

## RESEARCH ARTICLE

# Using principles of cognitive science to improve science learning in middle school: What works when and for whom?

Christian D. Schunn<sup>1</sup>  | Nora S. Newcombe<sup>2</sup>  | Louis Alfieri<sup>1</sup> | Jennifer G. Cromley<sup>2</sup>  | Christine Massey<sup>3</sup>  | Joseph F. Merlino<sup>4</sup>

<sup>1</sup>University of Pittsburgh, Pittsburgh, USA

<sup>2</sup>Temple University, Philadelphia, USA

<sup>3</sup>University of Pennsylvania, Philadelphia, USA

<sup>4</sup>The 21st Century Partnership for STEM Education, USA

## Correspondence

Christian Schunn, University of Pittsburgh, 3939 OHara St., Pittsburgh, PA 15260, USA.  
Email: schunn@pitt.edu

## Funding information

Institute of Education Sciences, Grant/Award Number: R305C080009

## Summary

Four principles of cognitive science were used to make systematic revisions in middle school science instructional modules from two kinds of curriculum: one popular textbook series and one popular hands-on series (two modules each). Schools were randomly assigned to 1 of the 3 arms (cognitive science modifications with professional development, active control with professional development, or business-as-usual). Two cohorts of students were followed in each arm for each setting. There were significant benefits of the cognitive science intervention, but the nature of effects varied for the two settings and curricula. For the text-based curriculum, positive effects of cognitive science modifications were concentrated in classrooms with lower proportions of underrepresented minority students. For the hands-on curriculum, there were positive effects that were not linked to school composition. Participation in the active control did not significantly improve student learning. Implications for policy and research are discussed.

## KEYWORDS

cognitive science, contextual interactions, curricular modifications, middle school science

## 1 | INTRODUCTION

It is vital to make science accessible to all learners (Alberts, 1999; Greene, 2008; Marincola, 2006). However, the United States, along with other countries, has lagged in student performance in science at the fourth- or eighth-grade levels, and improvements have not been evident even as policy makers voice concern (Provasnik et al., 2012). Supporting science performance likely requires several strategies, ranging from student-centered social programs to administrative reform and restructuring. One important potential strategy, however, involves using principles from cognitive science to guide the design of instructional materials, given that many principles have already been well established in small-scale efficacy studies. For example, comparing cases (e.g., analogs are compared to highlight the similarities/features that make the cases analogous; Alfieri, Nokes-Malach, & Schunn, 2013) has been shown to be an effective method in laboratory studies and in short classroom interventions. In this paper, we report findings from a large study designed to evaluate the effectiveness of four principles taken from cognitive science, implemented together as a package. Analyses examine generality of effects across contextual factors.

### 1.1 | Four principles for improving science curricular materials

We modified curricular materials using a set of four principles that we derived from cognitive science. We chose this approach over other possibilities for several reasons. First, consider an alternative approach in which we evaluated one principle at a time. Although isolated cognitive principles are *theoretically* interesting to understand and evaluate, in *practice*, the cumulative and possibly interactive benefits of a family of interventions are likely to be more relevant to producing meaningful learning gains in an extended curriculum. In the world of educational policy, it is important to demonstrate cumulative impact, rather than staying purely at the grain size of isolated principles. Second, another alternative might have been to implement every change we could find for which cognitive science provides some basis in theory and data. However, such an effort would likely have overwhelmed the teachers asked to put the curriculum into practice.

The four principles in this project were (a) identifying misconceptions that provide barriers to learning, (b) using contrasting cases to enhance learning, (c) teaching students to read visualizations, and (d) using spaced testing to consolidate learning. They were selected to

span a range of learning phases (from learning to retention) and learning content (from conceptual to visual): identify critical conceptual barriers, make science ideas salient, improve reasoning practices, and increase retention. Furthermore, they worked well together, for example, identifying potential misconceptions had implications for contrasting cases, changes in visualizations, and warm-up questions. We briefly present both the prior empirical evidence for each principle and examples of how the principle was used to modify curriculum materials in our study.

### 1.1.1 | Identifying misconceptions and student prior knowledge

Students are likely to enter science instruction holding entrenched misconceptions (e.g., teleological beliefs about the mechanisms of evolution; Bishop & Anderson, 1990), insufficiently developed naïve ideas (e.g., overly simplistic causal models; Chi, 1992), or a lack of understanding of scales for very large or very small magnitudes (Tretter, Jones, & Minogue, 2006). When students have misconceptions (entrenched misunderstandings that conflict with new information; Carey, 1991, 2009; Chi, 1992), they need to be addressed and corrected to maximize learning. Teacher-driven explanations to correct these misunderstandings are often not enough to undo them. An example from the current intervention involved misconceptions about which features distinguish living and nonliving things. For example, movement is often thought to be an indicator of life, and consequently, it is thought that seemingly nonmoving things (trees, plants) must be nonliving. Such common misconceptions relevant to each curricular module were identified using the research literature, and then addressed with case comparisons and/or visualization exercises (discussed below).

