%
	8th	1203	13.3	0.6	51%	44%	14%	19%	26%	1%

Note: Ethnicities sum to more than 100% due to multi-ethnic responses.

Table 2
Example items for each scientific sensemaking sub-construction from the dolphins scenario.

Sub-construct	Sample Question
Generating testable questions	Elijah wonders if the temperature of the water makes a difference in how much dolphins play. Which question is the best to ask to investigate this? a) Do dolphins play in warm water? b) Which other animals live in the same part of the ocean as dolphins? c) Do dolphins live in warm or cold water? d) Do dolphins play more when the water is warm or cold?
Designing investigations or experiments	You are wondering which type of dolphin eats the most amount of food per day. What is the best evidence you could get to answer this question? a) You observe how much Bottlenose dolphins eat in a day. b) You find a scientific study that says that Bottlenose dolphins swim over 18 miles per hour. c) You measure the amount of food eaten by 50 dolphins of each type. d) You ask several people who work at an aquarium to estimate how much food each type of dolphin eat.
Interpretation of data and/or tables	 *Seth is looking at some data from an experiment. Dolphins were given different amounts of food, and then scientists measured how much food they ate.* Seth says that dolphins are full after they eat 12 lb of fish. Which piece of evidence in the graph makes Seth think this is true? a) Dolphins can eat 16 lb. b) Dolphins eat only 4 lb when given 4 lb. c) Dolphins given 16 lb only eat 12 lb. Dolphins given 12 lb then eat 12 lb.
Constructing mechanistic explanations	Maria and Celia both think:**Dolphins are affected most by the amount of noise.**Many dolphins left the cove when there was a lot of noise.Maria says: Dolphins cannot hear each other when there is a lot of noise, so they leave.Celia says: Dolphins leave because it is noisy, so when there is a lot of noise they leave.Whose reasoning for why the dolphins leave the cove is more scientific? a) Celia because she repeats the important idea. b) Maria because she explains how the noise causes a problem. c) Celia because she uses data collected from a study. d) Maria because I would also leave if my environment was noisy.
Nature of Science	Dr. Powers is investigating how dolphins communicate with each other. Which of these would be an important part of her work as a scientist? a) Ask people if they have a favorite type of dolphin. b) Talk to other scientists about dolphins. c) Decide if dolphins are more popular than sharks. d) Write imaginative stories about dolphins.

measures of scientific sensemaking. Further, we conducted a linking study to allow us to compare scores from the pre assessment to the modified version used as the post assessment. The sample participants for each of these three studies are described in the following section.

2.2. Participants

School districts for the main study were recruited from two different regions in the United States to ensure diversity of participants, curricula, and cultural contexts (see Table 1; see *Measures* section for details on demographic instruments). Most science teachers in the selected schools agreed to participate. The first regional group included three public school districts in the Western US, having a high proportion of recent immigrants, English Learners, and Hispanic/Latino and Asian students. This sample included 27 sixth-grade classes and 32 eighth-grade classes from five schools with typical middle school arrangements (i.e., 6th–8th grade). From publicly available data, the school demographics varied widely: students qualifying for Free/Reduced Lunch ranging 24%–92%; Underrepresented Minority (not Caucasian or Asian) ranging 13%–94%.

The second regional group was drawn from a mid-size city in the Eastern U.S. characterized by a high proportion of African Americans. This sample included 26 sixth grade classes and 21 eighth grade classes from six public schools with widely varying student demographics (Free/Reduced Lunch 38%–84%; Underrepresented Minority 32%–99%) and configurations (four schools were 6th–12th grade, three of these were magnet schools focusing on the Arts, Science & Technology, and International Languages; and two schools had typical

6th–8th grade middle school arrangements).

To gather further validity evidence for our measure of scientific sensemaking, the research team examined responses on existing open-ended assessments that were part of the classroom curriculum broadly shared across the second regional group. Small modifications were made to some of the classroom assessments before deploying them so that each open-ended question consistently called upon a part of scientific sensemaking (e.g. asking questions, arguing from evidence). Rubrics were developed to measure the components of sensemaking in the open responses. This data was obtained from 348 students (46% Female), coming from six urban public schools all from the second regional group, while still representing a range of school configurations and student demographics (45% Caucasian; 42% African American; 5% Asian; and 8% Hispanic).

To determine if instructional practice might change levels of SSM, it was necessary to establish equivalency across the two versions of the SSM instrument using a linking study (i.e., place pre and post scores are on the same scale). 121 7th grade students were recruit for a linking study from an urban school not participating in our larger study, but in one of the participating school districts and similar to many of the schools in the larger study (64% Caucasian; 16% African American; 8% Asian; 5% Hispanic; and 7% Multiracial).

2.3. Measures

Scientific sensemaking. The scientific sensemaking instrument consists of 12 items all contextualized within a particular scenario. Items are presented in a specific order to represent a coherent investigation into a topic (e.g., going from asking questions to analyzing data to evaluating explanations). For this study, the two contextualizing scenarios involved dolphins and monkeys respectively. Below, in Table 2, we offer several example items from the dolphins scenario. Although the items are contextualized with meaningful science content, they rely on broad, rather than specific knowledge. For example, it assumes that respondents know that dolphins are animals and that they live in water. It does not assume that students know how much food dolphins eat, what parts of the world they live in, or any other rare or specialized dolphin knowledge. Each item is associated with a sub-construct of scientific sensemaking included in our assessment. Student responses are recoded as correct (1) or incorrect (0), for each item. See the online Appendix for the full version of both the Dolphins and Monkeys instrument and the coding of each item.

A series of steps were followed to generate an instrument that could provide valid inferences for the purposes identified here (see, e.g., Kane, Crooks, and Cohen, 1999; Kane, 2006). In this section, we describe, the instrument iterative development process, including data collection with open-ended items, expert review, and cognitive labs conducted with pilot students. We conclude with the psychometric properties of the instrument.

First, the definition of the construct and the development of an initial instrument was conducted by a team of researchers with diverse expertise in scientific research and science learning. This development model ensured capturing the way scientists think. Further, once an initial item set was developed, we invited multiple rounds of outside experts in the field to provide advice and feedback on our instrument. The initial instrument involved a number of open-response items (e.g., asking students to suggest good questions to investigate). A quantitative study using this instrument revealed that there was a very high correlation between performance on only the multiple-choice items and the whole instrument (Bathgate et al., 2015). Since coding open-responses is very time-intensive and the current investigation required a large number of participants, a revised instrument was created that used only multiple-choice items. Information taken from student responses to the open-ended items to construct the closed-ended items (e.g., kinds of questions asked, mechanisms considered, arguments provided).

After revisions were made, we used cognitive labs to establish a strong link between the cognitive process, the observed response, and the interpretation of that response (Leighton, 2004). In particular, researchers asked pilot students taking the survey one-on-one with the researcher to articulate the items in their own words and explain why they have chosen their answer. The responses were audio recorded and transcribed. Researchers reviewed these transcripts to verify the clarity of the items, to understand uniqueness of interpretation, and to determine if there were discrepancies between the item and the response options or evidence of gender or ethnicity bias in response options. Revisions were made as needed, and the process was repeated until the assessment questions consistently demonstrated internal validity. Note that this same systematic qualitative and quantitative instrument development process was also followed for all the scales used in the current study (Dorph et al., 2016, Bathgate & Schunn, 2017a, Vincent-Ruz & Schunn, 2017, Dorph, Bathgate, Schunn, & Cannady, 2018).

Because the scientific sensemaking assessment is new, and the inferences drawn from in this study depend heavily on the measure itself, we present the validity evidence for the measure including its psychometric properties and its relationship to other measures of sensemaking. These provide evidence of the validity of the inferences drawn from the instrument, as well as evidence of the coherence of the sub-dimensions as an overall scientific sensemaking ability.

Reliability. Armor's theta coefficients were produced using *R*. The resulting θcoefficients based on the full 12-item scales were 0.89 and 0.91 for the Dolphins scenario and Monkeys scenario, respectively (see Table 3 for mean scores and standard deviation).

Confirmatory factor analysis. Confirmatory factor analyses (CFA) were conducted on each of the two scenarios through Mplus (Version 7.11) utilizing the mean- and variance-adjusted weighted least square (WLSMV) estimator. A one-factor model was produced for each scale, then examined for adequately large factor loadings and model fit statistics. Factor loadings based on the full 12-item scales were generally satisfactory for both scenarios, ranging between 0.38 and 0.77 for the Dolphins scenario, and 0.36 to 0.86 for the Monkeys scenario (Costello and Osborne, 2011). Model fit statistics were also satisfactory for both scales (Dolphins: CFI = 0.99, TLI = 0.98, RMSEA = 0.03; Monkeys: CFI = 0.99, TLI = 0.99, RMSEA = 0.03; Byrne, 2001; Hu & Bentler, 1999). Overall, the results indicated that a unidimensional model was a satisfactory fit to the data for each scale.

