JOURNAL OF RESEARCH IN SCIENCE TEACHING

Research Article

The Increasingly Important Role of Science Competency Beliefs for Science Learning in Girls

Paulette Vincent-Ruz and Christian D. Schunn

Learning Research and Development Center, University of Pittsburgh, Pittsburgh, Pennsylvania

Received 28 August 2015; Accepted 3 January 2017

Abstract: The number of women studying STEM careers and pursuing graduate degrees has not changed in the last decade (National Student Clearinghouse Research Center, 2015; Science & Engineering Degree Attainment: 2004–2014). Most prior research to explain this problem has focused on the topics of identity, access, pedagogy, and choice (Brotman & Moore, 2008; Journal of Research in Science Teaching, 45, 971–1002). Additional research is needed on how internal and external factors interact with one another to demotivate girls and young women from pursuing science careers. Here, we show how girls' competency beliefs are an essential foundation for science content learning during middle school and how these effects of competency beliefs are mediated by in and out-of-school factors. We recruited over 2,900 6th and 8th grade students from two different regions in the United States. At two different time points, students completed surveys asking about their stance toward science such as competency beliefs in science, willingness to engage in argumentation, and choice preferences toward optional science experiences. We also collected a reasoning ability measure, and pre- and post-tests on science content knowledge. Moreover, students also reported on their cognitive behavioral engagement during a sampled science class on two separate occasions. Multiple regression and mediation analyses show that as boys grow older, their willingness to engage in argumentation and to participate in science experiences suppresses the role of competency beliefs on their learning science content. By contrast, as girls grew older, they showed an increasing need to have high competency beliefs to achieve strong content learning gains. Our results demonstrate that despite girls' willingness to participate in scientific argumentation and to take part in science experiences, they probably do not receive enough support in their environment to access the benefits of these experiences, and hence they have a stronger need to have high competency beliefs in order to achieve significant growth in science learning. J Res Sci Teach 9999:XX–XX, 2017
Keywords: competency beliefs; gender; science learning; development

The challenge of bringing women into science is hardly new. Over the last 20 years, many researchers and policymakers have sought causes and effective solutions for the constant scarcity of women in many scientific careers (Chipman, Krantz, & Silver, 1992; Clark Blickenstaff, 2006; National Student Clearinghouse Research Center, 2015; Rayman & Brett, 1995). Despite these efforts, the rate at which women graduate in the United States with a college science, technology, engineering, or math (STEM) degree has made little progress overall. While social and biological sciences now have women and men graduating in similar numbers, areas such as engineering still

Correspondence to: P. Vincent-Ruz; E-mail: pav22@pitt.edu
DOI 10.1002/tea.21387
Published online in Wiley Online Library (wileyonlinelibrary.com).

have large imbalances with only 20% women in graduating classes (National Student Clearinghouse Research Center, 2015). This gender gap is most prominent in higher education, but the gap also occurs at other life stages, suggesting a number of underlying factors.

Early on, in the elementary years, girls outperform boys in science and mathematics (Pomerantz, Altermatt, & Saxon, 2002). But then a gap favoring boys develops by the end of middle school (P. A. Muller, Stage, & Kinzie, 2001). This transition from relative strength to relative weakness suggests that middle school may be a critical time for girls, with potentially long-term consequences for their relationship with science. Around the same time period, there is also a significant drop in interest in science (Maltese & Tai, 2009; Tucker, Hanuscin, & Bearnes, 2008). These motivational changes may be implicated in the relative decline in abilities.

There are three popular accounts for the gender gap in science skills and knowledge: (i) males are predisposed to learn about mechanical systems because they focus on objects at an early age (Baron-Cohen, 2003; Leinbach, Hort, & Fagot, 1997); (ii) men have greater spatial and numerical abilities (Caplan, MacPherson, & Tobin, 1985; Kimura, 2000); and (iii) men are more variable in their cognitive abilities and thereby producing more "brilliant" male scientists (Nowell & Hedges, 1998). However, these accounts are primarily rooted in historical and social perceptions of gender (Kleinman, 1998) rather than in empirical evidence (Spelke, 2005).

One line of empirical investigations of the gap have emphasized a number of potential causes for the gender gap related to how girls engage with science (Brotman & Moore, 2008): (i) equity and access (i.e., reduced access for women) (Carlone, 2004); (ii) curriculum and pedagogy (i.e., biased curricula or instructional support for women) (Hughes, 2000); (iii) the culture of science (i.e., exclusionary of women or their needs) (Legewie & DiPrete, 2014; Moss-Racusin & Dovidio, 2012); and (iv) identity (i.e., women are less likely to find female role models and culturally the stereotype of scientists represents a white male) (Archer et al., 2012). A second line has emphasized motivational factors as a source of gender differences in science (Ghee, Keels, Collins, Neal-Spence, & Baker, 2016; Webb-Williams, 2014). These two different literatures (one of opportunity and another of motivation) rarely speak to one another but that together could provide better theoretical leverage for understanding the gender gap in STEM domains: how motivations shape opportunities to develop abilities and knowledge.

The Central Role of Science Competency Beliefs in Self-Regulated Learning

Underachievement in science is often associated with a lack of underlying skills. However, these outcomes are often the results of a lack of self-regulated learning rather than of lack of ability (Labuhn, Zimmerman, & Hasselhorn, 2010; Mahmoodi, Kalantari, & Ghaslani, 2014). Self-regulated learning is "an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided, and constrained by their goals and the contextual features in the environment" (Boekaerts, Pintrich, & Zeidner, 2007, p. 453). Students' self-regulation in the form of asking questions can stimulate scientific reasoning skills (Chin, 2006) and thereby increase performance in science-related tasks (Chin & Osborne, 2008). Similarly, students self-reflection on scientific concepts leads to better learning outcomes (Beeth, 1998).

These self-regulatory skills depend in part on the beliefs learners hold about themselves (Pajares, 2002). By interpreting the results of their actions in a subject, students refine their beliefs about their capabilities in that subject, which in turn affect the way they approach similar situations in the future. Especially in free choice situations, learners can avoid tasks based on these competency beliefs (Pajares, 2002), also called self-efficacy beliefs (Bandura, 1986). Within the influential Expectancy Values Theory, Eccles, Barber, Updegraff, and OBrien (1998) frame these choices that influence learning as achievement choices (Eccles, 2011; Eccles & Jacobs, 1986).

Competency beliefs have proved to be an important predictor of science achievement in higher education (Andrew, 1998), high school science (Britner, 2008), and middle school science (Britner & Pajares, 2006). Retrospective interviews with men and women selecting STEM careers point to competency beliefs as an important factor in these career choices (Zeldin, Britner, & Pajares, 2008). In middle school, competency beliefs are also an important predictor of participation in optional science learning experiences (Sha, Schunn, & Bathgate, 2015). Furthermore, motivation researchers have established the importance of competency beliefs for self-regulation (Zimmerman, Boekarts, Pintrich, & Zeidner, 2000; Zimmerman & Bandura, 1994), because learners need to feel competent in their abilities during multiples stages of self-regulation, such as during forethought, performance, and self-reflection (see Schunk & Ertmer, 2000, for a review).

Gender Effects on Competency Beliefs in Science

The size and even direction of gender differences in competency beliefs have varied across studies (Webb-Williams, 2014), though there is some agreement that these discrepancies are mostly content related (Pajares, 2002) (e.g., girls have higher competency beliefs in their writing abilities, and boys have higher competency beliefs in their science abilities). Bandura (1986) argued that a student's competency beliefs develop through interpreting information from four different sources that are reviewed below. Each of these sources may be influenced by gender.

Mastery Experiences. Successful outcomes generally raise confidence while those interpreted as unsuccessful generally lower it. More importantly, the accuracy of these interpretations is dependent upon competency beliefs: the more competent a learner believes themselves to be, the more the learner will attribute successful performance to their own competencies versus luck, simplicity of the task, or support from others (Bouffard-Bouchard, Parent, & Larivee, 1991). If girls develop lower competency beliefs in science from another source, then they will make inaccurate interpretations of their successes in later science activities, further lowering their competency beliefs.

Vicarious Experiences. Role models can have large effects on competency beliefs, especially when the learner identifies with them by means of similar characteristics and traits. Due to the traditional image of scientists as white males, many girls do not feel capable of pursuing science careers (Selimbegovic, Chatard, & Mugny, 2007). Additionally, female role models have to engage in a personal connection with the girls in order to effectively shape their beliefs (Buck, Clark, Leslie Pelecky, Lu, & Cerda Lizarraga, 2008), meaning that the lack of local role models (vs. only having access to books highlighting women in science) is particularly problematic.

Social Persuasion. Young children are not adept at making accurate self-assessments of their performance and, therefore, rely on adults as mediators of this process (Pajares, 2002). Parents tend to underestimate the scientific skills of girls (Halpern et al., 2007). Similarly, teachers can convey expectations that mathematics and science will be difficult for girls, and school counselors often discourage girls from pursuing careers in science (Zohar & Bronshtein, 2005). These adult messages have a negative effect on girls' developing competency beliefs development in science, especially when they are too young to make accurate self-assessments.

Psychological States. Anxiety and tension are negative sensations that hinder an expectancy of success in particular activities. Girls are consistently more anxious while performing science-related tasks (Britner, 2008), possibly because of societal expectations set for performance in science (Pajares, 2002).

The literature reviewed above establishes various possible causes of emerging gender differences in competency beliefs. Now, we turn to likely consequences of these gender differences for different aspects of self-regulated learning. In particular, we describe two kinds of choices shaped by competency beliefs that could influence science learning in adolescents (optional science learning and in-class engagement).

Choice Preferences for Optional Science Learning

Cumulative participation in optional science learning activities (e.g., additional classes, science clubs, science camps, relevant fiction, and non-fiction reading) can have a large impact on students' science content knowledge (Feder, Shouse, Lewenstein, & Bell, 2009), but only if learners choose to do these optional activities. Gender differences in choice are found in preferences within science toward particular topics, such a general tendency for girls to prefer biology and boys to prefer physical sciences (Stark & Gray, 1999). But there are also general gender preferences against overall participation in optional science opportunities. As a result, by 6th grade, boys have typically had more optional science experiences than girls (Jones, Howe, & Rua, 2000).

Willingness to take risks increases steadily among men, allowing them to make more choices perceived as risky when they are older; by contrast, women suffer a greater decline in risk-taking choices from the teenage years to age 30 (Marianne, 2011). Choosing science may be thought of as a risk when, due to a lack of female role models, girls feel they have to negotiate their feminine normativity by engaging in an activity that is perceived as inherently male (Kleinman, 1998). Moreover, competency beliefs have been linked to women's disposition to avoid risky decisions or choices (Marianne, 2011).

Although gender differences in competency beliefs are well establish in science, additional research is required to examine whether competency beliefs are particularly problematic for girls in influencing their optional learning choices. That is, we know that, on average, girls tend to have lower competency beliefs than boys in science. However, we do not know whether competency beliefs also shape girls' choices differently than they shape boys' choices (e.g., are competency beliefs even more important for girls in predicting science choices given the greater "risk" associated with making counter-normative science choices for girls?). Furthermore, it is not yet clear that differential participation in optional science learning is sufficiently large to produce the gender gaps in science achievement. Such experiences may be more important for driving changes in interest or identity than in driving changes in science knowledge, which may be more driven by required school experiences.

Cognitive/Behavioral Engagement During Science Class

School engagement is considered critical for avoiding low levels of general academic achievement and school dropout (Fredricks, Blumenfeld, & Paris, 2004). The research literature commonly divides engagement into three different types: behavioral, cognitive, and affective. In order to achieve understanding of the complex science material provided in class, students must engage in a behavioral way (i.e., actually completing the assigned hands-on or written tasks) and in a cognitive way (i.e., minds-on) to think deeply about the science content embedded in the activities (Pintrich, Zusho, Schiefele, & Pekrun, 2001). Although conceptually distinct, in practice, cognitive, and behavioral engagement tend to co-vary to such a high extent in middle school science that we treat them as one combined construct (Bathgate & Schunn, 2016). Less closely aligned is affective engagement, which relates to emotional experiences during learning activities and is predictive of growing (or declining) interest (Fredricks et al., 2004). Thus,

cognitive-behavioral engagement in particular is an important indicator of self-regulated learning, and a focus in our study.