### 1.1.2 | Case comparisons

Case comparisons (i.e., the simultaneous contrast of cases to help learners to understand similarities and differences) have been shown to lead consistently to improved learning (Alfieri et al., 2013), to greater learning than even the sequential study of five cases (Gentner, Loewenstein, & Thompson, 2003), and transfers to tests and problem solving (Gick & Holyoak, 1983). For example, in our modified materials, before the existing instruction explaining the characteristics of all members of the plant kingdom, students were asked to compare individual members of the kingdom mounted on sets of cards in order to

notice their commonalities and to distinguish those from the characteristics unique only to specific members of the kingdom. When a new organism within the plant kingdom is introduced, students could then transfer their knowledge of those commonalities and consequently have appropriate expectations as to the characteristics of this new organism. See Figures 1 and 2 for examples.

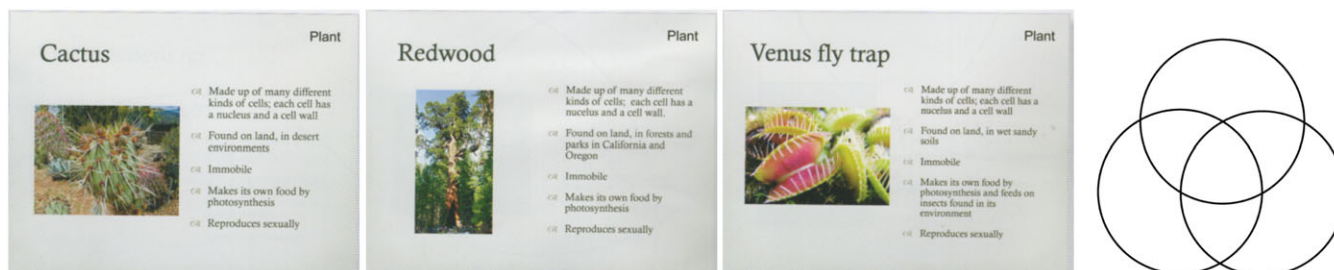
### 1.1.3 | Visualization exercises

Science curricula are filled with complex graphs and images (drawings, photographs, diagrams, etc.) that are critical to content understanding. Teachers are often surprised when skilled students are unable to interpret components of complex diagrams (Hegarty, Kriz, & Cate, 2003). Many students lack the skills to appreciate images fully (Berthold & Renkl, 2009), and such weaknesses can lead students to ignore them entirely (Bartholomé & Bromme, 2009). Well-designed exercises can assist students in learning to decode these images successfully (e.g., how to interpret arrows, captions/labels, scale, or colorization).

We created visualization exercises that helped teachers guide discussions of images and their components (e.g., conventions and interpretations) to clarify what students were seeing, to explain typical conventions within such representations, and to increase the likelihood that students would be able to discern the important information from such representations independently in the future. Initially, students were asked to attend only to single characteristics (e.g., relative scales or colorization) of simpler images (e.g., photographs of organisms), but as a unit progressed, discussions transitioned to increasingly complex images (e.g., the processes of photosynthesis and cellular respiration). Figures 3 and 4 display topics of earlier and later visualization activities.

### 1.1.4 | Spaced testing

Forgetting is a general instructional problem. Revisiting previous concepts at delay even when equating for total study time can greatly reduce forgetting (the “spacing” effect; Rohrer & Pashler, 2007), and testing can help students remember content better than simply reviewing it for the same amount of time (the “testing” effect; Rowland, 2014). We combine the two effects using two forms of spaced testing: daily warm-up questions and end-of-section quizzes. Daily warm-up questions prompted students to explain a previous day’s main ideas to recapitulate the purpose and content of the previous activity(s). The warm-up question was generally chosen to connect



**FIGURE 1** The three cards present students with cases of Venus fly trap, cactus, and redwood plants. Students are asked to compare the three by completing a Venn diagram to recognize which characteristics are shared by members of the plantae kingdom [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

If each set of parent Labrador retrievers has 16 puppies, how many will have black coats? How many will have yellow coats? Fill in your predictions, and then the actual numbers.