IRT analysis. We conducted an Item Response Theory (IRT) analysis to ensure a distribution of difficulty levels across items. In the Appendix, we include a Person-Item map, which shows the relationship between respondents' abilities and item difficulty on the Monkeys version of the SSM scale to provide a sense of which items were challenging to students across different SSM ability levels. Since the correlation between IRT scores and mean scores is very high (r = 0.99) for both the scenarios, mean scores are used for the remaining analyses.

Relationship to embedded assessments. Seeking to understand the relationship between our measure of scientific sensemaking and other ways of assessing scientific sensemaking in different contexts, we compared scores on our instrument to those on embedded classroom assessments. In a separate study, we obtained classroom assessments that were linked to the content of learning in the classroom and called

Table 3
Means, SD and reliability (Alpha for rating scales and Theta for dichotomous accuracy items) for each of the student assessments. SSM means are raw, prior to rescaling.

Scale	# of items	Alpha	Theta	M	SD
Cognitive/Behavioral Engagement	5	0.80		2.9	0.5
Affective Engagement	3	0.84		2.7	0.6
Pre-Content Knowledge Assessment	18		0.6	42%	21%
Post-Content Knowledge Assessment	18		0.7	49%	20%
SSM-Dolphins	12		0.89	6.9	3.1
SSM-Monkeys	12		0.91	5.7	2.7

upon some of the same process of scientific sensemaking (e.g. asking questions, arguing from evidence) measured in our scale. The classroom assessment was open-ended and rubrics were developed to measure the components of sensemaking, rather than content knowledge. The rubrics were first compared across multiple raters until sufficient inter-rater reliability was reached (weighted kappa = 0.86). The scores from the two methods of measuring sensemaking correlated at 0.37 ($p < 0.05$), indicating a statistically significant correlation with sensemaking scores with open-ended responses. Taking into account the noise from coding brief open-ended assessments that also captured only some of the sub-constructs in sensemaking and did so in a different content domain, the medium-sized, yet statistically significant, correlation implies that our multiple-choice assessment provides information similar to what would be obtained from an open-ended scientific sensemaking assessment.

Engagement in science learning activities. The engagement survey has two scales that are administered immediately after a learning activity to measure engagement during the activity to eliminate memory biases (Dorph et al., 2016). The instrument, built on theoretical and empirical research on engagement in classrooms (e.g., Carini, Kuh, & Klein, 2006; Finn, Pannozzo, & Voelkl, 1995; Fredricks, et al, 2004; Fredricks, et al, 2011), was also developed through extensive iterative development of expert review, cognitive labs, quantitative testing (including exploratory and confirmatory factor analysis and IRT analyses), and comparison with observation of student engagement in several contexts (Dorph et al., 2016; Vincent-Ruz & Schunn, 2017). The first scale has 5 items and measures students' self-reported behavior and cognitive engagement during a science learning activity (Fredricks, Blumenfeld, & Paris, 2004). Example items from this scale include; I was focused on the things we were learning all of the time; I was busy doing other tasks; Time went by quickly. Response options are: YES!, yes, no, NO!. The second scale has 3 items and measures students' affective engagement during the same learning activity, and uses the same response options. The items include: I felt happy; I felt excited; I felt bored. Each scale is unidimensional and has sufficient reliability (alpha > 0.8; see Table 3), and the entire instrument can be completed in less than 5 min. Since IRT analyses revealed equal distance between Likert response (see http://www.activationlab.org/tools/ for confirmatory factor analyses, and item response theory analyses) and the correlation between IRT scores and mean scores are high, mean scores are used for analyses.

Content knowledge assessment. Given the wide range of classrooms recruited for this study, different pre/post-content knowledge tests were used to assess how much the students learned from their classroom instruction during the period of the study. Each classroom was given an assessment that matched the content taught in their curriculum (e.g., a biology assessment for a biology-focused curriculum and a chemistry assessment for a chemistry-focused curriculum). Test-specific z-scores were used to make the scores comparable across forms, thereby measuring relative student gains within each class adjusted across classes for the content covered by their teacher. Each test form consisted of 18 multiple choice questions drawn from released TIMSS (Mullis, Martin, Gonzalez, & Chrostowski, 2004), AAAS (Laugksch & Spargo, 1996), and MOSART (Sadler et al., 2010) items (e.g. "What is the primary energy source that drives all weather events, including precipitation, hurricanes, and tornadoes?" (a) the Sun, (b) the Moon, (c) Earth's gravity, or (d) Earth's rotation). To create the test forms, teachers described the specific content learning goals for the instructional unit in a survey, and a draft assessment was created by selecting matching items from the items banks. These drafts were shared with each teacher and each teacher crossed out items that they did not expect to be included in their instruction over the study period. The process resulted in five different tests for 6th grade and four different tests for 8th grade. These assessments had acceptable reliability (see Table 3 for mean reliability, theta). Note that reliability of content assessments is typically lower at pretest given a higher rate of guessing due to lack of prior exposure to the content. Table 3 also includes mean percent correct at pre and posttest to show that students generally found the tests difficult, but there were not floor or ceiling effect problems that would artificially reduce variability across students.

Table 4
Teacher log survey questions regarding activity type found in the curriculum.

Question	Materials used
Students read (alone or aloud) from a book or other informational text.	Traditional
Students listened to a lecture or presentation.	Traditional
Students watched a live or video-based demonstration.	Traditional
Students used an interactive or simulation on the computer.	Hands-on
Students did a hands-on activity.	Hands-on
Students used tools that scientists use (microscope, beakers, pipettes, etc.).	Hands-on

Background survey. Participants provided demographic information in a survey that asked them about their sex, date of birth, and race/ethnicity. Students were asked to select among six different racial/ethnicity options with which they identified, and were allowed to choose more than one. From the ethnicity data, a binary variable called Underrepresented Minority was created: Students who choose only White, only Asian, or White and Asian, were coded as "0"; all others were coded as "1".

Curriculum materials used. To assess whether SSM predictiveness of content learning varied by type of classroom instruction, teachers completed a teacher log asking them to rate what percentage of classroom time over the prior week was dedicated to traditional instruction vs. hands-on classroom activities (see Table 4). Teachers completed this log once per month, giving us four teacher self-assessments. Teachers responded by rating the percent of time that week devoted to each task. The scale was found to be unidimensional with all items having a factor loading greater than 0.6. Mean responses across the teacher logs were computed for each teacher within each item, and then a ratio of time spent doing Hands-on type activities vs. Traditional Instruction was calculated. We found that teachers reported teaching behaviors throughout the spectrum from hands-on to traditional. Further, there is an important role for different materials and formats in instruction, instruction that was purely hands-on would not be sensible and demonstrations can play an important role in inquiry. The problems arise when students are given too high a ratio of materials consumed in relatively passive ways. We therefore split teachers in the top 35th percentile of the scale distribution into the "hands-on" materials category and teachers in the bottom 35th percentile as a part of the "traditional" materials category. The rest of the teachers were classified as "mixed" to refer to the use of mixed instructional materials; this group was removed for the analysis of interactions by materials used to provide a strong test of this effect.

Classroom discourse. Another feature of classroom instruction that may influence whether SSM predicts content learning is the kind of classroom discourse. Therefore, when researchers were in classrooms to administer the engagement surveys, they also implemented an observation protocol focused on the classroom instruction. In particular, they rated the type of instruction in the classroom (dialogical vs. direct instruction). This dimension was assessed by observation rather than teacher self-report because of demand characteristics associated with this dimension (i.e., teachers realize that dialogical instruction is considered best practice). To minimize rater fatigue, observers were asked to observe for 5-minute periods at multiple times during the class and record how long during those five minutes they observed different forms of classroom dialog (see Table 5). The scale was found to be unidimensional with all items having a factor loading greater than 0.6. Mean ratings were computed for each teacher, and then a ratio of time spent doing dialogic vs. direct instruction was calculated. Again, we found classroom discourse practices throughout the spectrum from dialogical to direct instruction. We therefore split teachers based on top

Table 5
Classroom discourse observation protocol.