Competency beliefs are thought to be an important predictor of the cognitive-behavioral engagement that students display in the classroom (Pintrich et al., 2001; Pintrich & De Groot, 1990). However, more research is required to investigate whether competency beliefs are equally important for cognitive-behavioral engagement across gender, since more self-regulation may be required when completing work that is counter gender stereotypes. In addition, cognitive-behavioral engagement may be disrupted by worrying about performance in science (Nosek et al., 2009; Shapiro & Williams, 2012).

Research Questions

In this paper, we propose that the gender achievement gap in science that emerges during adolescence is a consequence of girls' competency beliefs having a growing effect on self-regulation in the form of achievement choices (Figure 1), rather than from inherent differences in scientific reasoning abilities. Specifically, we asked the following research questions in a study of 6th and 8th graders:

- RQ 1. To what extent are scientific reasoning abilities higher in boys than in girls in public urban middle schoolers? Based on the literature, we expect to find at most small differences. But given plausible connections between reasoning ability and content learning and between reasoning ability and competency beliefs, it is especially important to first rule out gender differences in reasoning ability in our sample.
- RQ 2. Are competency beliefs more strongly associated with science content learning outcomes for girls than boys? Note that this question is not about whether competency beliefs are higher or lower, but rather about whether they are differentially important for learning.
- RQ 3. Does differential choice of optional science experiences mediate a gendered relationship between competency beliefs and science learning? We consider two possible forms of the effect: competency beliefs may have a greater effect on choice preferences in girls or choice preferences may have a larger effect on science content learning in girls.
- RQ 4. Does differential cognitive-behavioral engagement during science class mediate a gendered relationship between competency beliefs and science learning? Similar to RQ3, we test two possible forms of the effect: competency beliefs may have a greater effect on in-class cognitive-behavioral engagement in girls or in-class cognitive-behavioral engagement may have a larger effect on science content learning in girls?

Methods

Participants

Drawing on data from the ALES14 dataset (Activated Learning Enables Success 2014) of public urban middle school students in the United States, participants were sampled from two different regions in the United States purposely selected to have different diverse urban student populations but also differential cultural emphases on science and technology. One group came from Pittsburgh, a mid-sized city in the Eastern United States with a high proportion of African American residents. A total of 27 sixth grade classes and 21 eighth grade classes from six urban public schools were selected to represent a range of school configurations and student demographics. Two schools had typical middle school arrangements (i.e., 6–8th grade), and four schools included 6–12th grade, three of which were magnet schools focusing on Arts, Languages, and Science and Technology, respectively. The demographics across the schools varied widely:

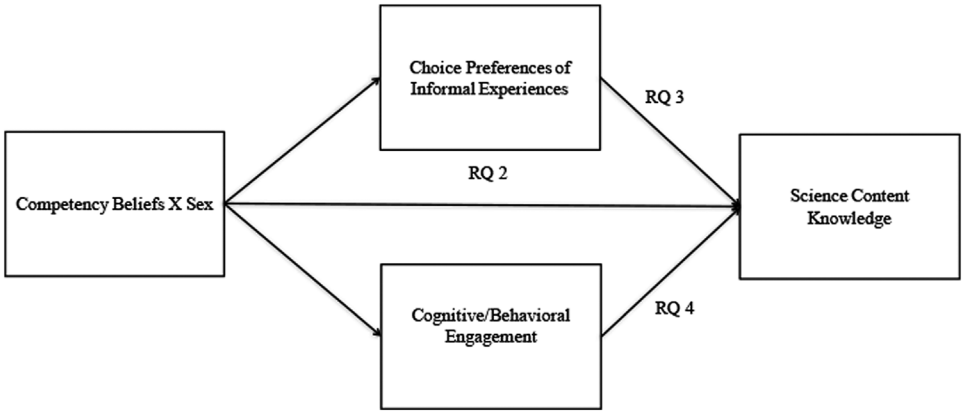

Figure 1. Hypothesized research question path models.

38–84% Free/Reduced Lunch; 32–99% Under-represented Minorities. The second group came from the Bay Area in the Western United States, a region with strong emphasis on technology and innovation, and a high proportion of diverse recent immigrants. A total of 27 sixth grade classes and 32 eighth grade classes were recruited from five urban schools, all in typical middle school arrangements (i.e., 6–8th grade). The demographics across the schools also varied widely: 24–92% Free/Reduced Lunch; 13–94% Under-represented Minorities. Table 1 presents overall demographic characteristics of each group. Overall, the sample is similar on key demographic distributions relevant to science education (sex and race/ethnicity; (Archer et al., 2012, Brown, 2004; Oakes, 1990) of all US urban middle school students, except for a slight over-representation of African Americans and under-representation of Hispanic/Latino and Asians (Lippman, McArthur, & Burns, 2004): 50% White, 25% Hispanic/Latino, 16% African American, 5% Asian, and 1.4% Other. Furthermore, US urban schools vary widely in racial and economic composition, motivating our inclusion of a wide range of demographics across schools, including schools with

Table 1
Participant demographics by location and grade, including number of participants, mean and standard deviation for age, percentage female, and percentage indicating each race/ethnicity

		Age (Years)			Race/Ethnicity				
Location	Grade	*M*	SD	% Female	White (%)	Asian (%)	African American (%)	Hispanic/Latino (%)	Other (%)
Pittsburgh	6th $N = 421$	11.5	0.7	51	47	5	49	8	1
	8th $N = 402$	13.4	0.7	52	55	7	47	11	1
Bay area	6th $N = 520$	11.3	0.6	52	41	14	21	28	1
	8th $N = 655$	13.3	0.6	51	44	14	19	26	1
Total	$N = 1{,}998$	12.3	0.6	51	46	11	31	20	1

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

higher percentages of Hispanic/Latino and Asians. School focus and type of curriculum have also been implicated as important in science learner (Kirschner, Sweller, & Clark, 2010; Marx et al., 2004), and thus it was important to capture diversity of this element as well.

Sample sizes varied across measures due to student absence across data collection points. A total of 1,998 students are used as the primary sample in this study. The percentage of missing item data for all the scales employed had a mean of 0.7% and no higher than 3% for any item. We, therefore, did not use data imputation methods, since they are typically recommended for datasets with an average of 4–15% missing data (Gold & Bentler, 2009). Instead, missing items were dropped from the computation of mean scores, and students simply needed to have at least half the items on a scale for a mean to be computed.

Parents provided written consent for child participation. Even though there was no individual assent procedure for students, no strong pressure as placed on the children to complete the measures. The research methodology for this project was approved by The University of Pittsburgh and University of California, Berkley Institutional Review Boards.

Data Collection Procedure

Students completed surveys across five time points:

(1) Three surveys (competency beliefs, scientific sensemaking, choice preferences) were administered during one class period in early Fall.
(2) Pre-test on content knowledge followed by a demographics survey were administered during another class period, 1 to 2 weeks later.
(3 and 4) Cognitive-behavioral engagement during a given science class was measured at the end of class on two different days in October and November.
(5) The post-test on content knowledge was administered in one class period in late January or early February of that school year. This assessment was part of their grade.

The motivation surveys were administered first to avoid artificial reductions in student's attitudes toward science stemming from low performance on the content pre-test. Similarly, the demographics survey was presented after the pre-test on content knowledge to minimize the effects of stereotype threat (the perceived threat from recognizing that one is a member of a group that is associated with a negative performance in science) on the sampled girls or under-represented minority students' responses (Steele & Aronson, 1995). The cognitive/behavioral engagement survey was applied multiple times at the end of a class to reflect on just that class rather than retrospectively across the whole semester to improve the accuracy of self-report.

Scale Development and Validation

New measures were developed for the larger dataset used in this study because of the following concerns. First, given the diverse population being studies, it was important that all the measures were at an appropriate reading level for middle schoolers of widely varying abilities. Second, the surveys needed to measure the constructs for the general domain of science, not school in general or specific science subjects. Third, to support complex statistical analyses, the psychometrics of the instruments needed to be strong (loaded as a single factor, measuring students well across ability levels). Fourth, and most importantly for our research questions here, the items needed to be clearly understood in construct-relevant ways by middle schoolers of different sex and ethnicities. No prior instrument met all of these criteria. Furthermore, beyond having met these four criteria, we ensured that the instruments were valid on the following

dimensions (Hardesty & Bearden, 2004): (i) items reflected what they were intending to measure as judged by experts and students (face validity); (ii) items exist for each aspect of each construct (content validity); (iii) and the overall scales pass empirical tests of validity (e.g., discriminant, convergent, concurrent, and predictive validity).

Face Validity. Measures were iteratively developed and refined, beginning with items reported in the literature from closely related constructs and adapted to the middle school and science context as needed. We shared our items with a panel of experts in motivation and science learning and improved the scales based on their feedback. Since we developed many different scales that each required different pools of experts, it was not feasible to obtain enough experts on every scale to select items according to a quantitative thresholds (e.g., validated by at least four of six experts). While we considered loosening our definition of experts (e.g., to include senior graduate students), the need to consider adapted items for different age groups led us to be conservative in selection of only senior experts and use their feedback in the following way. In particular, experts completed forms requesting feedback: (i) at the item level, on whether the item matched the motivational construct and on wording appropriateness, for the age group and (ii) at the scale level, on whether the overall scale covered the core aspects of the motivational construct. We also conduced cognitive interviews with students (Desimone & Le Floch, 2004) to ensure the items were being understood by students as intended. These interviews consisted of think-alouds in which approximately four-to-six male and female students from diverse backgrounds read each item, explained in their own words what the item was asking, and explained the reasoning behind their elected response to the item. If students misunderstood the intent of an item or responded using reasons that were inconsistent with the scale's intent, items were modified and new interviews were conducted. This process ensured that the items reflected what they meant to measure both from a literature perspective and from the student population we intended to survey.

Content Validity. Several steps were completed to ensure that the items fully cover the theoretical content domain of a construct. In our selection process, we first surveyed the literature to identify the essential components that our items were required to cover. For example, Competency Beliefs refer to both how an individual rates their skills in science and how confident they are in performing science activities. Therefore, multiple items were selected for each of those two components. Next, we identified items already published on the literature that covered these components and modified them in order to make them age appropriate, content specific and of adequate reading level. Finally, we collected feedback from experts on each construct to ensure full coverage of the domain with the selected and modified items.

Discriminant Validity. Discriminant validity refers to the extent to which scales are separable from related but different theoretical constructs, which is especially important for motivational scales because they are often high correlated with one another such that a scale could potentially be measuring the wrong construct. Theoretically, many of the constructs we are measuring should be related to one another (e.g., competency beliefs should be moderately related to actual abilities, and interest levels with competency beliefs), but not so highly (e.g., close to above 0.8 correlations) such that the construct becomes indistinguishable from these related constructs. It is important to have both constructs measured at the same time point to avoid apparent discrimination stemming from temporal instability. For example, the Competency Beliefs scale was found to be moderately but not highly correlated with science interest (a related but different motivational construct $r = 0.6$; Dorph, Cannady, & Schunn, 2016). Additional discriminant validity correlations for other scales can be found in Table 2.

Table 2
Scale reliabilities, means, standard deviations, and Pearson intercorrelation matrix

Scale	α/Theta	*M*	SD	Competency Beliefs	Scientific Sensemaking	Cog./Beh. Engagement	Choice Preferences	Pre-Test	Post-Test
Competency beliefs	0.85	2.8	0.5	1					
Scientific sensemaking	0.89	57%	25%	0.30	1				
Engagement	0.80	2.9	0.5	0.26	0.10	1			
Choice preferences	0.84	2.4	0.5	0.49	−0.03	0.22	1		
Pre-test	0.6	42%	21%	0.30	0.48	0.08	0.06	1	
Post-test	0.7	49%	20%	0.28	0.51	0.12	0.06	0.56	1

Predictive Validity. Predictive validity refers to how a measure should be able to predict something that is theoretically related as an outcome. These constructs should be measured at different time points to capture an outcome and to not be confused with discriminant validity. For example, competency beliefs were found to be correlated with science choice preferences (an output of competency beliefs $r = 0.5$). Similarly, a measure of scientific reasoning (in this paper, we refer to this construct as scientific sensemaking) should be able to predict how well a person performs in a science test. This aspect of for some of our scales is reflected in the analyses presented in the results section (e.g., in Table 2, for additional predictive validity information; see Dorph et al., 2016; Lin & Schunn, 2016; Sha, Schunn, Bathgate, & Eliyahu, 2016).