Family	Parents	Predicted # (out of 16)		Actual # (out of 16)	
		Black	Yellow	Black	Yellow
A	Black-Yellow				
B	Black-Black				
C	Black-Black				
D	Yellow-Yellow				
E	Yellow-Black				

What do you think "dominant" means? \_\_\_\_\_

\_\_\_\_\_

What do you think "recessive" means? \_\_\_\_\_


\_\_\_\_\_

Which color coat is dominant? \_\_\_\_\_

How do you know? \_\_\_\_\_

Which color coat is recessive? \_\_\_\_\_

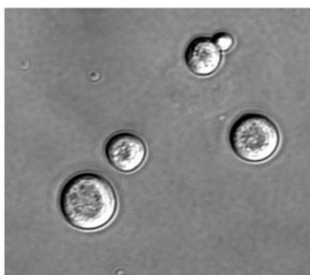
How do you know? \_\_\_\_\_



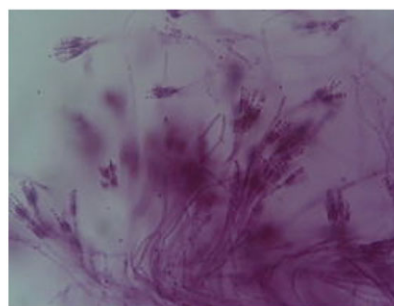
**FIGURE 2** This worksheet accompanies a lesson on heredity (i.e., dominant and recessive genes). Students are asked first to predict the color of the puppies before displaying the actual outcomes. Students can then compare families with each other (as cases) to determine how coat color is determined by heredity



Destroying angel



Yeast



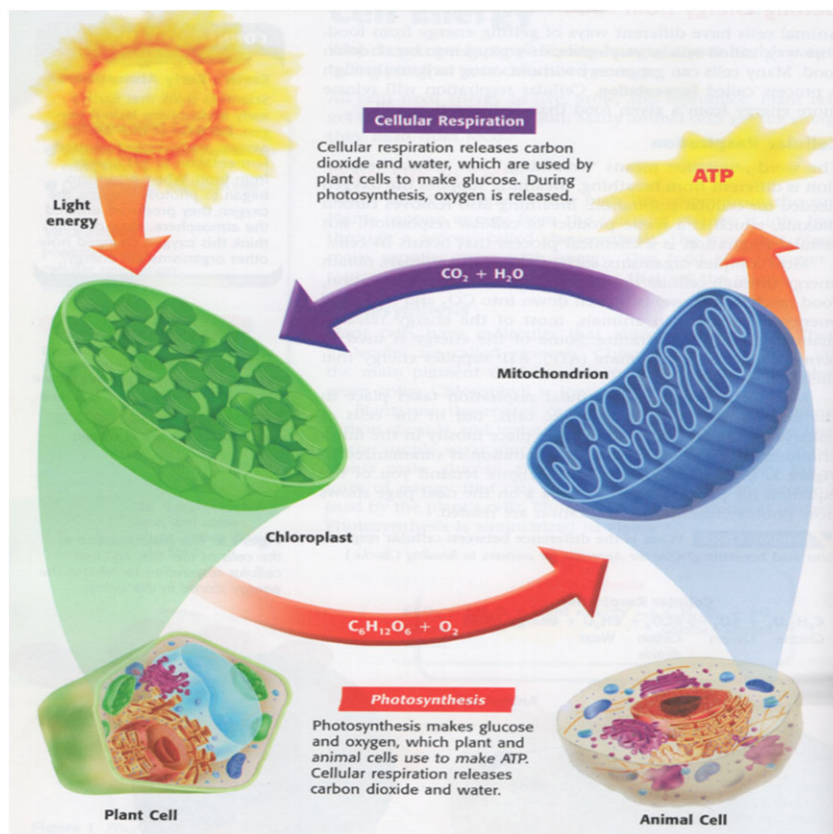
Penicillium

**FIGURE 3** While displaying the images of penicillium, yeast, and destroying angel, the teacher prompts students to consider and discuss the apparent and actual sizes of the organisms (explaining that discrepancy by introducing the term, relative scale) [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

with previous content because the day's activities were intended to build on that content. End-of-section quizzes included questions from both the current and previous sections of the unit. These additions to the curriculum provide additional exposure to crucial content, but their key mode of action has been shown to be their spacing, because massed practice and review is not effective.