Item	Teaching type
Students and teachers are talking to each other: Large group	Dialogic
Students and teachers are talking to each other: Small group or one-on-one	Dialogic
Students and teachers are talking to one another about the lesson	Dialogic
No one is talking	Direct Instruction
Teacher is directing the talk in the room. Student talk is limited to call & response/IRE	Direct Instruction

35th percentile of the scale distribution into "dialogical", the low 35th percentile was classified as "direct instruction" classroom discourse. The rest of the teachers were classified as mixed; this group was removed for the analysis of interactions by discourse to provide a strong test of this effect.

2.4. Procedure

Students completed surveys across six time points (see Table 6 for a summary), two early in the Fall semester, two in the middle, and two at the end of the semester. As demonstrated in Table 6: (1) The Scientific Sensemaking Dolphins scenario was administered during one class period in early Fall. (2) The pre-content knowledge assessment, followed by a demographics survey (to avoid stereotype threat Spencer, Steele, & Quinn, 1999), were administered during another class period, one to two weeks later. (3) & (4) Cognitive-behavioral and affective engagement were measured at the end of class on two different days in October and November. (5) The Scientific Sensemaking Monkeys scenario was administered during one class period in in late January or early February, approximately 4 months after the SSM instrument based on the Dolphins scenario. (6) A week later, the posttest on content knowledge was administered in one class period.

For the study linking the scores from the Dolphins scenario to the Monkeys scenario we gave a different group of students the Dolphins and Monkeys scenario-based assessments one week apart. We used a counterbalancing of scenario order: half of the students were randomly assigned to take the Dolphins scenario first while the other half took the Monkeys scenario first.

3. Results

***RQ1:* Do variations in initial SSM abilities predict science content learning gains across diverse learning contexts and content?**

To address RQ1, we begin with an overall test of whether scientific sensemaking is an independent predictor of students' content learning. For this set of analyses, we use a statistical significance threshold of $p < 0.05$ for main effects and interactions. First, we calculated the correlations between students' scores on the scientific sensemaking assessment with their scores on the content knowledge pre and posttests. These results are shown in Table 7. Scientific sensemaking is correlated to both the pre and posttests of content knowledge (0.43 and 0.44, respectively). These correlations are not so high as to cause multicollinearity problems in our multiple regressions. Further, these medium-sized correlations indicate that the instruments are likely measuring related, but separate constructs.

Table 7
Inter-correlation of pre and posttest content knowledge with pretest sensemaking scale scores. All correlations were significant at the $p < 0.01$ level.

	Knowledge Post-Test Score	Scientific Sensemaking
Knowledge Pre-Test Score	0.56	0.43
Knowledge Post-Test Score	–	0.44

We ran a 2-two level hierarchical linear model (HLM) to control for nested data effects within classrooms and schools using the lme4 package in *R* (Bates et al., 2014). The conditional R^2 was computed using the method provided by Nakagawa and Schielzeth (2013). It estimates the proportion of variance explained by the whole model accounting for the fixed and random factors. Model 1, below in Table 8, represents the fully unconditional model. In this model, we found an intraclass correlation of 0.17, implying that 17% of the total variation in the sample is between classrooms rather than between individuals. Given the magnitude of the ICC, we conducted the rest of our analyses in HLM. Model 2 is the HLM model predicting the content posttest score using only the content pretest score as a predictor. Finally, in Model 3 we added in scientific sensemaking as a predictor. Table 8 shows how, even though prior content knowledge accounts for the greatest amount of variation in the content knowledge posttest score, scientific sensemaking accounts for 8% of the variance and the standardized beta is in similar magnitude to that of the pretest. These results are consistent to those found by Bathgate et al. (2015) using a prior version of this measure and a different but similar dataset (ALES11).

Is scientific sensemaking more important for some science topics than others? We account for the possibility that the biology-focused topic of the scientific sensemaking instrument (Dolphins) may have a differential predictive utility for students studying biology compared to those studying a physics or chemistry curriculum. To test this possibility, we selected two groups of 8th grade students from our dataset (the 6th graders could not be cleanly divided into sufficiently large topic subgroups). The first group had studied a biology curriculum (6 teachers, 24 classrooms n = 602). The second group studied an introductory chemistry curriculum (6 teachers, 12 classrooms n = 352). We conducted a *t*-test by topic to verify there were not different underlying distributions between the two groups on initial scientific sensemaking scores (mean difference = 0.42, $p = 0.44$). Then we tested for an interaction between initial scientific sensemaking scores and the curriculum's topic in predicting posttest content knowledge scores, while controlling for pretest content scores. The interaction with curriculum topic was small and not statistically significant (SSM X Bio $\boldsymbol{b} = -0.10$, $p = 0.1$), and the main effect of scientific sensemaking remained statistically significant (SSM $\boldsymbol{b} = 0.20$, $p < 0.001$). Fig. 1 shows the nature of this relationship. In order to better understand how different levels of scientific sensemaking may relate to curriculum topic, we binned scientific sensemaking in three groups using quantile binning

Table 6
Administration time of year and amount of class time for each survey instrument.

Administration	Class time	Survey
Mid September	1 class period	Scientific Sensemaking Survey (Dolphins Scenario)
Early October	1 class period	Pre-content knowledge assessment & Demographics
October	once during the last 5 min of class	Engagement
November	once during the last 5 min of class	Engagement
Late January/Early February	1 class period	Scientific Sensemaking Survey (Monkeys Scenario)
Late January/Early February	1 class period	Post-content knowledge assessment

Table 8
Models testing the effect of scientific sensemaking on knowledge post-test scores, controlling for knowledge pre-test scores.

	Model 1: Baseline		Model 2: Pre-test Only		Model 3: + Sensemaking	
	β	*p*	*β*	*p*	*β*	*p*
Knowledge Pre-Test Score			0.51	< 0.001	0.40	< 0.001
Scientific Sensemaking					0.29	< 0.001
R^2 conditional			0.34		0.42	
ICC	0.17		0.06		0.06	

against the estimated marginal means of posttest z-scores after controlling for other covariates. This grouping was for visualization purposes only; the regression test of the interaction used the continuous SSM measure. Further note that a negative posttest z-score on the graph is not indicative of students unlearning something they already knew, but rather, that they scored below their classroom's mean. The graph shows how students in both groups achieve high content learning gains when they have higher initial scientific sensemaking skills.

Are there important differences in the distribution of scientific sensemaking scores by sex, ethnicity or grade-level? We then proceeded to test whether sex, grade, or race/ethnicity act as moderators on the way sensemaking predicts posttest content knowledge scores (Table 9). We found no moderation of scientific sensemaking's relationship to learning by sex, grade, or race/ethnicity, implying that SSM plays an important role in science content learning across subgroups, including those subgroups who are traditionally underrepresented and marginalized in science education. We provide a visual representation of these relationships in Fig. 2 by plotting the estimated marginal means for each group controlling for pretest z-scores; the regression test of the interaction used the continuous SSM measure.

Does the relationship between scientific sensemaking and learning gains vary by the instructional style in the classroom? Scientific sensemaking could be more important for learning in classrooms focused on hands-on or inquiry-based instruction than in classrooms with traditional instruction. For example, a classroom that consists of primarily struggle free activities (like the ones typical of textbook or lecture-based instruction) may lead students to a view of science as being a passive compiling of new information. Such a view of science may not call upon the use of scientific sensemaking abilities (Schweingruber, Keller, & Quinn, 2012). Further, scientific sensemaking abilities may have a larger role in student-centric classrooms, where students are active participants and have greater opportunity to make sense of information than in teacher-centric classrooms. To test these possibilities, we conducted two hierarchical linear regression models in order to test for possible moderation effects of classroom discourse or material-use on scientific sensemaking's impact on knowledge posttest scores (Table 10; see Fig. 3). Again, the regression analysis uses the continuous variable, the binning of low, medium and high SSM is for display purposes. None of these moderation effects were statistically significant (discourse $\boldsymbol{b} = 0.05$, $p = 0.22$; materials, $\boldsymbol{b} = -0.05$, $p = 0.26$). Generally speaking, the higher SSM students made almost 0.7 standard deviations greater gains than did lower SSM students (controlling for differences in prior content knowledge), which is a moderate effect. Note that the lack of statistically significant interactions also held when the discourse and materials-used measures were treated as continuous measures rather than categorical measures.