Psychometric Properties. In addition to testing validity, we conducted pilot studies with hundreds of students to assess basic psychometrics. In addition to testing scale reliability, item-response theory (IRT) analyses were used to make sure items had varying "difficulty" levels (i.e., measured participants well across ability levels) and that questions were equally relevant by sex and ethnicity. For example, items were removed that showed differential discriminability for girls than boys (see supporting information for DIF analyses on the competency beliefs scale). IRT analyses were also used to ensure the 4-point Likert scale could be treated as an interval scale and simple means across items were appropriate to use. The scale validation process took place across years of iterative work across each of the validity and psychometric checks, in which poorly functioning items were replaced with new items that were then validated at the item and overall scale level. The results presented below reflect the final versions of the scales used in this study.

Scale Reliability and Validity Analyses

Reliability, means, and inter-correlations of the resulting scales are provided in Table 2. EFA and CFA fit statistics as well as IRT analyses are reported for the competency beliefs scale, the central construct in the current study (Table 3). See http://www.activationlab.org/tools/ for confirmatory factor analyses and IRT analyses for the rest of the scales; all of the scales had strong psychometric properties.

The Chronbach alphas for competency beliefs, cognitive-behavioral engagement, and choice preferences are all above 0.75, meaning they are all have acceptable internal consistency. Armor's Theta is reported for scales with dichotomous items (scientific sensemaking and content knowledge tests). Because the content knowledge tests had different forms, we reported a weighted mean across forms. Reliability is typically lower at pre-test on content knowledge given

Table 3
Competency beliefs items, response scoring, EFA factor loadings, and Rasch fit statistics

Item Code	Item	Response Options and Coding	EFA	Rasch Model	
			Factor 1 Loading	Unweighted MNSQ	Weighted MNSQ
CB01	I can do the science activities I get in class.	4 = all the time; 3 = most of the time; 2 = half the time; 1 = rarely	0.66	1.02	1.04
CB02	If I went to a science museum, I could figure out what is being shown in.	4 = all areas; 3 = most areas; 2 = a few areas; 1 = none of it	0.70	0.94	0.94
CB03	I can understand science information on websites for kids my age.	4 = all websites; 3 = most websites; 2 = a few websites; 1 = none of them	0.74	0.94	0.93
CB04	If I did my own project in an after school science club, it would be.	4 = excellent; 3 = good; 2 = ok; 1 = poor	0.71	0.94	0.94
CB05	If I were working on a class science project, I could understand the science in books for adults.	4 = all; 3 = most; 2 = some; 1 = a little	0.71	0.99	0.99
CB06	I think I am very good at: figuring out how to fix a science activity that didn't work.	4 = YES! 3 = yes; 2 = no; 1 = NO!	0.66	0.99	0.99
CB07	I think I am very good at: coming up with questions about science.	4 = YES! 3 = yes; 2 = no; 1 = NO!	0.58	1.11	1.11
CB08	I think I am very good at: doing experiments.	4 = YES! 3 = yes; 2 = no; 1 = NO!	0.62	1.04	1.06

that many students have had no prior exposure to the content. No scales were so highly correlated as to cause multi-collinearity problems in the multiple regressions.

The total sample of 1,998 sixth and eighth grade students were split into two groups. With the first group, we ran an EFA to verify that there not strong alternative factors. Adequate fit to a unidimensional model was determined by a satisfactory visual inspection of the scree plot, sufficiently large factor loadings on each item (>0.30) (Table 4). With the second group, we ran a CFA to test the performance of each scale. We used three fit statistics to validate our measurements: (i) CFI: the comparative fit index tests how well the data the hypothesized unidimensional scale (values of 0.95 or above are considered satisfactory); (ii) TLI: the Tucker–Lewis index analyzes the discrepancy between the χ^2 statistic of the hypothesized model and the one of the null model. It ranges from 0 to 1 with values of 0.95 or more considered as satisfactory; (iii) RMSEA: the root mean square error of approximation index determines how well our model reproduces the data. It ranges from 0 to 1 with smaller values indicating a better model fit. Values of 0.06 or less are considered satisfactory (Osborne & Costello, 2009). Finally,

Table 4
CFA and Rasch statistics for scales

	Exploratory Factor Analysis			Confirmatory Factor Analysis			Rasch Model
Scale	Dimensions (Eigenvalue > 1)	Min Factor Loading	Max Factor Loading	Comparative Fit Index (CFI)	Tucker–Lewis Index (TLI)	Root mean square error of approximation (RMSEA)	EAP/PV[a]
Competency beliefs	1	0.57	0.74	0.97	0.96	0.073	0.83
Scientific sensemaking	1	0.56	0.83	0.90	0.90	0.06	0.77
Cognitive/behavioral engagement	1	0.58	0.73	0.87	0.82	0.10	0.65
Choice preferences	1	0.38	0.86	0.93	0.91	0.07	0.85

[a]EAP/PV reliability is the explained variance according to the estimated model divided by the total individuals variance.

with the whole dataset, the items were fit to a partial credit Rasch Model using R to calculate the infit (unweighted) and outfit (weighted) mean error square statistics (these item-level statistics are only reported for the Competency Beliefs measure, the central construct of this manuscript; see http://www.activationlab.org/tools/ for item-level statistics for the rest of the scales). Mean squared errors above one indicate that there is more noise in the data than the model can predict (e.g., a value of 1.2 indicates 20% more noise). We also calculated the person-separation reliability statistic (EAP/PV), which determines the inter-item reliability of the construct. As with Cronbach's alpha, values of 0.80 are considered sufficient (Wright & Stone, 1979). CFA and Rasch fit statistics are presented in Table 4.

As Table 4 shows, all our scales fit a one-dimensional model with factor loading of at least of 0.4, except for one item in the Choice Preferences scale with a value of 0.38. The competency beliefs scale has an RMSEA slightly above what is normally deemed acceptable, however, given the excellent CFI and TLI statistics, the scale has an overall acceptable model fit overall. The cognitive-behavioral engagement scale has mediocre model fits statistics, suggesting further improvements would be useful. But, as shown later, the reliability is sufficiently strong to reveal meaningful patterns in the data.

Measures

Competency Beliefs (CB). Competency beliefs were measured as a mean score across eight 4-point Likert scale items (Table 3). The scale asks students to consider specific contexts, typical for their age, that requires them to make sense of scientific information (e.g., in a class, at a museum, or at home). It includes items that measured competency beliefs for performing scientific activities (e.g., "I can do the science activities I get in class: All the time-most of the time-Half the time-Rarely"), and competency beliefs for performing scientific skills (e.g. "I think I am very good at defending my opinion when others disagree with me: YES!-yes-no-NO!"). We used Item Category Characteristic Item Curves (ICC) to determine an adequate number of Likert items for the scale (see online Supplemental Materials Figure S1) and we generated a Wright Map to assess the

convergent validity of the scale (Figure 2). A Wright Map plots the difficulty of the items against the overall distribution of the ability score of the sampled individuals. The Wright Map shows that each of the items and each of the possible response options are distributed along the sample distribution, meaning that we adequately distinguished between individuals with low, moderate, and high competency beliefs.

Person-item scores are the "ability" estimates produced by the IRT analysis, and are similar to factor scores produced by factor analysis. These estimates for each individual take into account the individuals responses and the relative "difficulty" of each item and each scale level of each item. However, we use simpler mean scores instead of these person-item scores in our analyses for two reasons. First, the ICC graphs showed that the response options of each item functioned approximately as an interval scale. That is, the psychological gap between levels was generally equivalent across the options (e.g., the difference between endorsing response option 1 vs. 2 was similar to the differences between 2 vs. 3). Second, the correlation between the person-item scores and the mean scores across the eight items was of 0.98, making them functionally equivalent. Similar patterns were found for all the scales (i.e., equal-sized gaps in the ICC graphs and correlations of person-item scores and mean scores close to 1) and thus means are used for all the scales.

Scientific Sensemaking (SSM). The survey used in this study is a minor variation from a previously validated instrument (Bathgate, Crowell, Schunn, Cannady, & Dorph, 2015). Scientific sensemaking is a science reasoning measure developed to assess how students engage with science-related content as a sensemaking activity using methods aligned with the accepted practices of science (Bathgate et al., 2015). The measure consists of 12 multiple choice items

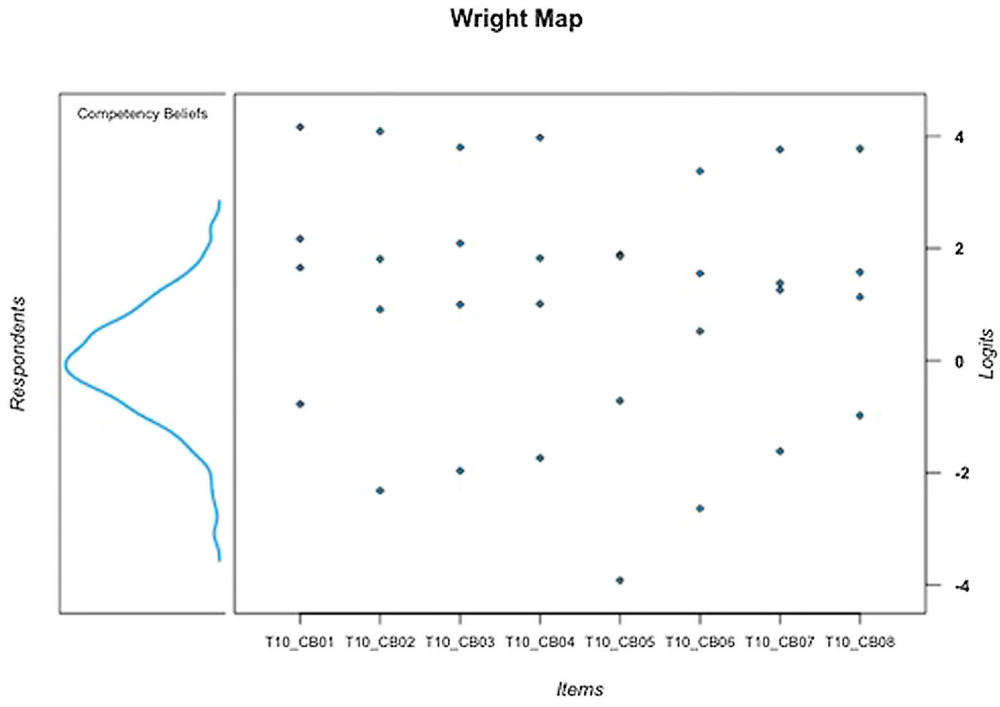

Figure 2. Competency beliefs Rasch model Wright map.

embedded in a scenario about saving endangered dolphins, a topic of broad interest to middle school students (Bathgate, Schunn, & Correnti, 2014). In addition, the scenario provides enough information so students with varying prior knowledge of dolphins have the opportunity to perform well if they have the underlying skills. The tested component skills are generating testable questions, designing investigations or experiments, adequate interpretation of data, constructing mechanistic explanations about natural and physical phenomena, and understanding of the nature of science (see online Supplemental Materials Table S1 for items, scoring, and Wright map). The score is reported in terms of percentage of correct responses.

Cognitive/Behavioral Engagement (Ecbh). Cognitive/behavioral engagement was measured with four Likert scale items (YES!-yes-no-NO! scored as 4, 3, 2, 1, or the reverse depending on the item) on a survey administered at the end of two science classes. To reduce memory encoding problems, the instrument asks students to reflect on the particular activities on a given day (e.g., "During today's activity I was focused on the things we were learning most of the time"). Measuring this at two different time points provided an assessment of a student's average engagement in science class. Behavioral engagement is related to the behaviors necessary for the learner to complete the task or off task behaviors. Cognitive engagement relates to the thought processes and learning attention necessary for meaningful processing of information. Although the cognitive and behavioral engagement can be differentiated theoretically, the responses were so highly correlated in these middle school science classroom contexts that separating those scores was not meaningful (see online Supplemental Materials Table S1 for items, scoring, and Wright map).

Choice Preferences (CP). Choice preferences for optional science learning experiences was also measured as a mean of ten items on a Likert scale (YES!-yes-no-NO! scored as 4, 3, 2, 1). These items provided students with future choices about participating in common optional learning experiences involving science at home, at school, or in other locations (e.g. "I would like to attend a science camp next summer"). The choices ranged from situations that could happen in the immediate future to choices about preferences for the next year (see online Supplemental Materials Table S1 for items, scoring, and Wright map).