**1.2 | Key contextual variables**

Although there is reason to suppose that all four of these principles work, at least in isolation and in small-scale implementation, there are several contextual factors to consider in a real-world trial that may affect the results of these interventions. We selected three important variables of this kind to structure our analyses.



**FIGURE 4** While displaying the diagram, the teacher leads a discussion that explores with students the use of shadows to indicate the zoom-out convention, the use of arrows, the use of colors within both the cells and arrows, and the use of labels [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

### 1.2.1 | Immediate versus delayed effects

Cognitive science-based interventions, centered in learner materials, are often relatively straightforward curricular modifications. Even so, the extent to which teachers are already entrenched in existing teaching practices may increase the amount of time required to see changes in practices, and thus improvements in learning outcomes. For this reason, we designed the research to include teachers who would be teaching the modified curriculum for a second time, as well as implementing it on an initial pass.

### 1.2.2 | Hands-on versus textbook science

In the United States, middle school science curricula vary greatly in terms of whether learning is done primarily through textbook reading (with classroom discussion and worksheet activities linked to the readings) or whether learning is done primarily through hands-on investigations. It is usually possible to cover more content via textbook than via hands-on inquiry (Mayer, 2004). It is possible that cognitive science interventions might be differentially effective for different types of curriculum. For instance, the interventions involved active learning, and their benefits might be weaker or stronger within a curriculum that already involves active learning. For this reason, we implemented our cognitive science modifications both for a text-based curriculum and for a hands-on curriculum, and conducted two parallel trials.

### 1.2.3 | Learning context

Several disruptive situational factors might undermine the effect of an intervention. For example, teachers might not receive professional development in how to implement a curriculum (Abdal-Haqq, 1996;

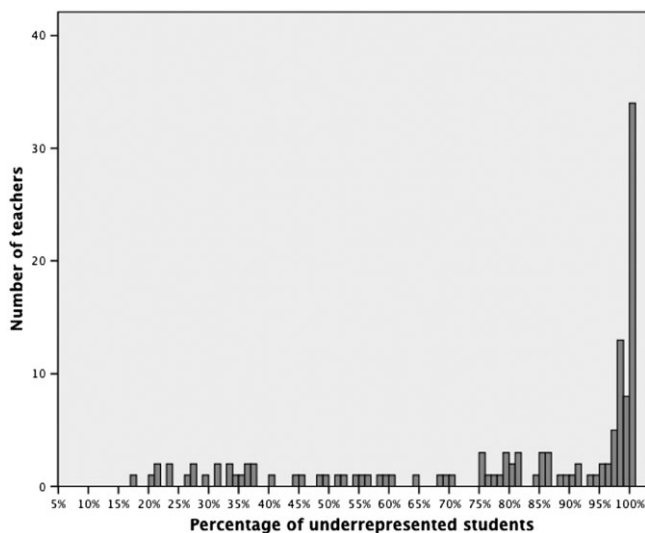
Cambone, 1995; Corcoran, 1995), teachers might not have sufficient control of their classes to implement the intervention fully (Emmer & Stough, 2001), students may not be attending school often enough to receive an intervention (Spencer, 2009), or students may be failing to pay attention due to hunger or lack of sleep (Symons, Cinelli, James, & Groff, 1997). Disruptive factors co-occur in complex combinations and can occur at many levels—either specific to the classroom (e.g., a group of students are collectively unruly), to the teacher (e.g., the teacher has poor classroom management skills), or to the school (e.g., extracurricular announcements regularly take priority over quiet classroom time). Sadly, socioeconomic status and ethnicity are correlated with disruptive factors. In the United States, students from ethnicities traditionally underrepresented in science careers (Barton & Coley, 2010; Beede et al., 2011; Byars-Winston, Estrada, & Howard, 2008) experience a variety of challenges, including less skilled teachers (Lankford, Loeb, & Wyckoff, 2002; Shen, 1997) and often-disrupted classrooms (Ingersoll, 2004; Weiner, 2003). Note that these disruptive factors are at the classroom, teacher, or school levels, and thus the effects of these factors are associated with the proportion of students who are from underrepresented minorities, not with individual student characteristics. Such disruptive factors might also interact with curriculum type.

### 1.3 | Analytical concerns in investigating effect moderation

We aimed to analyze this randomized controlled trial of our intervention considering the factors just discussed. There are two different

ways to do so. First, the effects can be disaggregated at the student level (e.g., students from ethnicities traditionally underrepresented in science careers; hereafter referred to as *underrepresented students*). States, districts, schools, and teachers are required to report their student performance data in the United States in this way. But such disaggregation, although policy relevant, is more distant from contextual factors than ideal; it puts together underrepresented students in supportive contexts with underrepresented students in nonsupportive contexts. Second, the effects can be disaggregated into classrooms grouped by overall student characteristics (i.e., classrooms with a higher/lower proportion of students from underrepresented ethnicities). Such a disaggregation, although not following the typical reporting requirements, better matches the decision context of policy makers (i.e., they make decisions about whole schools or districts).