Are all the sub-constructs of scientific sensemaking predictive of learning? Having found that the overall scientific sensemaking score is an important predictor of content knowledge posttest scores across learning contents, contexts, and student populations, we wanted to address the possibility that not all the components of scientific sensemaking are predictive of science learning. Because we had too few items for interpreting data, we merged it with engaging in argumentation since the interpreting item also involved some engagement in argumentation. Thus, we created four sub-scores of scientific sensemaking, one for each of following sub-constructs: (1) asking good questions, (2) seeking mechanistic explanations for natural and physical phenomena & engaging in argumentation about scientific ideas, (3) designing investigations, and (4) understanding the changing nature of science. In calculating these sub-scores, we do not mean to imply that these are sufficient measures of each of these constructs to inform instruction nor do we mean to imply that these are the only important

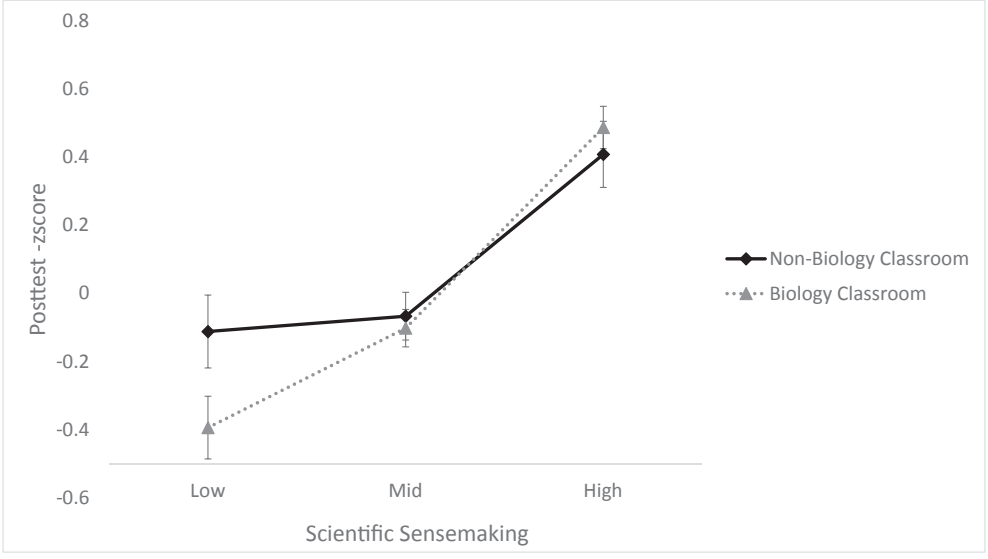


Fig. 1. Estimated marginal means of adjusted posttest scores, controlling for pretest z-scores, for the interaction of binned scientific sensemaking scores and curriculum topic.

Table 9
Models testing the interactions of scientific sensemaking with sex, race and grade on knowledge posttest scores, controlling for knowledge pretest scores.

	Model 4: Grade Interaction		Model 5: Sex Interaction		Model 6: Ethnicity Interaction	
	β	*p*	*β*	*p*	*β*	*p*
Knowledge Pre-Test Score	0.40	< 0.001	0.44	< 0.001	0.43	< 0.001
Scientific Sensemaking	0.31	< 0.001	0.29	< 0.001	0.27	< 0.001
Grade	0.02	0.62				
SSM × Grade	−0.05	0.23				
Sex			0.00	0.84		
SMM × Sex			−0.04	0.34		
Underrepresented Minority					−0.09	0.03
SMM × Underrepresented Minority					−0.01	0.74
R^2 conditional	0.41		0.42		0.44	

aspects of scientific sensemaking. Rather, we are testing whether each of the components of the larger construct are themselves contributing to science learning, taking into account the large correlations between the sub-construct scores (see Table 11). Model 9 (see Table 12) shows the predictive relationships of each sub-construct with knowledge posttest scores (asking questions, $b = 0.09$, $p < 0.001$; design of experiments, $b = 0.08$, $p < 0.001$; arguing with evidence, $b = 0.12$, $p < 0.001$; nature of science, $b = 0.11$, $p < 0.001$). This implies that each sub-construct is a statistically significant predictor of science learning and uniquely contributes to knowledge posttest scores.

RQ2: Can SSM abilities be improved through high cognitive engagement in science instruction?

To determine whether gains in SSM occurred, it was necessary to determine the precise relationship across the pre and post SSM scenarios. Methodologically, this calculation is achieved through a linking study that allows us to convert the score from one assessment to a score on the other assessment. We describe the results of the linking substudy first, and then return to analyses of change in SSM scores from pre to post in the larger study.

Table 10
Models testing interactions effects on knowledge posttest scores of scientific sensemaking with curriculum materials and classroom dialogue.

	Model 7: Classroom Dialogue Interaction		Model 8: Curriculum Type Interaction	
	β	*p*	*β*	*p*
Knowledge Pretest Score	0.39	< 0.001	0.41	< 0.001
Scientific Sensemaking	0.28	< 0.001	0.23	< 0.001
Classroom Dialogue	0.05	0.22		
SSM × Classroom Dialogue	0.00	0.91		
Curriculum Materials			−0.05	0.26
SMM × Curriculum Materials			0.08	0.17
R^2 conditional	0.38		0.42	

Scientific sensemaking measure linking substudy. To link the scores, we used the item response theory (IRT) True Score Method to determine the corresponding ability estimate on the Monkeys scenario for the

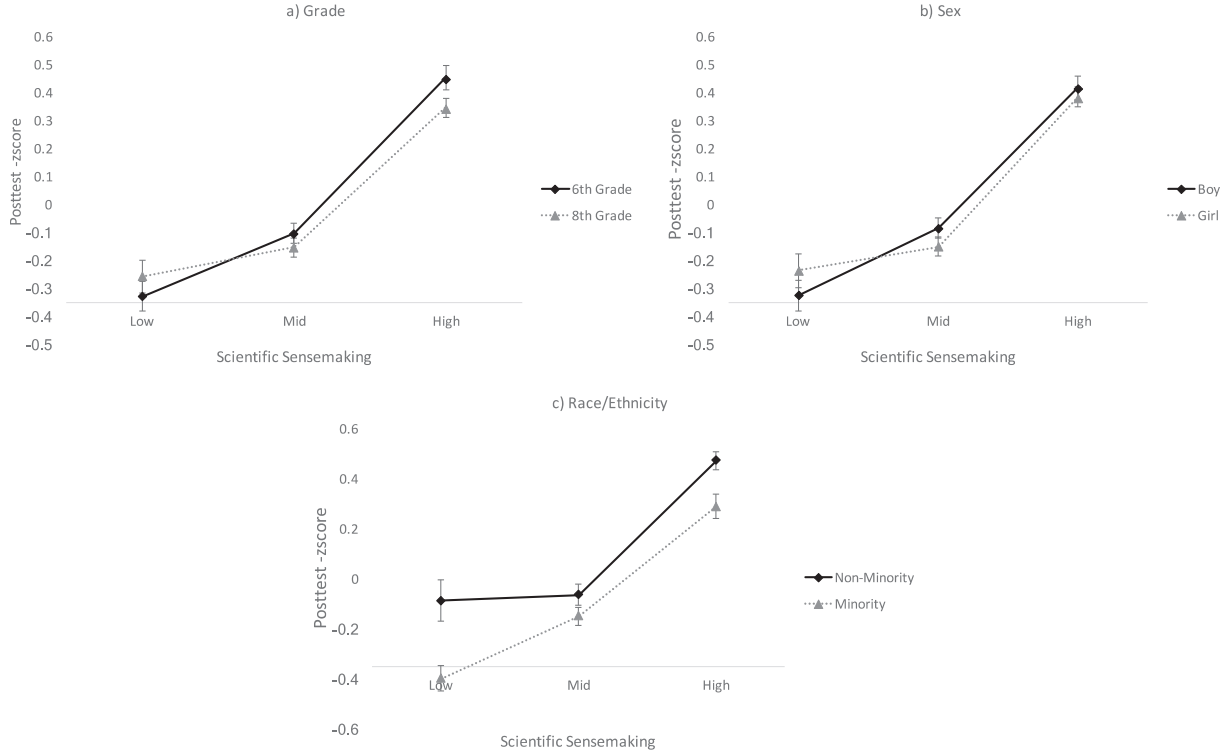


Fig. 2. Estimated marginal means of adjusted post-test knowledge scores, controlling for pretest z-scores, for the interaction of binned scientific sensemaking with (a) grade, (b) race/ethnicity, and (c) sex.