Content Knowledge (CK). Pre/post-content knowledge tests were used to assess how much the students learned from their classroom instruction in the 4-month period of the study; each classroom was given an assessment that aligned with their curriculum. These tests were constructed to (i) assess the central conceptual content of each participating school's curriculum; (ii) assess the big ideas rather than only simple factual knowledge; and (iii) consume no more than one class period to complete. 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?" (i) the Sun, (ii) the Moon, (iii) Earth's gravity, or (iv) Earth's rotation). The curriculum road maps of each district and grade were analyzed to identify the topics that were going to be taught during the studied 4-month period. Participating teachers also verified the list of topics, since teachers sometimes adjust districts plans. The process resulted in five different tests for 6th grade and four different tests for 8th grade. To make the scores comparable across different classrooms and grades, we calculated the *z*-scores by subtracting the relevant test form mean and dividing by the relevant test form standard deviation.

Background Characteristics. Participants provided basic 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 with which they identified, and were allowed to choose

more than one. From the ethnicity data, a binary variable called Under-Represented Minority was created, with a 0 for Whites and Asians, and 1 otherwise.

Furthermore, the questionnaire included three scales related to socio-economic status (SES). SES is often assessed in research through the convenient simple indicator of free/reduced lunch status, which is a distal measure of the learning-relevant resources provided in the home. Instead, we included more direct measures of the socio-economic factors that support science learning: parental education (e.g., below High-School, graduated from High-School, College) of each parent, learning-relevant physical resources in the home (e.g., Internet, computer, calculator, etc.), and family support for learning (e.g., When I work on homework at home, I have someone who can help me with it if I need help). Parental education was coded as a categorical predictor based on highest educational level across both parents. Home resources consisted of a mean score across seven items. Family support consisted of a mean score across five items.

Curriculum Types. To assess whether patterns in the data were caused by differences in type of classroom instruction included in the curriculum, teachers completed a teacher log survey once a month asking them to rate what percentage of classroom time over the last week was dedicated to textbook/presentation versus hands-on kinds of classroom activities (see Table 5). Mean responses were computed for each teacher, and then a percentage of time spent doing hands-on type activities was calculated for each classroom ($M = 54\%$, $SD = 18\%$).

Classroom Dialogue. When researchers were in classrooms to administer the cognitive behavioral engagement surveys, they also rated the amount and type of observed dialogue in the classroom. To avoid rater fatigue, observers were asked to observe for 5-minute intervals at multiple times during the class and record how long during those 5 minutes they observed various forms of classroom dialog (see Table 6). Mean ratings were computed for each teacher, and then a percentage of time spent doing dialogic (vs. direct instruction) was calculated ($M = 43\%$, $SD = 19\%$).

Model Building Procedure

The reported analyses involve ANCOVAs. HLM analyses were also conducted to ensure that the obtained patterns held when controlling for classroom clustering (students nested within classrooms), but we only report the simpler, non-nested models because there were no differences on the coefficients for any of the central constructs and variance explained due to classroom clustering was below 5%. Models were systematically built in ways supported by theory and practice (Box, 1979) rather than via automated

Table 5
Teacher logged activities for hands-on versus textbook centered activities

Self-Reported Teaching Activity	Activity Type
Students read (alone or aloud) from a book or other informational text.	Textbook
Students listened to a lecture or presentation.	Textbook
Students watched a live or video-based demonstration.	Textbook
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

procedures such as stepwise and backwards regression (Whittingham, Stephens, Bradbury, & Freckleton, 2006):

(1) Beginning with key control predictors: other variables that likely also had a strong relationship to final content scores (e.g., pre-test *z*-score, school).
(2) Adding key research question predictors: the variables of interest for a given research question (e.g., scientific sensemaking, competency beliefs, sex).
(3) Adding secondary control predictors: variables that according to prior research could have also have an effect on final content scores (e.g., grade, age, SES). These variables were only kept in the model when statistically significant.
(4) Checking rival hypothesis: other variables that could offer alternative explanations to observed patterns (e.g., nature of classroom instruction across grades).
(5) Including interactions if necessary: interactions of relevance to the current research question (e.g., sex × competency beliefs).

Analyses were conducted and reported in the following order corresponding to the four research questions:

(1) For every initial student characteristic, we conducted an independent sample *t*-test between girls and boys separately within each grade in order to assess initial differences in underlying distributions by sex (RQ1).
(2) We constructed a multiple regression model to test whether competency beliefs are more strongly associated with science content learning outcomes for girls than boys (RQ2).
(3) We conducted moderated mediation and mediated moderation analyses using classroom engagement and choice preferences for optional science learning to identify underlying mechanisms for interactions obtained in the previous step (RQ3 & RQ4).

Results and Discussion

Quantifying Initial Differences Between Boys and Girls Across Grades in the Sample

Before assessing whether competency beliefs are differentially important for content learning by sex, it is important to first test to what extent there are different underlying distributions of any of the critical variables, and therefore, trivial causes of differential importance. Furthermore, given emerging changes in science performance by sex around this age, it is important to assess the size of differences in scientific sensemaking as well as competency beliefs. Table 7 shows means and standard deviations by sex and grade. We used Cohen's *d* as an effect size and interpret them using the standard cut-offs for small, medium, and large effect sizes (Cohen, 1992), although we

Table 6
Classroom observation items for the dialogic vs. direction instruction activities scale

Classroom Observation Activity	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 and response/IRE	Direct instruction

Table 7

Means, standard deviations for content knowledge scores, competency beliefs, scientific sensemaking, cognitive-behavioral engagement, and choice preferences, separately by sex and grade, along with Cohen's d, t-test, p-value, and 95%CI for the effect size in the difference by sex

Variable	6th Grade					8th Grade				
	Boys *M* (SD)	Girls *M* (SD)	Cohen's *d*	*p* - value	CI of the Effect Size	Boys *M* (SD)	Girls *M* (SD)	Cohen's *d*	*p* - value	CI of the Effect Size
Adjusted pre-test *z*-scores	−0.02 (1.0)	0.02 (1.0)	−0.03	0.56	(−0.39, −0.21)	0.04 (0.9)	0.03 (1.0)	−0.01	0.81	(−0.1, 0.08)
Adjusted post-test *z*-scores	0.04 (1.0)	0.01 (1.0)	−0.03	0.61	(−0.39, −0.21)	0.09 (0.9)	0.05 (1.0)	−0.04	0.46	(−0.13, 0.05)
Competency beliefs	2.89 (0.5)	2.93 (0.5)	0.08	0.15	(−0.01, 0.17)	2.80 (0.5)	2.82 (0.6)	0.04	0.45	(−0.05, 0.13)
Scientific sensemaking	54% (24%)	53% (24%)	−0.1	0.1	(−0.19, −0.01)	**65% (23%)**	**59% (26%)**	**−0.25**	**<0.001**	**(−0.34, −0.16)**
Cognitive behavioral engagement	3.12 (0.6)	3.06 (0.6)	−0.11	0.06	(−0.2, −0.02)	2.87 (0.5)	2.9 (0.5)	0.05	0.31	(−0.04, 0.14)
Choice preferences	2.61 (0.6)	2.66 (0.6)	0.09	0.13	(0, 0.18)	**2.28 (0.5)**	**2.40 (0.5)**	**0.23**	**<0.001**	**(0.14, 0.32)**

note that these cut-offs are by convention rather than strictly logical (Lakens, Hilgard, & Staaks, 2016). Examination of the effect sizes suggest that the sex differences on the variables of interest are at most minimal. These results are consistent with previous literature suggesting that there are not intrinsic differences by sex when it comes to cognitive abilities in science (Spelke, 2005), despite the fact that boys in the sample have directionally higher scores on all the ability and knowledge measures. In no cases are the standard deviations (or interquartile ranges) meaningfully different by sex or age, suggesting that restricted range issues cannot be the cause of differential predictiveness by sex in either grade.

Do Age and Sex Moderate the Impact of Competency Beliefs or Sensemaking on Learning?

Having established similar distributions, we turn to predictive relationships. We first examined the extent of between classroom variance through linear mixed methods. The fully unconditional model of students nested within classrooms showed that the variance accounted for classroom clustering (students nested within classrooms) was below 5%. Hence, we report the remaining analyses in terms of multiple regression instead of linear mixed models (although identical results are obtained for all key predictors when linear mixed models are used). Next, we verified with a paired *t*-test that the content assessments showed statistical significant evidence of overall content learning from pre ($M = 41\%$ correct, $SD = 17\%$) to post ($M = 50\%$ correct, $SD = 20\%$) with a medium size effect ($N = 2{,}348$, $p < 0.001$, $d = 0.48$) in an 18 question assessment. This appears to represent a relatively low mean 1.6 item change over a 4-month period between pre- and post-assessments. In general, the mean amount of science learning in urban schools can be quite low (Apedoe, Reynolds, Ellefson, & Schunn, 2008; Doppelt, Mehalik, Schunn, Silk, & Krysinski, 2008), further reinforcing the need for studies into motivational barriers to learning. However, 16% of our sample reported learning gains of five items or more. That is, despite a mediocre average gain of knowledge over a 4-month period, a sizable number of students had substantial gains.

For the regression analyses, to make the content test-scores comparable across forms, we *z*-scored the content assessment data within each test form. We began with a test of whether scientific sensemaking (abilities) and competency beliefs (beliefs about abilities) are each independent predictors of students' content learning. As a model-building process, this set of analyses used an alpha level of 0.1 as a threshold for main effects and interactions to err toward controlling for all useful predictors. We conducted a multiple linear regression predicting the content post-test with competency beliefs and sensemaking as independent variables and controlling for pre-test *z*-score and also including dummy variables for the school to which the students belonged. As results from Model 1 show (see Table 8), scientific sensemaking scores is a strong predictor of science content learning, replicating findings from Bathgate et al. (2015) with an earlier measure of scientific sensemaking. In addition, even when controlling for actual abilities, competency beliefs is also predictive of growth final content knowledge. Similar results are obtained with more complex models that include under-represented minority status or other student background characteristics.

Next, we examined whether sex or grade moderated the relationships of sensemaking or competency beliefs with learning: we conducted separate ANCOVAs with post-test scores as the dependent variable and included interaction terms of grade (coded as 0 = 6th grade, 1 = 8th grade) and sex (coded as 0 = boy, 1 = girl) with the variables of interest. We also centered the measures of scientific sensemaking and competency beliefs to avoid multicollinearity issues that are particularly problematic in interaction analyses. Results of Model 2 in Table 8 shows that there is no moderation of grade or sex on the effects of scientific sensemaking predicting learning; girls

Table 8
ANCOVA models predicting post-test knowledge z-scores

	Model 1		Model 2		Model 3		Model 4	
Predictor	β	p	β	p	β	p	β	p
Pre-test score	0.35	<0.001	0.36	<0.001	0.36	<0.001	0.36	<0.001
Scientific sensemaking (SSM)	0.33	<0.001	0.36	<0.001	0.35	<0.001	0.31	<0.001
Competency beliefs (CB)	0.08	0.08	0.07	<0.001	0.08	0.03	−0.04	0.45
Sex			−0.02	0.71	0.04	0.80	−0.04	0.2
Grade			0.03	0.66	0.15	0.29	−0.02	0.54
Grade × sex			−0.04	0.60	−0.33	0.06	−0.003	0.93
Sensemaking × sex			0.02	0.78				
Sensemaking × grade			−0.06	0.42				
Sensemaking × grade × sex			0.01	0.90				
Competency beliefs × sex					−0.04	0.76	0.06	0.2
Competency beliefs × grade					−0.17	0.22	0.05	0.28
Competency beliefs × grade × sex					0.30	0.08	0.005	0.10
School	−0.03–0.14		−0.06–0.10		−0.03–0.14			
Curriculum types							−0.083	<0.001
Classroom dialogue							0.076	<0.001
F	121.4		81.2		80.8		80.20	
R	0.43		0.43		0.43		0.37	

and boys in this study were equally able to use their sensemaking abilities for learning science content.