Note that the individual, classroom, teacher, and school factors that could influence intervention efficacy are likely to be correlated. For example, students most likely to lack learning support resources at home are also more likely to be vulnerable to low expectations, more interruptions in classrooms, classmates with less interest in science, and teachers with less teaching experience and knowledge of science (Condon & Roscigno, 2003; Henry, Bastian, & Smith, 2012; Lankford et al., 2002). This complexity has two implications for studying the ways in which context moderates the effectiveness of the intervention. First, large correlations among potential moderators preclude pulling them apart statistically. Second, the highly unequal distribution of students to schools means that traditional interaction analyses within one dataset are underpowered because of problems specific to the Hierarchical Linear Modeling (HLM) approaches typically used for analyzing large randomized control trials (RCTs). HLMs often have low statistical power at the teacher level despite high statistical power at the student level. Disaggregation by ethnicity causes problems at the teacher level because of the highly uneven distribution of student ethnicities across teachers. For example, as shown in Figure 5, almost one-third of the teachers in our textbook sample



**FIGURE 5** A histogram from the textbook curriculum randomized control trial illustrating that almost one-third of our teachers have classes that are composed entirely of students who are of an underrepresented minority ethnic status in science

taught only underrepresented minority (URM) students. Analyses of within-teacher interactions by ethnicity lose these teachers entirely, greatly reducing statistical power and representativeness to the larger sample.

As a related issue, interaction analyses of variables are limited in power by the cell associated with the weakest power. Thus, if an effect is unreliable for one group but highly reliable for another group, both main effects and traditional interaction analyses will fail to find any statistically significant outcome. Because disruptive factors can be highly variable across teachers and schools within our subgroups, it may be that nothing can be concluded from one subgroup but strong inferences can be drawn for another subgroup. Further, because ethnicity likely captures complex, multi-component underlying effects, many of which are not actually at the student level, student level regressors may be differentially predictive of performance across contexts. For these reasons, we did not adopt the current standard approach to building a single model for the larger dataset with interaction terms because it (a) has low statistical power and (b) will therefore often fail to inform the practice or research community of important patterns. Instead, we built separate models.

## 2 | METHODS

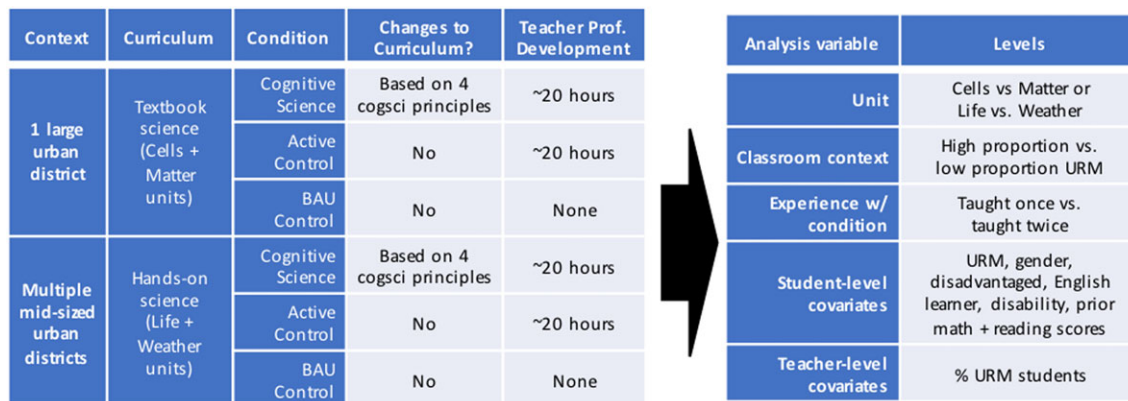
### 2.1 | Overview

We analyze data from two large RCTs: one conducted in a large urban district in the US Northeast using a textbook curriculum and another conducted in several mid-sized urban districts in the US Southwest using a shared hands-on curriculum. Each RCT examines learning on two different 3-month-long instructional modules from the curriculum and involves the same three conditions: a cognitive science condition in which the modules were modified using the cognitive science principles, an active control condition in which teachers received professional development on underlying science content, and a business-as-usual control condition (see Figure 6 for a study overview). Schools were randomly assigned to condition within regions of the large urban district in the textbook RCT and within school districts in the hands-on RCT. Because curricula are adopted at the level of school districts, these parallel trials of necessity also involved different geographic and sociocultural settings.