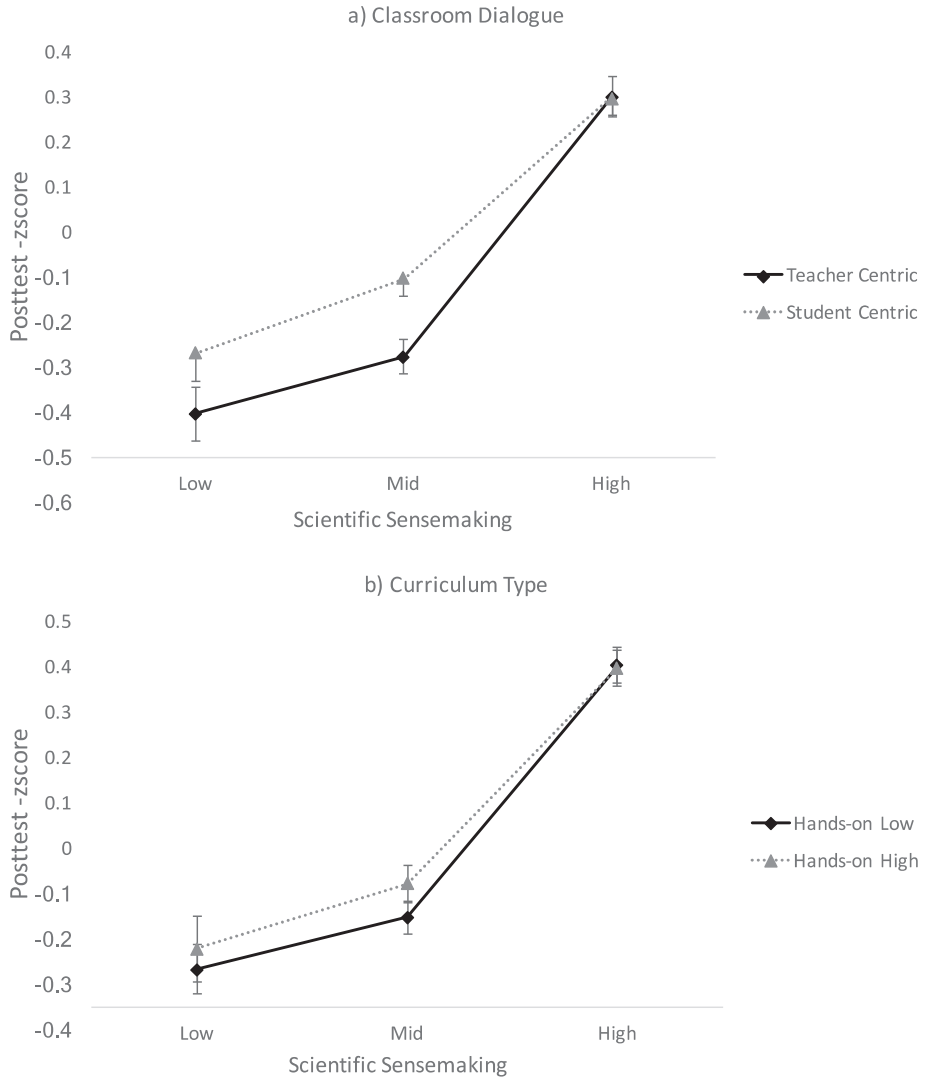


Fig. 3. Estimated marginal means of adjusted post-test scores, controlling for pre-test z-scores, for the interaction of binned scientific sensemaking and (a) observed classroom dialogue type or (b) teacher-reported curriculum materials used.

Table 11
Polychoric correlations between scientific sensemaking sub-construct scores.

	Design of Experiments	Explanations & Argumentations	Nature of Science
Asking Questions	0.69	0.53	0.72
Design of Experiments	–	0.68	0.81
Explanations & Argumentations		–	0.72

ability estimate on the Dolphins scenario. Below we describe the process (Kolen & Brennan, 2014):

1. Each item on both test forms was scored as either correct (1) or incorrect (0).
2. Using a 2-parameter model in IRTpro, we calculated estimates for each item's difficulty (how difficult the item is to answer correctly) and discrimination (how well the item distinguishes high ability students from low ability students). This was done for both test forms at the same time, and inserted into the model as if they were all a part of the same test given all at the same time. This allows for the item parameters (difficulty & discrimination) for both assessments to be estimated on the same scale.
3. Once the initial estimates were calculated, we fixed, or "anchored", the item parameters (discrimination and difficulty) for all the items on the Dolphins scenario. Using these as the anchor, we estimated the score for each respondent on the Monkeys scenario using the Newton-Rapson Method, an iterative process for finding the overall best model fit. Said another way, given the ability estimate for each student on the Dolphins scenario, we can calculate the probability they get the correct answer for each of the items on the Monkeys scenario. Summing these probabilities gives us the ability estimate on the Monkeys scenario and a way to link ability estimates across the two forms. This method allows us to convert scores between forms of two assessments measuring the same construct. A second order equation was fit to the data ($R^2 = 0.99$) allowing for conversion of scores across the two scenarios.

Instructional impact on scientific sensemaking. With linking scores established, we then could examine how much scientific sensemaking changes through instruction. To act as a resource for learning, it needs to have some stability over time. But if it does not change at all during instruction (especially for cases of instruction that was inquiry-based), it has little practical value for educators. In other words, we wish to assess whether SSM is semi-malleable. This semi-malleability can be shown through test-retest reliability and analysis of change with

Table 12
Models testing the effect of scientific sensemaking individual sub-constructs on knowledge posttest scores.

	# of Items per Sub-construct	Model 9: Sensemaking Sub-Constructs	
		β	p
Knowledge Pretest Score		0.40	< 0.001
Asking Questions	2	0.09	< 0.001
Design of Experiments	3	0.08	< 0.001
Arguing with Evidence	3	0.12	< 0.001
Nature of Science	4	0.12	< 0.001
R^2 conditional		0.42	
ICC		0.06	

instruction. In terms of test-retest reliability, the correlation between pre and post sensemaking scores is moderate (r = 0.58), and mean scores changed from 4.2 to 6.9 (in adjusted scores in units of the Monkeys assessment), representing a Cohen's *d* of 0.51. This level of stability is high enough to argue for its stability (i.e., that both test forms are measuring the same construct and students carry their initial SSM abilities throughout the semester of instruction) but not so high that one could argue that students' scientific sensemaking did not change during those four months.

To test whether scientific sensemaking appears to be modifiable through classroom experience, we conducted a multiple linear regression with cognitive-behavioral and affective engagement predicting post-scientific sensemaking scores while controlling for pre-scientific sensemaking scores (see Table 13). Cognitive-behavioral engagement accounts for a statistically significant 5% of the variance in post-scientific sensemaking scores, but affective engagement was not statistically significant. The results are as expected since cognitive/behavioral engagement in classroom activities are more likely to be related to sensemaking skill development than affect during the lesson which may be influenced from on or off task behaviors.

4. General discussion

We sought to unpack the generality of foundational scientific practices for learning science content knowledge across instructional contexts and whether students' level of mastery of these practices can be meaningfully characterized with an overarching scientific sensemaking level, rather than only measuring isolated practices. We provided evidence in support of our hypothesis that: (1) there is a general overall mastery level of core scientific practices; (2) it supports learning in new science content areas; and (3) it may be modifiable through instruction. In the next section, we discuss these findings and the limitations of our study. We then go on to discuss the instrument itself and scientific sensemaking's relationship to PISA scientific literacy competencies and to the NGSS. We then describe the implications of this instrument and these findings for the field, specifically for the design of science learning experiences that focus on building a sensemaking activity.

Table 13
Models testing the effect of engagement on scientific sensemaking posttest scores, controlling for pretest scientific sensemaking scores.

	Model 10: Baseline Model		Model 11: + Engagement	
	β	p	*β*	p
Scientific Sensemaking	0.51	< 0.001	0.55	< 0.001
Cognitive/Behavioral Engagement			0.09	< 0.001
Affective Engagement			0.01	0.57
R^2 conditional	0.35		0.40	
ICC	0.08		0.07	

4.1. Do variations in initial SSM abilities predict science content learning gains across diverse science contents, learners, and instructional methods?

Across our diverse sample of learners and a diverse set of science topics, we found that scientific sensemaking is a positive covariate of science learning. Learners both in biology and non-biology classrooms with higher scientific sensemaking abilities demonstrated greater increases in science content knowledge even though the SSM scenarios involved biology content. This suggests that scientific sensemaking may be a broadly transferable skill across content areas within science. It may be that the focus on practices as sensemaking (rather than rituals or disconnected processes) in particular facilitates transfer (Nokes-Malach & Mestre, 2013).