By contrast, the results of Model 3 are suggestive of a three-way interaction between competency beliefs, grade, and sex that we explore in depth given its close connection to the core research questions. Note that variations of Model 3 with fewer or more control variables consistently result in this three-way interaction trend with statistical significance below $p = 0.1$. Figure 3 shows the nature of the underlying three-way interaction by binning competency beliefs in thirds (low, moderate, and high) and plotting the estimated marginal means for each group, controlling for scientific sensemaking, school, and pre-test z-scores. At 6th grade, boys and girls in the sample achieved similar content knowledge (adjusted post-test z-scores) when they had similar competency beliefs. But by 8th grade, the boys acquired science content knowledge regardless of their competency beliefs while the girls in the sample showed a heightened sensitivity to competency beliefs. Note that a negative post-test is not indicative of a student unlearning content but that it simply reflects the use of z-scores and having a below-average mean.

Next, we conducted an ANCOVA test to see whether difference in teaching styles across grades might account for the three-way interaction. In particular, we classified teachers into three different types of instruction (textbook, mixed, hands-on) based on teacher log data and into three types of dialogue used in the classroom (dialogic, mixed, direct instruction) using the observation protocol data. We repeated the three-way interaction analysis including the curriculum types and classroom dialogue variables (Model 4). Both controls were statistically significant, and the three-way interaction remained with a similar regression coefficient and significance below $p < 0.1$. Furthermore,

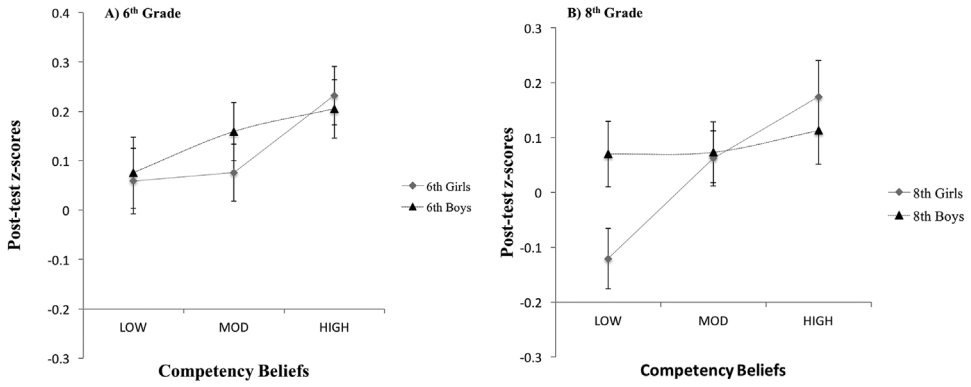

Figure 3. Estimated marginal means of adjusted post-test scores by binned competency beliefs and gender controlling for scientific sensemaking, school, and pre-test z-score for 6th and 8th graders.

Model 4 does not explain more variance than Model 3, so we elected to continue the next set of analysis using the simpler Model 3 as a base.

To unpack the three-way interaction that was of central theoretical focus in this paper, and specifically test the statistical significance of the underlying two-way interactions suggested in Figure 3, we ran ANCOVAS on the interaction of competency beliefs with sex separately by each grade. For the 6th graders ($F = 54.1$, $R^2 = 0.46$), the interaction between sex and competency beliefs was not statistically significant (CB $\times$ Sex $\beta = -0.041$, $p = 0.77$), similar to the visual pattern in Figure 3A. But, for the 8th graders ($F = 54.6, R^2 = 0.41$), the moderation effect of sex on competency beliefs was statistically significant (CB $\times$ Sex $\beta = 0.29$, $p = 0.03$), similar to the visual pattern in Figure 3B. Again these central results held with models with different control variables.

Having established that the sex moderation effect is only present in 8th grade, we proceeded to look for evidence of possible underlying mechanisms for this relationship within other data obtained with the 8th-grade sample. For this subsequent set of analyses which explore underlying mechanism, we applied a Bonferroni correction to produce an overall alpha of 0.05. That is, in testing two possible mediation analysis, the results had to be $p < 0.025$ to be considered statistically significant.

A Brief Introduction to Moderated Mediation and Mediated Moderation

The interaction of a third variable (sex) with a predictor variable (competency belief) is often called a moderated relationship. However, this statistical moderation result provides no information about the underlying mechanism. In general, mediation analyses are often used to provide insight into underlying mechanisms for the relationship between variables. To specifically explain a moderation effect, however, the mediation analyses are somewhat more complex, and so we provide a brief tutorial here (see online Supplemental Materials for a more detailed tutorial). Two qualitatively different patterns can explain the moderated relationship: mediated moderation or moderated mediation. For mediated moderation (also known as a conditional indirect effect), the mediating variable acts as the mediator between predictor and outcome in only some levels of the moderating variable. For example, if competency beliefs only influence cognitive/behavioral engagement in girls but not boys (and engagement is relevant to learning), then engagement explains the moderating role of sex on competency belief to content learning relationships as a mediated moderator. For moderated mediation, a mediating relationship is always present but it is

stronger for one group than for another. For example, perhaps choice of optional science learning experiences is a stronger mediator for girls than boys (Preacher, Rucker, & Hayes, 2007) (i.e., competency beliefs drive participation in optional science learning more strongly in girls than boys).

To test these possible models, one proceeds by estimating three mediation equations derived from the mediation model proposed by Baron and Kenny (1986) and following the procedure outline by Muller et al. (2005), essentially testing the significance and strength of the mediation connections as a function of the moderator. An extensive explanation of the mathematical models and procedure is available in the supporting information. Even though we tested both possibilities for research questions 3 and 4, to simplify the presentations of these complex results, we only report here the support for the models found to be explanatory. The analysis proceeds by estimating three regression equations (Baron & Kenny, 1986):

$$Y = \beta_{10} + \beta_{11}X + \beta_{12}\mathrm{Mo} + \beta_{13}X\mathrm{Mo} + \varepsilon_1 \quad (1)$$

$$\mathrm{Me} = \beta_{20} + \beta_{21}X + \beta_{22}\mathrm{Mo} + \beta_{23}X\mathrm{Mo} + \varepsilon_2 \quad (2)$$

$$Y = \beta_{30} + \beta_{31}X + \beta_{32}\mathrm{Mo} + \beta_{33}X\mathrm{Mo} + \beta_{34}\mathrm{Me} + \beta_{35}\mathrm{MeMo} + \varepsilon_3 \quad (3)$$

where X is the independent variable of interest (competency beliefs), Mo is the moderator (sex), Me is the mediator variable (engagement or optional learning experiences), Y is the dependent variable (content post-test scores), β represents the various regression weights, and ε represents the residual error terms in the regressions. The pattern of obtained β values are then evaluated for evidence of moderated mediation or mediated moderation (see details below).

Do Out-of-School Factors Explain the Moderated Relationship Between Competency Beliefs and Content Learning?

In the first hypothesized mediation model (see Figure 4), Choice Preferences for Optional Science Learning Experiences mediated Competency Beliefs' relationship to post-test z-scores (i.e., students with higher competency beliefs learn more because they are willing to do more optional science learning). The key results of the estimated mediation equations shown in Figure 4 were evaluated in terms of evidence for mediated moderation and moderated mediation (see online Supplemental Materials Table S4 for full model details).

A moderated mediation relationship was found, as shown by having met three conditions:

(1) Competency beliefs as a significant main effect on path a ($\beta_{\mathrm{CB-CK}} = 0.29, p < 0.001$).
(2) The standardized coefficient of the interaction of competency beliefs with sex on path a ($\beta_{\mathrm{CB^*Sex-CK}}$) minus the interaction of competency beliefs and sex on path c ($\beta_{\mathrm{CB2^*Sex-CK}}$) must be different from zero ($\beta_{\mathrm{CB^*Sex-CK}} - \beta_{\mathrm{CB2^*Sex-CK}} = 0.29–0.004 = 0.286 \neq 0$).
(3) The product of the standardized coefficients of Choice preferences on path c ($\beta_{\mathrm{CP-CK}}$) and the interaction between competency beliefs and sex on path b ($\beta_{\mathrm{CB^*Sex-CP}}$) and/or the product of the standardized coefficient between the interaction of choice preferences and sex on path c ($\beta_{\mathrm{CP^*Sex-CK}}$) and competency beliefs on path b ($\beta_{\mathrm{CB-CP}}$) must be different from zero ($\beta_{\mathrm{CP-CK}}\beta_{\mathrm{CB^*Sex-CP}} = 0.001 \approx 0$; $\beta_{\mathrm{CP^*Sex-CK}}{}^{*}\beta_{\mathrm{CB-CP}} = 0.13$).

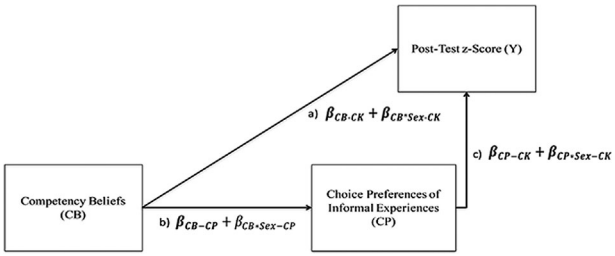

Figure 4. Estimated marginal means of adjusted post-test scores by binned choice preferences and gender controlling for scientific sensemaking, school, and pre-test *z*-score for 8th graders.

As the pattern of results shown above, Choice Preferences appears to have mediated the pattern between Competency Beliefs, sex, and science content through a moderated mediation mechanism. In other words, students' Choice Preferences mediated the relationship of Competency Beliefs to post-test *z*-scores for the girls but not for the boys.

For both sexes, Competency Beliefs predicted Choice Preferences to an equal extent. But, as shown in Figure 5, choice preferences appear to have benefitted content learning in the girls but not in the boys of this sample. It may be that the boys tended to actually participate in (and hence benefit from) relevant informal learning experiences whether they wanted to or not due to parental or peer expectations. However, the girls may have participated only when they had strong preferences (driven by competency beliefs) and then the girls learned more science content from this participation. In other words, the broken linkage for boys may been between preferences and participation. Future research that also measures actual participation and perceptions of risk at participating in optional learning experiences is needed to shed further light on the underlying mechanism.

Do Inside-of-School Factors Explain the Moderated Relationship Between Competency Beliefs and Content Learning?

We also tested another hypothesized mediation model (see Figure 6) that Cognitive-Behavioral Engagement mediated the relationship between Competency Beliefs' and post-test *z*-scores. The key results of the estimated mediation equations shown in Figure 7 were evaluated in terms of evidence for mediated moderation and moderated mediation (see online Supplemental Materials Table S5 for full model details).

A moderated mediation relationship was found, as shown by having met three conditions:

(1) Competency beliefs as a significant main effect on path a ($\beta_{\mathrm{CB-CK}} = 0.02, p < 0.001$).
(2) The standardized coefficient of the interaction of competency beliefs with sex on path a ($\beta_{\mathrm{CB^*Sex-CK}}$) minus the interaction of competency beliefs and sex on path c ($\beta_{\mathrm{CB2^*Sex-CK}}$) must be different from zero ($\beta_{\mathrm{CB^*Sex\text{-}CK}} - \beta_{\mathrm{CB2^*Sex-CK}} = 0.26\text{–}0.001 = 0.259 \neq 0$).
(3) The product of the standardized coefficients of Cognitive/Behavioral Engagement on path c ($\beta_{\mathrm{Ecbh-CK}}$) and the interaction between competency beliefs and sex on path b ($\beta_{\mathrm{CB^*Sex-Ecbh}}$) and/or the product of the standardized coefficient between the interaction of Cognitive Behavioral Engagement and sex on path c ($\beta_{\mathrm{Ecbh^*Sex}}$) and competency beliefs on path b ($\beta_{\mathrm{CB-Ecbh}}$) must be different from zero ($\beta_{\mathrm{Ecbh-CK}}\boldsymbol{\beta_{\mathrm{CB}}}_{\mathrm{Sex-Ecbh}} = -0.001 \approx 0; \beta_{\mathrm{Eccbh^*Sex}}{}^{*}\beta_{\mathrm{CB-Ecbh}} = 0.14$).