### 2.2 | Curricula

#### 2.2.1 | Textbook

The textbook-based curriculum was the widely used textbook, *Holt Science and Technology* (Holt, Rinehart, & Winston, 2007). It is part of a Short-Course Series self-described as combining the content teachers need with an accessible design, student-friendly narrative, and vivid visuals. In comparison with the hands-on curriculum (see below), learning in the Holt curriculum is primarily driven by readings rather than active experimentation. We report here results from two Holt units: *Cells, Heredity, and Classification* (the biological science unit; subsequently referred to as Cells) and *Introduction to Matter* (the physical science unit; subsequently referred to as Matter), which were



**FIGURE 6** Study design overview. Left shows the contexts and curricula types in which the two randomized control trials were implemented, the conditions for each randomized control trial, and the features that varied across conditions (curriculum changes and teacher professional development). The right shows the key variables used in data analysis. URM = underrepresented minority [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

taught at the beginning of the seventh and eighth grade school years, respectively.

### 2.2.2 | Hands-on

The hands-on curriculum was also one that is commonly used at the middle school level, the *Full Option Science System* (FOSS; developed by The Lawrence Hall of Science). In comparison with the Holt curriculum, learning in the FOSS curriculum occurs primarily within and surrounding the activities rather than through reading. We report here results from two FOSS units: *Diversity of Life* (the biological science unit; subsequently referred to as Life) and *Weather and Water* (the physical science unit; subsequently referred to as Water). Both Life and Water were taught in the participating districts during the sixth grade school year, but at varying times across schools because they rotated the kits from school to school during the year.

## 2.3 | Participants

Participants in the textbook curriculum came from 97 schools in one urban district in the northeast of the United States. Given high year-to-year turnover in teachers and teaching assignments, the study involved 229 teachers across the two cohorts of students and two grade levels. Participants in the hands-on curriculum came from 116 teachers in 65 schools from 6 urban school districts in two different cities in a southwestern state in the United States. There were equivalent student attrition rates (enrolled but did not complete the end-of-unit test) of approximately 10% per year across conditions for both curricula. Teacher attrition rate from year to year was higher and more variable across curricula, with approximately 40% of teachers of the Hands-on curriculum and 50% of the teachers of the Textbook curriculum not completing the second year, primarily due to changes in teaching assignments within and between schools.

The (typical) relatively high attrition/reassignment rate in urban districts from year to year meant that there were a greatly reduced number of teachers who taught the given unit for a second time. Therefore, to examine the relative intervention effects as a function of teachers' experience implementing the curriculum, we collapse students of teachers who entered the study in the second cohort of

students with all students in the first cohort. We also did so for students in the control condition because the teachers who remain stable may be systematically different. In the textbook RCT, this aggregation approach produced approximately 6,400 students who were considered part of the first year of implementation (across conditions) and approximately 3,200 students in the second year of implementation (across conditions). In the hands-on RCT, the corresponding numbers were approximately 7,600 students in the first year and approximately 4,200 students for the second year of implementation.

To examine effects separately by context (high vs. low proportion of URM), the data from each implementation year were divided into subgroups by the teacher's proportion of URM students (greater or less than .80; other cutoffs produced similar results). Table 1 presents the number of students within each subgroup for each curricular module. Table 1 also includes critical individual demographic variables that previously have been associated with student test performance in science: gender, disadvantaged (student received free or reduced-price lunch), English language learners, previous achievement (average score in reading and mathematics on state tests), URM status (African American, Hispanic/Latino, and Native American), and disability status.

The study was approved by institutional IRBs. Teachers provided consent for their data, presented here in abbreviated form. The child data qualified as Exempt and therefore did not require parental consent. Parents were sent an informational letter, as recommended best practice.