We also tested for possible interactions with learner characteristics. As adolescence is a stage of great developmental change, for example, critical reorganization of regulatory systems (Steinberg, 2005) and large sex-specific physiological changes (Staurt, Shock, Breckenridge, & Vincent, 1953); it is possible these changes have influences on the relationship between SSM and content learning. We indeed found some main effects of race/ethnicity on post-content knowledge scores that are consistent with race/ethnicity achievement gaps reported previously in the literature (e.g., Quinn & Cooc, 2015). However, we found no evidence of interactions of scientific sensemaking with sex, age, or race/ethnicity. This may suggest a broad generality of SSM to be productive for learner's across learner characteristics typically implicated in science learning outcomes. Such findings are important for building a vision of effective science learning that is not culturally biased or prioritizing ways of reasoning that are not broadly productive; it also speaks against perceptions students, teachers, and parents might have regarding learning styles that specific students might not able to learn via engaging in the practices and instead would benefit more from traditional instruction.

Interestingly, SSM accounted for a higher percent of the variance than did classroom clustering, which embeds the effects of the many important variations among teachers (years teaching, instructional approaches, amount of content covered, etc.). We specifically tested whether distinct instructional styles could position students to use their scientific sensemaking skills more often and therefore demonstrate greater learning gains. We did find a main effect of classroom discourse on science learning; learners in student-centric classrooms tended to have greater learning gains than learners in teacher-centric classrooms—note this learning effect cross-validates the measure of instructional style. However, we found no interaction effects with classroom discourse or instruction type (hands-on vs. traditional). This suggests that any benefit of scientific sensemaking in supporting science learning is the same across instructional content, discourse, or practice because this skill is consistent across these environments.

The finding of benefits across instructional style is theoretically and practically important. From a theoretical perspective, it provides support for conceptualizations of scientific sensemaking as a general and broad approach to learning, rather than a simple collection of skills that would each be independently acquired and individually useful in particular learning contexts. For example, the ability to design studies would only be relevant to cases in which students are allowed to design studies. By contrast, approaching science learning as sensemaking and having a broad set of competencies in the practices appears to enable students to better understand that content whether they are actively creating and analyzing data or whether they are reading about or listening to presentations about data patterns and explanations from history.

From a practical perspective, the broad usefulness of SSM suggests that focusing on improvements in SSM at younger grades will benefit learners for many years even if the older grades are slower to change to newer instructional approaches, or more generally that learners who

experience alternations between reform-oriented instruction and traditional instruction should show benefits during both traditional and reform instruction from earlier gains in SSM.

4.2. Can SSM abilities be improved through high cognitive engagement in science instruction?

We argue that our scientific sensemaking instrument is assessing a malleable set of skills and understandings rather than a relatively fixed ability like general intelligence. If we were measuring general intelligence, we would have found higher correlation scores between pre and post sensemaking scores. Further, we found that cognitive/behavioral engagement in classroom activities was predictive of increases in scientific sensemaking scores, implying that they can be changed through cognitive and behavioral practice in a science learning setting. Similarly, the finding that affective engagement was not predictive of changes in scientific sensemaking provides further discriminative validity to the hypothesized causal relationship. Note, however, that we do not argue that affective engagement is unimportant. Rather, we argue that affective engagement is important for motivational development, such as growing (or declining) interest in science (Fredricks et al., 2004; Vincent-Ruz & Schunn, 2017). Collectively, these results point to the semi-malleable nature of the scientific sensemaking measure and differentiates the underlying construct from general intelligence.

5. Limitations

The current study has several limitations that constrain the inferences we can draw from our results. Here we discuss those limitations and the influence they have on potential inferences. We developed our measure by selecting a number of scientific practices that have previously been argued to be important for students' science learning. However, we did not include all practices implicated in prior literature, most saliently the understanding and creation of scientific models (Lehrer & Schauble, 2006). Given that scaffolded instruction around modeling has been shown to further student knowledge acquisition (Mulder, Bollen, de Jong, & Lazonder, 2016), we intend to include this practice in future versions to have a broader measure of the construct of scientific sensemaking.

The sample, although large and including of a diverse population of students, did not include a population that was fully representative of the US in that it included schools from only two regions in the US, it sampled no rural schools, and only included public schools. Levels of sensemaking skills might have been substantially lower or higher in other regions or school contexts, and may require a broader range of item difficulties in the assessments. Similarly, the generalization of the findings to instruction outside of the US is not established. Theoretically, the components of SSM are based on a broadly-shared understanding of reform science instruction (see discussion of overlap with PISA scientific literacy competencies below). But the details of the SSM assessments may not fit the expectations and common content understandings of learners in other countries. Early pilot work in English-speaking Africa suggests some wording changes are required to the scenarios based on cultural context differences, but the overall scenarios were meaningful to those learners.

Analyses conducted in this study were fundamentally correlational in nature and, therefore, do not strongly establish causality. We did rule out reverse causal effects such as content knowledge being the cause rather the consequence of scientific sensemaking by collecting the variables in temporal sequential order. However, only a few other student characteristics were included in the presented regressions. The presented study also collected many other student level covariates which may be drivers of student learning, such as motivational variables (i.e., fascination with science, valuing science for self and society, competency beliefs in science), characteristics of the home support for learning (e.g., whether the family has physical resources for learning, whether the family supports learning), and optional science learning experiences outside of class (e.g., summer camps, museum visits, gardening). While a full discussion of these variables is outside the scope of this article, we have tested the extent to which any of these variables were functioning as drivers of student learning. Most variables had no relationship to content learning either on their own or with SSM included as a predictor (Chung, Cannady, Schunn, Dorph, & Vincent-Ruz, 2016; Dorph et al., 2016, Dorph, Bathgate, & Schunn, 2017, Vincent-Ruz & Schunn, 2017, Bathgate & Schunn, 2017a). The one exception is competency beliefs, which also predict content learning, but at a relatively weak level that leaves a large predictive role for SSM.

There are also limitations on the inferences we can make about how broadly SSM can support learning, both in content and across settings. For example, the study sites were finite with a fairly narrow range of instructional content (i.e., 6th and 8th grade, in three school districts). This puts limits on inferences regarding any role SSM can play in predicting learning in other learning environments, especially for tasks that require additional content knowledge. That is, we cannot say how well SSM can support a learner to succeed in a learning environment that requires content knowledge beyond what they currently hold. It is unclear how well and across what domains SSM would support "figuring out" the additional necessary content knowledge.

Further, performance on any assessment is at least partially driven by youth motivation for success on the measure and test taking abilities. Given that all measures used in this study were collected using the same approach, the influence of motivation and test taking abilities is likely to uniformly influence scores and may play a role in the underlying relationships found. This is, at least in part, mitigated by using classroom content linked assessments for knowledge gain and a relatively short, meaning fairly low burden, assessment of SSM.

5.1. The scientific sensemaking instrument

The development process and psychometric characteristics support the use of the scientific sensemaking instrument to assess middle school student's levels of scientific sensemaking. The instrument's reliance on common, rather than rare or specific, content knowledge makes it an incomplete measure for assessments of student ability to integrate practices with difficult content, but the instruments' integration of many of the practices (within NGSS or PISA scientific literacy competencies) and content knowledge can serve as a model for the development of such assessments that focus on specific and targeted content knowledge. Further, the results of student performance on an SSM instrument could be used to guide instruction. Recognizing that this tool lightly measures across the sub-constructs of scientific sensemaking, it remains an open question if a broader assessment, one that sampled the sub-constructs more thoroughly, would be helpful to understand scientific sensemaking's role in the learning process while not overly burdening the respondents.

Future versions of this tool will include items assessing skills and practices associated with the roles of models in scientific sensemaking. Further, more difficult versions of the assessment will be developed to avoid ceiling effects and extend the range of use of the assessment to high school students.

5.2. Scientific sensemaking, scientific literacy, and the next generation science standards

We argue that a learner who is strong in scientific sensemaking is able to understand the scientific practices and the role these practices play in understanding of natural and physical phenomena (i.e., the nature of science). Part of arguing for the necessity of each of these practices and NOS requires testing whether each of them was independently predictive of science learning. We calculated sub-scores for each of the tested sub-constructs and tested whether they were

Table 14
Comparison of PISA (2015) Competencies for Scientific Literacy, NGSS practices, and the SSM sub-constructs.

PISA Scientific Literacy Competencies	NGSS Practices	Scientific Sensemaking Sub-Constructs
	Asking Questions	Asking Good Questions
Evaluate and Design Scientific Enquiry	Planning and Carrying out Investigations	Designing Investigations
Interpret Data Scientifically	Analyzing and Interpreting Data	Interpreting Data Tables & Graphs
Explain Phenomena Scientifically	Constructing Explanations	Seeking Mechanistic Explanations
Interpret Evidence Scientifically	Engaging in Argument from Evidence	Engaging in Argumentation about Scientific Ideas
(Embedded with the Knowledge dimension)	(Embedded with Practices)	Understanding the Changing Nature of Science
	Developing and Using Models	(Future Versions)
	Using Mathematics and Computational Thinking	
	Obtaining, Evaluating, and Communicating Information	

predictive of posttest content knowledge scores while controlling for pretest content scores. All of these sub-constructs were statistically significant predictors of learning gains. Additional research would be needed to examine which other possible sub-constructs would also predict content learning.