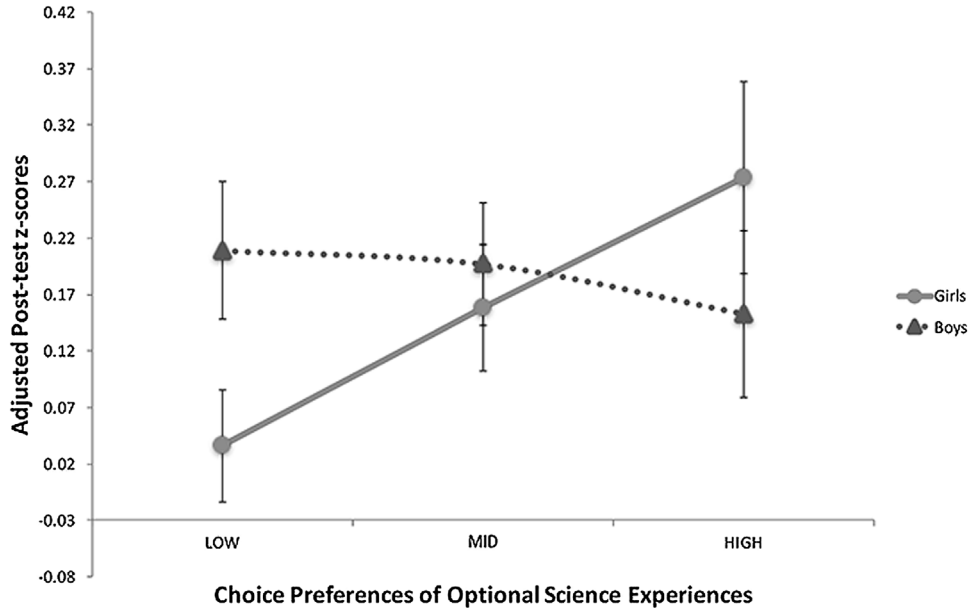

Figure 5. Estimated marginal means of adjusted post-test scores by binned choice preferences and gender controlling for scientific sensemaking, school, and pre-test *z*-score for 8th graders.

In particular, even though for both sexes, Competency Beliefs predicted Cognitive/Behavioral Engagement to roughly an equal extent, cognitive/behavioral engagement appears to have related to amount of science content learning for the girls but not for the boys in this sample (see Figure 7). Since it is surprising that the boys were able to succeed in the science classroom regardless of level of cognitive/behavioral engagement, we discuss possible explanations in the General Discussion section.

General Discussion

We revisit each of our primary research questions to discuss theoretical and practical implications of our findings.

To What Extent Do Sex Differences in Scientific Sensemaking Abilities Appear at the Transition Into Adolescence?

Across a large, diverse sample of learners, we found no substantial intrinsic differences when it comes to scientific sensemaking skills in science of males and females for science learning in grades 6th and 8th (corresponding to ages 11−14). As Table 7 shows, even though there are differences emerging by 8th grade, the effect size is small, corresponding to heavily overlapping distributions. Furthermore, the interaction of Sex and Scientific Sensemaking in predicting learning was not statistically significant; that is, there is no evidence that the girls in the sample were less well positioned than the boys to use their underlying scientific sensemaking abilities to learn science content. This lack of reasoning ability differences is especially important since sensemaking and prior content knowledge account for a size fraction (37%) of the variance in content knowledge post-test scores. Middle school is often characterized as a time where all adolescents undergo critical reorganization of regulatory systems (Steinberg, 2005) and large

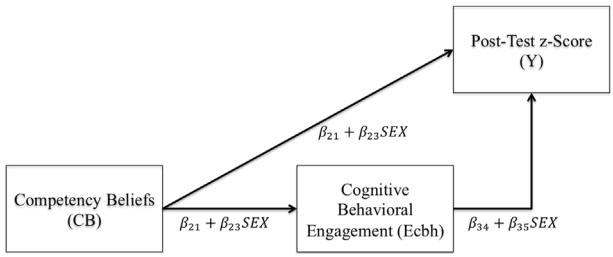

Figure 6. Hypothesized model for the mediation of competency beliefs interaction with student's gender by cognitive/behavioral engagement for predicting post-test scores of 8th graders.

sex-specific physiological changes (Staurt, Shock, Breckenridge, & Vincent, 1953). Since prior research has established that there are not sex differences pre-adolescence in cognitive abilities in learning, reasoning, and mechanical interactions (Spelke, 2005), adolescence as a stage of great developmental change would be the likely candidate where these differences could appear. That the observed 8th grade differences in reasoning ability were so small suggests that a social rather than physical mechanism was responsible for the effects; for example, just as competency beliefs appears to have shaped learning of science content, competency beliefs may also have shaped the development of scientific sensemaking.

We acknowledge that meta-analyses have reported lower achievement in science for girls (for ages 8–18) and not in other domains (Becker, 1989; Hattie, 2008). However, early sex differences in science achievement have been found to be limited to very high cognitive ability students and to

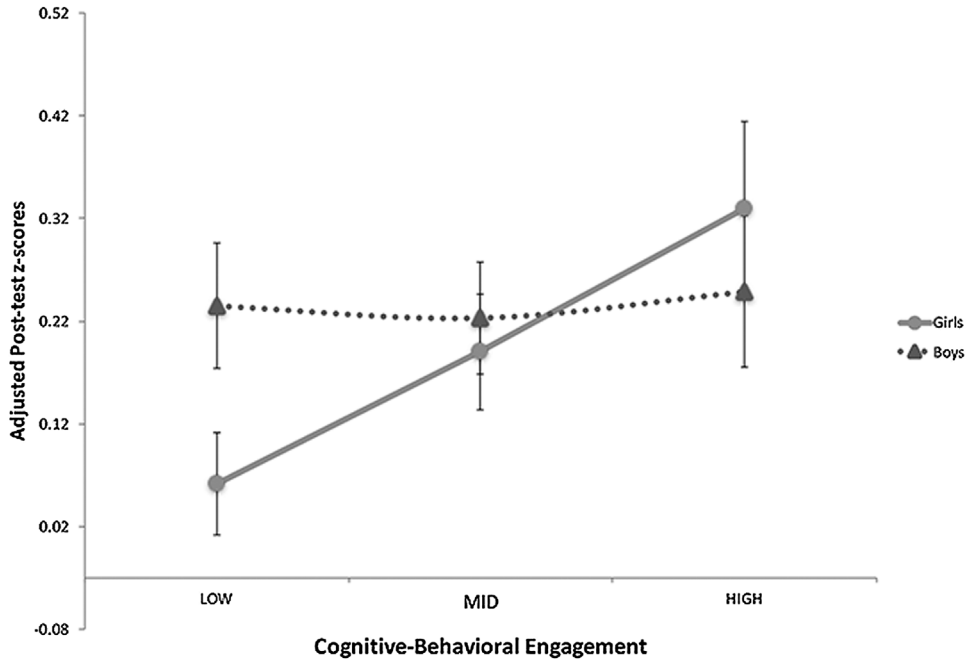

Figure 7. Estimated marginal means of adjusted post-test scores by binned cognitive-behavioral engagement and gender controlling for scientific sensemaking, school, and pre-test *z*-score for 8th graders.

physical science content (Dimitrov, 1999). Indeed, in life sciences, girls outperformed boys in 8th grade (Yenilmez, Sungur, & Tekkaya, 2006) and were similar in 10th grade (Sungur & Tekkaya, 2003). Moreover, research on "stereotype threat" (the fear of confirming negative stereotypes about your social, or sex group) (Leinbach et al., 1997; Shapiro & Williams, 2012) suggests that these performance gaps may be partly due to stereotypes invoking disruptive worrying behavior during the completion of the science assessments.

Competency Beliefs Are More Strongly Associated With Science Content Learning Outcomes for Girls Than Boys

Interestingly, we found no overall sex differences across our large samples of learners in mean competency beliefs in 6th or 8th grades. That is, a primary contribution of this paper is not about the mean differences in competency beliefs, but rather about the consequences of low competency beliefs. We contribute to a growing literature showing that beliefs about competencies (also called self-efficacy beliefs), not just the actual competencies, influence how much students learn (Andrew, 1998; Bandura, 1986; Zeldin et al., 2008; Zimmerman & Bandura, 1994); our models found that Competency Beliefs variation explain approximately 6% of the variance in science content learning, even when controlling for actual abilities and prior knowledge. We further uncovered an important variation in this relationship by sex and age: in our sample, competency beliefs predicted learning outcomes for girls but not for boys in 8th grade, but competency beliefs were (equally) predictive of learning outcomes across sexes in 6th grade. This grade by sex by competency beliefs interaction was not statistically robust, but it should be noted that three-way interactions are usually statistically underpowered analyses. Even with the current large sample, the power for this analysis was only 0.75 (i.e., 75% chance of rejecting the null hypothesis at the $\alpha = 0.05$). However, the two-way sex by competency belief interaction within the 8th grade data was statistically significant, and formed the focus of the follow-up analyses.

Why would the relationship of competency beliefs and learning be absent in 8th grade boys? We discuss likely mechanisms in the next sections, but it is also important to rule out more trivial methodological problems as possible sources. One possibility to investigate is whether our competency beliefs scale was no longer measuring the construct accurately for boys in 8th grade. However, the scales had similar (and acceptable) scale reliabilities within both 8th grade boys and girls (Chronbach alphas of 0.86 and 0.83, respectively). A second possibility is that there is restricted range in 8th grade boys' competency beliefs. Indeed, males tend to overestimate their abilities (Webb-Williams, 2014) and tend to have higher self-esteem (Kling, Hyde, Showers, & Buswell, 1999). However, the interquartile ranges in competency beliefs were the same across for boys and girls in 8th grade ([2.5,3.3] and [2.4,3.2] respectively). Furthermore, we considered whether the moderation effect could be a consequence of the different type of instruction could have an effect on student's achievement, self-efficacy, and self-regulation (Sungur & Tekkaya, 2003). However, we performed these analyses controlling for type of instruction (e.g. hands-on vs. textbook) and use of dialogue in the classroom (dialogic vs. direct instruction). These variables did not change the interaction pattern. We also performed these analyses in HLM (Bates, Mächler, Bolker, & Walker, 2014) to control for unmeasured teacher, classroom culture, and school effects (e.g. content difference, teacher experience, cohort effects, etc.) and we found the same results as with traditional regression.

In terms of theories of motivation and learning, they now need to be expanded to take into consideration the moderating role of sex and age. Existing models have regularly included the role of context variables in shaping motivational variables such as interest or competency beliefs (Eccles et al., 1998). Here, we find instead that context shapes the effects of at least one of these motivational variables.

Practically speaking, these results suggest that educators should find ways to reduce the sex-based threats associated with (required and optional) activities such that competency beliefs become less of a driver; this point is discussed further in sections below. Additionally, future research should focus on (i) what drives boys' success in science beyond competency beliefs and (ii) whether we find similar moderating effects by ethnicity.

The Role of Optional Science Experiences in Mediating the Relationship Between Competency Beliefs and Science Learning

Not surprisingly, following prior research and many influential theories of motivation and learning (Eccles et al., 1998; Eccles & Jacobs, 1986; Eccles & Wigfield, 2002; Sha et al., 2015), students' choice preferences for participating in optional science experiences were associated with competency beliefs. As noted in the literature review, prior research suggests sex differences in which out-of-school activities are selected when given the choice. Our instrument was carefully designed and iteratively improved using Differential Item Functioning IRT analyses to make sure there were no sex biases in the options being presented in the choice preferences survey instrument (i.e., the child made the choice because of its association with science preferences overall, not because the choice was gendered, such as a chemistry kit for making perfume).

Important for science policy, our findings further suggest that these optional science learning experiences were associated with science content learning, and specifically learning of the content being taught in class. Much of the prior literature on free choice learning or informal science learning has emphasized the importance optional experiences for increasing interest or identity in science (Brotman & Moore, 2008; Maltese & Tai, 2009; Zimmerman et al., 2000), leaving classrooms as the primary driver of core science content learning. Although causality cannot be strongly tested using correlational analyses, the key measures of relevance here were collected in a temporally sequential order in this study, ruling out reverse causal effects such as of content learning to initial competency beliefs. In addition, many likely confounds were addressed in the multiple regression analyses. But additional research will be required to rigorously test causality and further understand these effects.

More central to the sex mediation effect, we find the relationship between optional experiences and learning outcomes held for 8th girls but not for 8th grade boys in our sample (see Figure 3). But why would boys' actual participation in optional experiences not be predicted by competency beliefs, or at least to a much weaker extent than are girls' choices? Throughout their development, boys will have been exposed to more science-related role models (Eccles, 2011). Additionally, both boys and girls come to perceive many things as gendered—in this case, they will have learned that science-relevant activities are perceived as male (Leinbach et al., 1997). For girls, science will typically be perceived as an unlikely area of success (Zeldin et al., 2008). For these reasons, competency beliefs may have played a more important role for girls because these beliefs can help them take the risk (Marianne, 2011) of doing something different from what is expected of them (Kleinman, 1998). By contrast, science boys may have perceived science experiences as closely related to their sex, they may not have needed to feel confident in their abilities in order to choose to participate. In other words, it may be that competency beliefs matter more when there is perceived social risk. Future research is needed to uncover what factors matter when risk is low; for example, is science interest (or relative interest compared with other topics) the only factor that drives choice for boys?