## 2.4 | Materials and procedure

### 2.4.1 | Cognitive science condition

The cognitive science-based intervention incorporated three major components as described in the introduction—case comparisons, visualization exercises, and spaced testing in the form of daily warm-up questions and repeated/delayed questioning on quizzes and tests—that were interleaved into the base curricular units. Research on misconceptions informed the selection of which concepts needed the most support. Table 2 displays the number of class meetings (days) that included each component of the intervention and lists the specific days for one of the units. As can be seen, the intervention served as a

**TABLE 1** Student characteristics by unit, implementation year, and relative proportion of URMs in the classroom

Textbook-based	Cells	1st year implemented	Higher-URM-proportion classes		Lower-URM-proportion classes		Matter	Higher-URM-proportion classes		Lower-URM-proportion classes			
			n	M; SD	n	M; SD		n	M; SD	n	M; SD		
Textbook-based	Cells	1st year implemented	3,687	M = 7.5; SD = 3.3	2,782	M = 9.6; SD = 3.7	Posttest score Percent URM Female (%) Disadvantaged (%) English learners (%) z previous achievement Disability (%)	4,125	M = 7.6; SD = 3.4	2,290	M = 9.3; SD = 3.7		
			96	50	89	13		96	49	88	7	15	14
			50	89	13	96		49	88	7	15	14	
			89	13	96	49		88	7	15	14		
			13	96	49	88		7	15	14			
			96	49	88	7		15	14				
			50	89	13	96		49	88	7	15	14	
			89	13	96	49		88	7	15	14		
			13	96	49	88		7	15	14			
			96	49	88	7		15	14				
Hands-on	Life	1st year implemented	2,596	M = 8.3; SD = 3.0	4,983	M = 10.2; SD = 3.4	Water	2,728	M = 8.1; SD = 2.9	4,929	M = 10.2; SD = 3.3		
			92	51	80	12		92	51	81	11	8	7
			51	80	12	92		51	81	11	8	7	
			80	12	92	51		81	11	8	7		
			12	92	51	81		11	8	7			
			92	51	81	11		8	7				
			51	80	12	92		51	81	11	8	7	
			80	12	92	51		81	11	8	7		
			12	92	51	81		11	8	7			
			92	51	81	11		8	7				
2nd year implemented	Life	2nd year implemented	1,840	M = 7.8; SD = 3.5	1,302	M = 9.1; SD = 3.8	Posttest score Percent URM Female (%) Disadvantaged (%) English learners (%) z previous achievement Disability (%)	1,721	M = 8.5; SD = 3.7	1,547	M = 10.3; SD = 3.8		
			96	47	66	6		96	48	69	7	15	10
			47	66	6	96		48	69	7	15	10	
			66	6	96	48		69	7	15	10		
			6	96	48	69		7	15	10			
			96	47	66	6		96	48	69	7	15	10
			47	66	6	96		48	69	7	15	10	
			66	6	96	48		69	7	15	10		
			6	96	48	69		7	15	10			
			96	47	66	6		96	48	69	7	15	10
2nd year implemented	Life	2nd year implemented	944	M = 8.6; SD = 3.0	3,151	M = 10.2; SD = 3.4	Posttest score Percent URM Female (%) Disadvantaged (%) English learners (%) z previous achievement Disability (%)	1,216	M = 8.0; SD = 2.9	3,235	M = 10.0; SD = 3.2		
			92	50	42	54		92	51	35	60	9	7
			50	42	54	92		51	35	60	9	7	
			42	54	92	51		35	60	9	7		
			54	92	51	35		60	9	7			
			92	50	42	54		92	51	35	60	9	7
			50	42	54	92		51	35	60	9	7	
			42	54	92	51		35	60	9	7		
			54	92	51	35		60	9	7			
			92	50	42	54		92	51	35	60	9	7

Note. URM = underrepresented minority.

**TABLE 2** Days modified by cognitive science intervention

Intervention component	Percentage (# of days/# of days total)	Specific days
Case comparisons	29 (14/48)	1, 2, 3, 4, 11, 12, 13, 14, 23, 24, 35, 36, 44, & 45
Visualization exercises	31 (15/48)	1, 6, 7, 15, 16, 17, 19, 20, 25, 27, 28, 30, 33, 37, & 46
Warm-up questions	88 (42/48)	1–8, 10–21, 23–30, 32–40, 42–47
Assessments (quizzes/tests)	10 (5/48)	9, 22, 31, 41, & 48

supplement to the standard curriculum, not as a replacement, but total time on the unit was controlled across conditions. Case comparisons were designed to better set up the instruction found within each chapter (textbook) or extended investigation (hands-on), whereas visualization exercises were more dispersed as necessitated by the visual demands of the chapters/extended investigation. The intervention was integrated into the entire unit and its implementation took place during the same three- to four-month time span when each unit was ordinarily delivered.