The PISA 2015 assessment of scientific literacy (OECD, 2017) identified several general competencies, in addition to assessing knowledge and attitudes in several different contexts. Table 14 shows the relationship between the SMM subconstructs and the three broad competencies assessed in PISA. We divided the interpret data and evidence scientifically into the two components of data and evidence because that are separated in SSM and NGSS. Note the absence of asking questions and understanding the nature of science with the PISA framework. The nature of science is included with the knowledge dimension of PISA, but also to some extent with the evaluate aspect of evaluate and design scientific inquiry. Asking questions was explicitly part of the 2000 and 2006 PISA frameworks (OECD, 2017).

The Next Generation Science Standards identify eight practices that should be the aim of science instruction. These practices include asking questions and defining problems, developing and using models, planning and carrying out investigations, analyzing and interpreting data, using mathematics and computational thinking, constructing explanations and designing solutions, engaging in argument from evidence, and obtaining, evaluating and communicating information. As shown in Table 14, there is substantial overlap between these recommended practices and the currently-tested components of scientific sensemaking. NGSS and SSM have a key similarity of purpose: to support student learning of science content. But the two also have unique purposes that could lead to different components: the NGSS list is also meant to include all the core adult science practices that should be broadly shared to understand what science is and be able to participate in science (even though the practices should be gradually developed throughout K-12 schooling) whereas the SSM list was developed with the goal of focusing on the core components that are most needed in early science learning. For example, advanced mathematics and computational thinking are important skills for science learning in some parts of high school science and beyond, but may not be as useful for science instruction at earlier grades given the strong focus on qualitative science and more basic mathematics such as proportional reasoning in the earlier grades.

As noted earlier, the currently included sub-components of SSM were not meant to be a strong claim to exhaustively list all of the important elements for supporting science learning. Further research will be required to see whether the other sub-constructs are important in early science. We suspect that developing and using models (Lehrer & Schauble, 2000) is an important scientific skill that could be included in our instrument and we plan to incorporate modeling into future versions. In general, there can be no claim that the current list of sub-constructs included in SSM is an exhaustive list because the importance of other practices and understandings have not been tested in terms of (1) coherence with the current sub-constructs in SSM and (2) their broad support of science content learning in K-12 science.

Finally, similar to PISA scientific literacy competencies, NGSS does not include nature of science as a separate scientific practice, but rather recognizes that the nature of science is integrated into the active work entailed in the rest of the practices. This is an important difference between a list of practices and how we conceive scientific sensemaking. That is, scientific sensemaking is an ongoing process seeking coherence and meaning through construction and reconstruction of explanations across multiple representations and knowledge. This process requires more than practices; it also requires skills and knowledge, including epistemological understandings of the process of knowledge generation. We call the nature of science out separately in part because we think is essential for students' understanding not only of scientific content but also of the scientific practices that they need to use to engage in inquiry (Duschl & Grandy, 2012; Osborne et al., 2003; Sadler, Chambers, & Zeidler, 2004). In addition, we also found that students who had stronger grasps of the nature of science also had larger increases in content knowledge than their peers who did not have this understanding, even when controlling for their mastery of these other components.

5.3. Implications

The work presented here offers evidence of the role scientific sensemaking can play in a classroom. Scientific sensemaking includes many of the practices that are targets for reform instruction. That alone makes it worthy of investigation, assessment, and curriculum development. The fact that a student who can engage in scientific sensemaking is actually better positioned to learn scientific content only adds to scientific sensemaking's centrality in the modern science classroom. Teachers should be aiming to support the acquisition of scientific sensemaking skills in their students, and develop lessons that encourage students to engage in sensemaking activities as they explore new content. This is not a suggestion to teach SSM skills separately from content, but rather, that in developing SSM skills in a particular content area should then transfer to ability to use them for learning new scientific content.

The instrument may also be useful for informal educators, where scientific sensemaking skills are common targeted outcomes, rather than specific bits of content knowledge. In these settings, informal environments might be able to demonstrate impact on a construct that is useful for both the understanding of science and for the process of learning more science. While it is likely that a longer assessment or dynamic items and simulations could improve the precision of a measure of scientific sensemaking, the results here demonstrate that it is possible to measure scientific practices in content through relatively short pencil paper instruments. Short, static instruments are generally easier to administer, making them useful for a broader range of learning environments. Given that the role of scientific sensemaking in the process of learning content was independent of classroom discourse or classroom instructional practice, having an instrument that can be used in a variety of settings is appropriate.

Going forward, we wish to develop more scenarios, include items on

the use of models as a sensemaking activity and extend the range of difficulty to broaden the appropriate audience for this tool. We also aim to vet (i.e. support a validity argument) the items for use in with a broader audience and ensure that the items function similarly across the range of individuals in the population.

Acknowledgements

This work was supported by the National Science Foundation under Grant #1348666 and The Gordon and Betty Moore Foundation under Grant #4250.

References

American Association for the Advancement of Science (AAAS). (1994). Benchmarks for science literacy. Oxford University Press.

Allison, A., & Shrigley, R. L. (1986). Teaching children to ask operational questions in science. *Science Education, 70*(1), 73–80. https://doi.org/10.1002/sce.3730700109.

Andriessen, D. (2006). Combining design-based research and action research to test management solutions. Presented at the 7th World Congress Action Research.

Apedoe, X., & Ford, M. (2009). The empirical attitude, material practice and design activities. *Science & Education, 19*(2), 165–186. https://doi.org/10.1007/s11191-009-9185-7.

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of Social and Clinical Psychology, 4*(3), 359–373.

Bathgate, M., Crowell, A., Schunn, C., Cannady, M., & Dorph, R. (2015). The learning benefits of being willing and able to engage in scientific argumentation. *International Journal of Science Education, 37*(10), 1590–1612.

Bathgate, M. E., Schunn, C. D., & Correnti, R. (2014). Children's motivation toward science across contexts, manner of interaction, and topic. *Science Education, 98*(2), 189–215.

Bathgate, M., & Schunn, C. (2017a). Factors that deepen or attenuate decline of science utility value during the middle school years. *Contemporary Educational Psychology, 49*, 215–225.

Bathgate, M., & Schunn, C. (2017b). The psychological characteristics of experiences that influence science motivation and content knowledge. *International Journal of Science Education, 39*(17), 2402–2432.

Bell, P. (2004). Promoting students' argument construction and collaborative debate in the science classroom. *Internet Environments for Science Education,* 115–143.

Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education, 94*(5), 765–793.

Berland, L. K., Schwarz, C. V., Krist, C., Kenyon, L., Lo, A. S., & Reiser, B. J. (2015). Epistemologies in practice: Making scientific practices meaningful for students. *Journal of Research in Science Teaching.*.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational researcher, 18*(1), 32–42.

Byrne, B. M. (2001). *Structural Equation Modeling With AMOS.* Psychology Press.

Carey, S. (2000). Science education as conceptual change. *Journal of Applied Developmental Psychology, 21*(1), 13–19. https://doi.org/10.1016/S0193-3973(99)00046-5.

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in higher education, 47*(1), 1–32.

Chin, C., & Osborne, J. (2008). Students' questions: A potential resource for teaching and learning science. *Studies in Science Education, 44*(1), 1–39. https://doi.org/10.1080/03057260701828101.

Chung, J., Cannady, M. A., Schunn, C., Dorph, R., & Vincent-Ruz, P., (2016). Measures Technical Brief: Scientific Sensemaking. Retrieved from: http://www.activationlab.org/wp-content/uploads/2016/02/Sensemaking-Report-3.2-20160331.pdf.

Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163–183.

Costello, A. B., & Osborne, J. W. (2011). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research, & Evaluation, 2005*, 10 10, 7.

Cuccio-Schirripa, S., & Steiner, H. E. (2000). Enhancement and analysis of science question level for middle school students. *Journal of Research in Science Teaching, 37*(2), 210–224. https://doi.org/10.1002/(SICI)1098-2736(200002)37:2<210::AID-TEA7>3.0.CO;2-I.

Danielak, B. A., Gupta, A., & Elby, A. (2014). Marginalized identities of sense-makers: Reframing engineering student retention. *Journal of Engineering Education, 103*(1), 8–44.