Pragmatically, while there are many interventions devoted to providing girls access to science experiences (Brotman & Moore, 2008), the current findings call for a need to also focus on interventions that can build up girls' competency beliefs in science, as well as improving girls' social support systems that translate competency beliefs into productive use of learning resources.

These results also suggest that we must study not only willingness to participate but also opportunities to participate in optional learning, including an examination of the kinds of social pressures that prevent some learners from making effective use of outside-of-class learning opportunities.

The Role of Cognitive and Behavioral Engagement During Science Class in Mediating the Relationship Between Competency Beliefs and Science Learning

As with the optional experiences, extensive prior research suggests that cognitive/behavioral engagement in class is driven by competency beliefs (Gasiewski, Eagan, Garcia, Hurtado, & Chang, 2012; Pintrich et al., 2001), and that cognitive/behavioral engagement in class supports science learning outcomes (Zimmerman et al., 2000). The surprising finding in the current study is that this path was not found for 8th-grade boys. In particular, in the 8th-grade sample, we found (see Figure 7) a moderated mediation relationship in which cognitive-behavioral engagement was only mediating learning in girls.

How could it be that the boys not paying attention in class learned as much science content at the same rate as boys who are paying attention? It is possible that the boys were less able to monitor their learning processes, or less willing to report accurately. However, under those explanations, we would expect lower scale reliability, reduced variability, and/or lower standard deviations in boys than girls, but none of these expectations are supported by the data. Teachers tend to attribute boys' success in the classroom to innate ability rather than effort (Fennema, Peterson, Carpenter, & Lubinski, 1990), are more likely to call on male students' participation in class (Altermatt, Jovanovic, & Perry, 1998), and provide more directions and instructions (Serbin, O'Leary, Kent, & Tonick, 1973). This sex-differential treatment of teacher toward boys may have an effect on their achievement despite having low cognitive/behavioral engagement in class. However, the current study did not provide direct evidence for any mechanisms behind these engagement results, and future research should test underlying mechanisms, as well as verify the replicability and causality of the effect.

Generalizability of the Patterns

Another question involves generalizability of the observed patterns. This study purposely sampled public urban middle school students in the United States and teaching contexts that varied widely on several important dimensions (e.g., distributions of learners from low SES backgrounds, instruction that is student-centric or using hands-on materials, % of minority students in schools) and roughly matched typical racial/ethnic and sex backgrounds found urban public schools in the United States. This sampling provided us with the opportunity to have students from a wide variety of environments.

Even though the focus of this paper is on sex differences, the race/ethnicity diversity of our sample and its overall match to base-rates in US urban public schools is important because of previously found interactions of sex and ethnicity in science (Aschbacher, Li, & Roth, 2010; Hazari, Sadler, & Sonnert, 2013). Thus, these demographics are important aspects to consider in producing policy relevant findings. In general, we purposefully sampled to be representative on the dimensions that mattered the most for science learning (sex, ethnicity) while also sampling broadly on other dimensions implicated in the science learning literature (parental education and income, school structure, curriculum type (hands-on vs. textbook), and instructional approach (student vs. teacher centric). The observed relationships held despite this large underlying diversity of learning contexts, suggesting some level of generalizability. Indeed, many of our measures of contextual diversity had statistically significant effects on content post-test scores:

SES measures (Family support, Home Resources, Highest Parental Education Level) accounted for 3%, ethnicity accounted for 2%, and the classroom clustering accounted for 5% of the variance in content scores. These additional effects support the use of these measures as valid statistical controls. It is important to note that despite their statistical significance, the overall variance explained by background variables is small compared to other reports in the literature. Often achievement gaps are reported only in terms of simple associations with demographics. Our results show motivational factors are important drivers of demographic effects, leaving little unexplained variance remaining.

Limitations

Our sample did not involve data from rural or suburban populations or from private or religious schools, nor did we have enough schools of various subtypes to warrant using regression weights to produce generalization to the overall US public urban population. This sample has similar sex and racial/ethnic demographics to national averages, but conclusions should be limited to the target population in our sample. Furthermore, given the correlational nature of the analyses, interventions on student competency beliefs to track their effects on engagement and learning will be important to include in future research on this topic.

Implications

We now turn to implications for the current research for different education stakeholders.

Policy Makers. For over 60 years, US policymakers have worried about the output of scientific workers in the country given the importance of US technological leadership (Sargent, 2014). This concern has led to various education policies and funding of research programs aimed to increase the number of students taking advance math and science courses in high school along with other programs aimed to increase the entering numbers STEM majors in college (Maltese & Tai, 2011). In this paper, we focus on how differential motivation by sex has important effects on science achievement in middle school which in turn will have snowballing effects on student's advanced science course taking in high school. Policies aimed at pre-high school will likely be important to include to address current pipeline problems, especially related to lack of diversity in the pipeline. Furthermore, our focus on the role of motivation as an important factor in achievement choices suggest that policies focused on "access" will not be sufficient to increase the number of women in the STEM pipeline: even with equal access, motivational differences will continue to produce differential attrition.

Practitioners and Teachers. Results from this work suggest that practitioners and teachers should focus on pursuing interventions that can target girls' competency beliefs (Betz & Schifano, 2000; Luzzo, Hasper, Albert, Bibby, & Martinelli, 1999; Schunk & Ertmer, 2000), such as providing access to female role models in science (Marx & Roman, 2002) or doing problem-based learning activities of broad perceived relevance (Dunlap, 2005). Such interventions should increase girls' achievement choices and in turn achievement outcomes in science. In addition, our work suggests that practitioners should more closely monitor actual levels of participation in optional science learning, especially by sex, to better understand how well they are supporting development in science of all their learners.

Researchers. Our work extends prior research by revealing an important moderating effect of sex on motivational benefits for learning. Researchers must now think not only about simple sex effects on motivation or participation but also about differential effects of these motivational outcomes on participation. Furthermore, the interaction with sex by grade creates a need to

replicate other prior findings across the middle school years and to better understand these key developmental changes during early adolescence.

Conclusion

Prior work has highlighted the importance of motivational factors such as competency beliefs in shaping science learning beyond the effects of actual scientific sensemaking abilities, and much research on sex effects in science has suggested motivational explanations for explaining sex differences in outcomes. The current research extends this prior work by revealing an important moderating effect of sex on motivational benefits for learning. This moderation has two theoretically and pragmatically important aspects: why effects are large for girls and why effects are small or absent for boys. We provide evidence in support of two underlying mechanisms related to achievement choices: participating in optional science learning and participation in class. Further research will be required to both validate these explanations and develop new interventions to better support girls' participation in science given these emerging factors.

References

Altermatt, E. R., Jovanovic, J., & Perry, M. (1998). Bias or responsivity? Sex and achievement-level effects on teachers' classroom questioning practices. Journal of Educational Psychology, 90(3), 516.

Andrew, S (1998). Self-efficacy as a predictor of academic performance in science. Journal of Advanced Nursing, 27(3), 596–603. http://doi.org/10.1046/j.1365-2648.1998.00550.x

Apedoe, X. S., Reynolds, B., Ellefson, M. R., & Schunn, C. D. (2008). Bringing engineering design into high school science classrooms: The heating/cooling unit. Journal of Science Education and Technology, 17(5), 454–465.

Archer, L., DeWitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2012). "Balancing acts": Elementary school girls' negotiations of femininity, achievement, and science. Science Education, 96(6), 967–989. http://doi.org/10.1002/sce.21031

Aschbacher, P. R., Li, E., & Roth, E. J. (2010). Is science me? High school students' identities, participation and aspirations in science, engineering, and medicine. Journal of Research in Science Teaching, 47(5), 564–582.

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, N.J: Prentice-Hall.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51(6), 1173.

Baron-Cohen, S. (2003). The essential difference: The truth about the male and female brain. New York: Basic Books.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version, 1(7).

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., & Schunn, C. (2016). Disentangling intensity from breadth of science interest: What predicts learning behaviors? Instructional Science, 44(5), 423–440.

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

Becker, B. J. (1989). Gender and science achievement: A reanalysis of studies from two meta-analyses. Journal of Research in Science Teaching, 26(2), 141–169.

Beeth, M. E. (1998). Teaching for conceptual change: Using status as a metacognitive tool. Science Education, 82(3), 343–356. http://doi.org/10.1002/(SICI)1098-237X(199806)82:3<343::AID-SCE3>3.0.CO;2-C

Betz, N. E., & Schifano, R. S. (2000). Evaluation of an intervention to increase realistic self-Efficacy and interests in college women. Journal of Vocational Behavior, 56(1), 35–52.

Boekaerts, M., Pintrich, P. R., & Zeidner, M. (2007). Handbook of self-regulation. San Diego, Calif: Academic Press.

Bouffard-Bouchard, T., Parent, S., & Larivee, S. (1991). Influence of self-efficacy on self-regulation and performance among junior and senior high-school age students. International Journal of Behavioral Development, 14(2), 153–164.

Box, G. E. (1979). Robustness in the strategy of scientific model building. Robustness in Statistics, 1, 201–236.

Britner, S. L. (2008). Motivation in high school science students: A comparison of gender differences in life, physical, and earth science classes, Journal of Research in Science Teaching 45(8), 955–970. http://doi.org/10.1002/tea.20249

Brotman, J. S., & Moore, F. M. (2008). Girls and science: A review of four themes in the science education literature. Journal of Research in Science Teaching, 45(9), 971–1002. http://doi.org/10.1002/tea.20241

Brown, B. A. (2004). Discursive identity: Assimilation into the culture of science and its implications for minority students. Journal of Research in Science Teaching, 41(8), 810–834. http://doi.org/10.1002/tea.20228

Buck, G. A., Clark, V. L. P., Leslie-Pelecky, D., Lu, Y., & Cerda-Lizarraga, P. (2008). Examining the cognitive processes used by adolescent girls and women scientists in identifying science role models: A feminist approach. Science Education, 92(4), 688–707.

Caplan, P. J., MacPherson, G. M., & Tobin, P. (1985). Do sex-related differences in spatial abilities exist? A multilevel critique with new data. American Psychologist, 40(7), 786–799. http://doi.org/10.1037/0003-066X.40.7.786

Carlone, H. B. (2004). The cultural production of science in reform-based physics: Girls' access, participation, and resistance. Journal of Research in Science Teaching, 41(4), 392–414. http://doi.org/10.1002/tea.20006

Chin, C. (2006). Using self-questioning to promote pupils' process skills thinking. School Science Review, 87(321), 113–119.

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

Chipman, S. F., Krantz, D. H., & Silver, R. (1992). Mathematics anxiety and science careers among able college women. Psychological Science, 3(5), 292–295. http://doi.org/10.1111/j.1467-9280.1992.tb00675.x

Clark Blickenstaff, J. (2006). Women and science careers: Leaky pipeline or gender filter? Gender and Education, 17(4), 369–386. http://doi.org/10.1080/09540250500145072

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159.

Desimone, L. M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. Educational Evaluation and Policy Analysis, 26(1), 1–22.

Dimitrov, D. M. (1999). Gender differences in science achievement: Differential effect of ability, response format, and strands of learning outcomes. School Science and Mathematics, 99(8), 445–450.

Doppelt, Y., Mehalik, M. M., Schunn, C. D., Silk, E., & Krysinski, D. (2008). Engagement and achievements: A case study of design-Based learning in a science context. Journal of Technology Education, 10(2). http://doi.org/10.21061/jte.v10i2.a

Dorph, R., Cannady, M. A., & Schunn, C. D. (2016). How science learning activation enables success for youth in science learning experiences. Electronic Journal of Science Education, 20(8), 50–85.

Dunlap, J. C. (2005). Problem-based learning and self-efficacy: How a capstone course prepares students for a profession. Educational Technology Research and Development, 53(1), 65–83. http://doi.org/10.1007/BF02504858

Eccles, J. S. (2011). Understanding educational and occupational choices. Journal of Social Issues, 67(3), 644–648. http://doi.org/10.1111/j.1540-4560.2011.01718.x

Eccles, J. S., & Jacobs, J. E. (1986). Social forces shape math attitudes and performance on JSTOR. Signs, 11, 367–380. http://doi.org/10.2307/3174058

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. Annual Review of Psychology, 53(1), 109–132.