Dorph, R., Bathgate, M. E., Schunn, C. D., & Cannady, M. A. (2018). When I grow up: The relationship of science learning activation to STEM career preferences. *International Journal of Science Education, 40*(9), 1034–1057.

Dorph, R., Cannady, M. A., & Schunn, C. (2016). How science learning activation enables success for youth in science learning. *Electronic Journal of Science Education, 20*(8), 49–85.

Driver, R., Asoko, H., Leach, J., Scott, P., & Mortimer, E. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher, 23*(7), 5–12.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087.

Duncan, R. G., & Tseng, K. A. (2010). Designing project-based instruction to foster generative and mechanistic understandings in genetics. *Science Education, 95*(1), 21–56. https://doi.org/10.1002/sce.20407.

Duschl, R. A., & Grandy, R. (2012). Two views about explicitly teaching nature of science. *Science & Education, 22*(9), 2109–2139. https://doi.org/10.1007/s11191-012-9539-4.

Erduran, S., & Jiménez-Aleixandre, M. P. (2007). In S. Erduran, & M. P. Jiménez-Aleixandre (Eds.). *Argumentation in science education*Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-6670-2.

Ford, M. J. (2012). A dialogic account of sense-making in scientific argumentation and reasoning. *Cognition and Instruction, 30*(3), 207–245.

Finn, J. D., Pannozzo, G. M., & Voelkl, K. E. (1995). Disruptive and inattentive-withdrawn behavior and achievement among fourth graders. *The Elementary School Journal, 95*(5), 421–434.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109.

Fredricks, J., McColskey, W., Meli, J., Mordica, J., Montrosse, B., & Mooney, K. (2011). Measuring Student Engagement in Upper Elementary through High School: A Description of 21 Instruments. Issues & Answers. REL 2011-No. 098. Regional Educational Laboratory Southeast.

Gervais, W. M. (2015). Override the controversy: Analytic thinking predicts endorsement of evolution. *Cognition, 142*, 312–321. https://doi.org/10.1016/j.cognition.2015.05.011.

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16*(3), 235–266.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling a Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118.

Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing, & T. M. Haladyna (Eds.). *Handbook of Test Development* (pp. 131–153). Mahwah, NJ: Erlbaum.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating Measures of Performance. *Educational Measurement: Issues and Practice, 18*, 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x.

Kapon, S. (2017). Unpacking sensemaking. *Science Education, 101*(1), 165–198.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1–48. https://doi.org/10.1207/s15516709cog1201_1.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices.* Springer Science & Business Media.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning.* Mit Press.

Kuhn, D. (1992). Thinking as argument. *Harvard Educational Review, 62*(2), 155–179.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!. *Intelligence, 14*(4), 389–433.

Laugksch, R. C., & Spargo, P. E. (1996). Construction of a paper-and-pencil test of basic scientific literacy based on selected literacy goals recommended by the American Association for the Advancement of Science. *Public Understanding of Science, 5*(4), 331–359.

Lawson, A. E., & Weser, J. (1990). The rejection of nonscientific beliefs about life: Effects of instruction and reasoning skills. *Journal of Research in Science Teaching, 27*(6), 589–606.

Lehrer, R., & Schauble, L. (2000). *Modeling in mathematics and science* (pp. 101–159). New Jersey, NJ: Lawrence: Erlbaum.

Lehrer, R., Schauble, L., & Petrosino, A. J. (2001). Reconsidering the role of experiment in science education. Designing for Science: Implications From Everyday, Classroom, and Professional Settings, 251–278.

Lehrer, R., & Schauble, L. (2006). *Cultivating model-based reasoning in science education.* Cambridge University Press.

Leighton, J. P. (2004). The assessment of logical reasoning. *The Nature of Reasoning,* 291–312.

Michael, J. (2006). Where's the evidence that active learning works? *Advances in Physiology Education, 30*(4), 159–167.

Mulder, Y. G., Bollen, L., de Jong, T., & Lazonder, A. W. (2016). Scaffolding learning by modelling: The effects of partially worked-out models. *Journal of Research in Science Teaching, 53*(3), 502–523.

Mullis, I. V., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). TIMSS 2003 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades. ERIC.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142.

National Research Council (2007). *Taking science to school: Learning and teaching science in grades K-8.* National Academies Press.

National Research Council (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* National Academies Press.

Nokes-Malach, T. J., & Mestre, J. P. (2013). Toward a model of transfer as sense-making. *Educational Psychologist, 48*(3), 184–207.

National Science Teachers Association (NSTA, 2000). The Nature of Science—A Position Statement of NSTA. Washington, DC.

OECD (2017). PISA 2015 Science Framework. In: PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving, OECD Publishing, Paris. https://doi.org/10.1787/9789264281820-3-en.

Osborne, J., Collins, S., Ratcliffe, M., Millar, R., & Duschl, R. (2003). What "ideas-about-

science" should be taught in school science? A Delphi study of the expert community. *Journal of Research in Science Teaching, 40*(7), 692–720.
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.
Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3–14.
Quinn, D. M., & Cooc, N. (2015). Science achievement gaps by gender and race/ethnicity in elementary and middle school: Trends and predictors. *Educational Researcher, 44*(6), 336–346.
Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education, 92*(3), 499–525. https://doi.org/10.1002/sce.20264.
Sadler, P. M., Coyle, H., Miller, J. L., Cook-Smith, N., Dussault, M., & Gould, R. R. (2010). The Astronomy and Space Science Concept Inventory: Development and validation of assessment instruments aligned with the K-12 national science standards. *Astronomy Education Review, 8*(1), 010111.
Sadler, T. D. (2009). Situated learning in science education: Socio-scientific issues as contexts for practice. *Studies in Science Education, 45*(1), 1–42.
Sadler, T. D., Chambers, F. W., & Zeidler, D. L. (2004). Student conceptualizations of the nature of science in response to a socioscientific issue. *International Journal of Science Education, 26*(4), 387–409.
Sadler, T. D., & Zeidler, D. L. (2005). Patterns of informal reasoning in the context of socioscientific decision making. *Journal of Research in Science Teaching, 42*(1), 112–138.
Sampson, V. D., & Clark, D. B. (2006). Assessment of argument in science education: a critical review of the literature. *Presented at the Proceedings of the th international conference on Learning sciences* (pp. 655–661). Bloomington, Indiana: International Society of the Learning Sciences.
Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education, 92*, 447–472. https://doi.org/10.1002/sce.20276.
Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction, 23*(1), 23–55. https://doi.org/10.1207/s1532690xci2301_2.
Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32*(1), 102.
Schweingruber, H., Keller, T., & Quinn, H. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* National Academies Press.
Songer, N. B., Kelcey, Ben, & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching, 46*(6), 610–631. https://doi.org/10.1002/tea.20313.
Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*(1), 4–28.
Staurt, H. C., Shock, N. W., Breckenridge, M. E., & Vincent, E. L. (1953). *Physical Growth and Development*.
Steinberg, L. (2005). Cognitive and affective development in adolescence. *Trends in Cognitive Sciences, 9*(2), 69–74.
Stewart, B. M., Cipolla, J. M., & Best, L. A. (2013). Extraneous information and graph comprehension. *Campus-Wide Information Systems, 26*(3), 191–200. https://doi.org/10.1108/10650740910967375.
Toulmin, S. E., & Rieke, R. D. J. (1984). An introduction to reasoning.
Triona, L. M., & Klahr, D. (2010). Point and click or grab and heft: Comparing the influence of physical and virtual instructional materials on elementary school students' ability to design experiments. *Cognition and Instruction, 21*(2), 149–173. https://doi.org/10.1207/S1532690XCI2102_02.
Vincent-Ruz, P., & Schunn, C. D. (2017). The increasingly important role of science competency beliefs for science learning in girls. *Journal of Research in Science Teaching, 54*(6), 790–822.
Warren, B., Ballenger, C., Ogonowski, M., Rosebery, A. S., & Hudicourt-Barnes, J. (2001). Rethinking diversity in learning science: The logic of everyday sense-making. *Journal of Research in Science Teaching, 38*(5), 529–552.
Zhou, S., Han, J., Koenig, K., Raplinger, A., Pi, Y., Li, D., et al. (2016). Assessment of scientific reasoning: The effects of task context, data, and design on student reasoning in control of variables. *Thinking Skills and Creativity, 19*, 175–187. https://doi.org/10.1016/j.tsc.2015.11.004.
Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172–223. https://doi.org/10.1016/j.dr.2006.12.001.
Zimmerman, H. T., Reeve, S., & Bell, P. (2010). Family sense-making practices in science center conversations. *Science Education, 94*(3), 478–505.