Eccles, J. S., Barber, B. L., Updegraff, K., & OBrien, K. M. (1998). An expectancy-value model of achievement choices: The role of ability self-concepts, perceived task utility and interest in predicting activity choice and course enrollment. In Interest and learning: Proceedings of the seeon conference on interest and gender (pp. 267–280). Kiel: Institut für die Pädagogik der Naturwissenschaften.

Feder, M. A., Shouse, A. W., Lewenstein, B., & Bell, P. (2009). Learning science in informal environments: People, places, and pursuits. National Academies Press.

Fennema, E., Peterson, P. L., Carpenter, T. P., & Lubinski, C. A. (1990). Teachers' attributions and beliefs about girls, boys, and mathematics. Educational Studies in Mathematics, 21(1), 55–69.

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.

Gasiewski, J. A., Eagan, M. K., Garcia, G. A., Hurtado, S., & Chang, M. J. (2012). From gatekeeping to engagement: A multicontextual, mixed method study of student academic engagement in introductory STEM courses. Research in Higher Education, 53(2), 229–261.

Ghee, M., Keels, M., Collins, D., Neal-Spence, C., & Baker, E. Fine-tuning summer research programs to promote underrepresented students' persistence in the STEM pathway. CBE-Life Sciences Education, 15(3), ar28–ar28. http://doi.org/10.1187/cbe.16-01-0046

Gold, M. S., & Bentler, P. M. (2009). Treatments of missing data: A monte carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-Maximization. Structural Equation Modeling, 7(3), 319–355. http://doi.org/10.1207/S15328007SEM0703_1

Halpern, D. F., Aronson, J., Reimer, N., Simpkins, S., Star, J. R., & Wentzel, K. (2007). Encouraging girls in math and science (No. NCER 2007–2003). Washington DC: National Center for Education Research.

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development. Journal of Business Research, 57(2), 98–107. http://doi.org/10.1016/S0148-2963(01)00295-8

Hattie, J. (2008). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. New York, US: Routledge.

Hazari, Z., Sadler, P. M., & Sonnert, G. (2013). The science identity of college students: Exploring the intersection of gender, race, and ethnicity. Journal of College Science Teaching, 42(5), 82–91.

Hughes, G. (2000). Marginalization of socioscientific material in Science–Technology–Society science curricula: Some implications for gender inclusivity and curriculum reform. Journal of Research in Science Teaching, 37(5), 426–440. http://doi.org/10.1002/(SICI)1098-2736(200005)37:5<426::AID-TEA3>3.0.CO;2-U

Jones, M. G., Howe, A., & Rua, M. J. (2000). Gender differences in students' experiences, interests, and attitudes toward science and scientists. Science Education, 84(2), 180–192.

Kimura, D. (2000). Sex and cognition. Cambridge; London: MIT Press.

Kleinman, S. S. (1998). Overview of feminist perspectives on the ideology of science. Journal of Research in Science Teaching, 35(8), 837–844.

Kling, K. C., Hyde, J. S., Showers, C. J., & Buswell, B. N. (1999). Gender differences in self-esteem: A meta-analysis. Psychological Bulletin, 125(4), 470.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2010). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational Psychologist, 41(2), 75–86. http://doi.org/10.1207/s15326985ep4102_1

Labuhn, A. S., Zimmerman, B. J., & Hasselhorn, M. (2010). Enhancing students' self-regulation and mathematics performance: The influence of feedback and self-evaluative standards. Metacognition and Learning, 5(2), 173–194. http://doi.org/10.1007/s11409-010-9056-2

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. BMC Psychology, 4(1), 1. http://doi.org/10.1186/s40359-016-0126-3

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.

Legewie, J., & DiPrete, T. A. (2014). The high school environment and the gender gap in science and engineering. Sociology of Education, 87(4), 259–280. http://doi.org/10.1177/0038040714547770

Leinbach, M. D., Hort, B. E., & Fagot, B. I. (1997). Bears are for boys: Metaphorical associations in young children's gender stereotypes. Cognitive Development, 12(1), 107–130. http://doi.org/10.1016/S0885-2014(97)90032-0

Lin, P. Y., & Schunn, C. D. (2016). The dimensions and impact of informal science learning experiences on middle schoolers' attitudes and abilities in science. International Journal of Science Education, 1–22.

Lippman, L., McArthur, E., & Burns, S. (2004). Urban schools: The challenge of location and poverty. National Center for Education Statistics.

Luzzo, D. A., Hasper, P., Albert, K. A., Bibby, M. A., & Martinelli, E. A. J. (1999). Effects of self-efficacy-enhancing interventions on the math/science self-efficacy and career interests, goals, and actions of career undecided college students. Journal of Counseling Psychology, 46(2), 233–243. http://doi.org/10.1037/0022-0167.46.2.233

Mahmoodi, M. H., Kalantari, B., & Ghaslani, R. (2014). Self-regulated learning (SRL), motivation and language achievement of iranian EFL learners. Procedia—Social and Behavioral Sciences, 98, 1062–1068. http://doi.org/10.1016/j.sbspro.2014.03.517

Maltese, A. V., & Tai, R. H. (2009). Eyeballs in the fridge: Sources of early interest in science. International Journal of Science Education, 32(5), 669–685. http://doi.org/10.1080/09500690902792385

Maltese, A. V., & Tai, R. H. (2011). Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among U. S. students. Science Education, 95(5), 877–907. http://onlinelibrary.wiley.com.pitt.idm.oclc.org/doi/10.1002/sce.20441/full

Marianne, B. (2011). New perspectives on gender. In Handbook of labor economics, 4, (pp. 1543–1590). San Diego, CA: Elsevier.

Marx, D. M., & Roman, J. S. (2002). Female role models: Protecting women's math test performance. Personality and Social Psychology Bulletin, 28(9), 1183–1193. http://doi.org/10.1177/01461672022812004

Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., Fishman, B., Soloway, E., Geier, R., & Tal, R. T. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. Journal of Research in Science Teaching, 41(10), 1063–1080. http://doi.org/10.1002/tea.20039

Moss-Racusin, C. A., & Dovidio, J. F. (2012). Science faculty's subtle gender biases favor male students (Vol. 109, pp. 16474–16479). Presented at the Proceedings of the National Academy of Sciences. http://doi.org/10.1073/pnas.1211286109

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. Journal of Personality and Social Psychology, 89(6), 852.

Muller, P. A., Stage, F. K., & Kinzie, J. (2001). Science achievement growth trajectories: Understanding factors related to gender and racial–ethnic differences in precollege science achievement. American Educational Research Journal, 38(4), 981–1012. http://doi.org/10.3102/00028312038004981

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. TIMSS & PIRLS International Study Center. Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467.

National Student Clearinghouse Research Center. (2015). Science & Engineering Degree Attainment*:* 2004–2014 retrieved from https://nscresearchcenter.org/snapshotreport-degreeattainment15/

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., . . . Kesebir, S. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. Proceedings of the National Academy of Sciences, 106(26), 10593–10597.

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. Sex Roles, 39(1–2), 21–43. http://doi.org/10.1023/A:1018873615316

Oakes, J. (1990). Opportunities, achievement, and choice: Women and minority students in science and mathematics. Review of Research in Education, 16, 153. http://doi.org/10.2307/1167352

Osborne, J. W., & Costello, A. B. (2009). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Pan-Pacific Management Review, 12, 131–146.

Pajares, F. (2002). Gender and perceived self-efficacy in self-regulated learning. Theory Into Practice, 41(2), 116–125.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. Journal of Educational Psychology, 82(1), 33.

Pintrich, P. R., Zusho, A., Schiefele, U., & Pekrun, R. (2001). Goal orientation and self-regulated learning in the college classroom: A cross-cultural comparison. In F. Salili, C. Y. Chiu, & Y. Y. Hong (Eds.), Student motivation (pp. 149–169). US: Springer.

Pomerantz, E. M., Altermatt, E. R., & Saxon, J. L. (2002). Making the grade but feeling distressed: Gender differences in academic performance and internal distress. Journal of Educational Psychology, 94(2), 396–404. http://doi.org/10.1037/0022-0663.94.2.396

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. Multivariate Behavioral Research, 42(1), 185–227.

Rayman, P., & Brett, B. (1995). Women science majors: What makes a difference in persistence after graduation? The Journal of Higher Education, 66(4), 388. http://doi.org/10.2307/2943794

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.

Sargent, J. F., Jr. (2014). The U.S. Science and Engineering Workforce: *Recent, Current, and Projected Employment, Wages, and Unemployment*. Congressional Research Service. Available at: https://www.fas.org/sgp/crs/misc/R43061.pdf

Schunk, D. H., & Ertmer, P. A. (2000). Self-regulation and academic learning: Self-efficacy enhancing interventions.

Selimbegovic, L., Chatard, A., & Mugny, G. (2007). Can we encourage girls' mobility towards science-related careers? Disconfirming stereotype belief through expert influence. European Journal of Psychology of Education, 22(3), 275–290.

Serbin, L. A., O'Leary, K. D., Kent, R. N., & Tonick, I. J. (1973). A comparison of teacher response to the preacademic and problem behavior of boys and girls. Child Development, 44, 796–804.

Sha, L., Schunn, C., & Bathgate, M (2015). Measuring choice to participate in optional science learning experiences during early adolescence. Journal of Research in Science Teaching, 52(5), 686–709. http://doi.org/10.1002/tea.21210

Sha, L., Schunn, C., Bathgate, M., & Ben-Eliyahu, A. (2016). Families support their children's success in science learning by influencing interest and self-efficacy. Journal of Research in Science Teaching, 53(3), 450–472.

Shapiro, J. R., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. Sex Roles, 66(3–4), 175–183.

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science?: A critical review. American Psychologist, 60(9), 950–958. http://doi.org/10.1037/0003-066X.60.9.950

Stark, R., & Gray, D. (1999). Gender preferences in learning science. International Journal of Science Education, 21(6), 633–643.

Staurt, H. C., Shock, N. W., Breckenridge, M. E., & Vincent, E. L. (1953). Physical growth and development. In The Adolescent: A Book of Readings (pp. 88–138). Ft Worth, TX: Dryden Press, xviii, 798 pp. http://doi.org/10.1037/11402-004

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. Journal of Personality and Social Psychology, 69(5), 797.

Steinberg, L. (2005). Cognitive and affective development in adolescence. Trends in Cognitive Sciences, 9(2), 69–74. http://doi.org/10.1016/j.tics.2004.12.005

Sungur, S., & Tekkaya, C. (2003). Students' achievement in human circulatory system unit: The effect of reasoning ability and gender. Journal of Science Education and Technology, 12(1), 59–64.

Tucker, S. A., Hanuscin, D. L., & Bearnes, C. J. (2008). Igniting girls' interest in science. Science, 319, 1621–1622.

Webb-Williams, J. (2014). Gender differences in school children's self-efficacy beliefs: Students' and teachers' perspectives. Educational Research and Reviews, 9, 75–82.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology, 75(5), 1182–1189.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: Mesa Press.

Yenilmez, A., Sungur, S., & Tekkaya, C. (2006). Students' achievement in relation to reasoning ability, prior knowledge and gender. Research in Science & Technological Education, 24(1), 129–138.

Zeldin, A. L., Britner, S. L., & Pajares, F. (2008). A comparative study of the self-efficacy beliefs of successful men and women in mathematics, science, and technology careers. Journal of Research in Science Teaching, 45(9), 1036–1058. http://doi.org/10.1002/tea.20195

Zimmerman, B. J., Boekarts, M., Pintrich, P. R., & Zeidner, M. (2000). A social cognitive perspective. In Handbook of self-regulation, Oxford, UK: Academic Press, Chaper 2.

Zimmerman, B., & Bandura, A. (1994). Impact of self-regulatory influences on writing course attainment. American Educational Research, 31, 845–862.

Zohar, A., & Bronshtein, B. (2005). Physics teachers" knowledge and beliefs regarding girls" low participation rates in advanced physics classes. International Journal of Science Education, 27(1), 61–77.

Supporting Information

Additional Supporting Information may be found in the online version of this article.