Teachers used the existing curriculum materials except (a) they added contrasting cases (usually in the form of teacher PowerPoints and student worksheets), (b) brief visualization exercises (similarly a mix of PowerPoints and student worksheets) that usually directly connected to visualizations already found in the existing curricular materials, although sometimes were simplified for instructional effect, (c) daily warm-up questions, and (d) end-of-section tests (some shorter and some longer). Teachers dropped whatever tests they would have used and whatever warm-up questions they would have used (if any); note that prior research suggests their own tests and warm-up questions would have not involved content from prior weeks of instruction and teachers informally reported this change as well. To keep the same total number of weeks, we also had suggestions for content in the existing curriculum that seemed especially ineffectual (e.g., starting a unit by singing a song about the Grand Canyon). The teachers were given a teacher's manual that overviewed all the activities (newly inserted plus recommended existing materials) in a clear day-by-day fashion.

In the summer before implementing a modified unit, teachers attended 3 days of professional development that explained the rationale for the revised activities (the cognitive principles involved) and how to implement those activities. Each teacher received a binder that included a written introduction to the intervention: its scope, contents, and goals, along with a CD of prepared PowerPoint presentations and any specialized materials needed to complete activities in the modified curriculum. The materials were not required scripts, but rather suggestions; teachers were told they could rephrase the materials to meet the needs of their classrooms. Teachers also attended four after-school, small group follow-up sessions (one per month of the unit) to discuss challenges and successes.

Students participated in the study via their regular science classes, and also completed the regular textbook-based activities and tests included within the curriculum. Overall, the integration of the intervention into the standard curriculum was seamless and it is unlikely that students' subjective experiences were of participating in an experimental curriculum. Teacher surveys confirmed that the conditions had equivalent amounts of hours of science per week.

When new teachers moved into a cognitive science school in the second year, they received make-up professional development during the summer and the school year similar to that provided in the first year. Returning teachers were provided with a brief overview of minor improvements to the intervention based on teacher feedback, and with additional after-school follow-up sessions similar in format to those in the first year.

#### 2.4.2 | Active control condition

Teachers in the active control condition attended the same amount of professional development as the cognitive science teachers (3 days in the summer and four follow-up after-school sessions), but these sessions focused on the curriculum's underlying science content, rather than on pedagogy or principles of learning. The training sessions were provided by content experts who often implement these forms of content-deepening training for teachers (e.g., university faculty and museum educators) and were designed for teachers as adult learners. These teachers did not receive any modifications to the standard curriculum. Such a control condition is policy relevant in that science teachers often experience such relatively brief content-deepening professional development experiences in the summer or during the school year (Garet, Porter, Desimone, Birman, & Yoon, 2001). However, the primary purpose of this condition for the study was to rule out attributing any benefits of the cognitive science condition to a Hawthorne effect (i.e., improvements that stem from teacher *perceptions* of being in an experimental or higher quality condition), either as a function of professional development involvement or as a byproduct of participation in professional learning communities (Adair, 1984; Cook, 1967; Jones, 1992). Teachers in this condition self-reported enjoying the training and said that it was relevant to their classroom practices (Desimone & Hill, 2017). However, the amount of content provided in this intervention is relatively modest and thus is not a strong test of the benefits of teacher content knowledge on student learning.

Students of teachers in the content training condition attended their scheduled science classes and completed activities included in the standard curriculum. After the unit, students completed the end-of-unit test that was standard to all three arms/conditions of our study.

When new teachers moved into a content arm school in the second year, they received make-up professional development during the summer and the school year, similar to that provided in the first year. Returning teachers were provided additional content training matched in duration to the follow-up training provided to the cognitive science teachers.



### 2.4.3 | Business-as-usual condition

Teachers in the business-as-usual control arm received neither professional development nor the modified curriculum, but they consented to and were made aware of their participation within the larger study. Students attended their scheduled classes, completed only the activities included in the standard curriculum, and then completed our end-of-unit test. Teacher surveys verified that teachers implemented the core curricula and did not use any of the cognitive science modifications.

### 2.4.4 | End-of-unit tests

The end-of-unit assessments each involved 18 questions related to the curricular content (see online Appendix for a posttest for one of the curriculum units). The questions for each unit's assessment were developed by sampling items from various item pools (released state tests, released NAEP and TIMSS items; Porter, Polikoff, Barghaus, & Yang, 2013). The base units and the pool of possible test items were content analyzed using the Survey of Enacted Curriculum framework constructed by Porter and colleagues. To avoid any potential bias in question content or format towards the cognitive science condition, questions were sampled using an algorithm such that the content of the questions maximally matched the content of the base curriculum (i.e., with no regard to the modifications). The tests showed good reliability given the diversity of content embedded in each test (Cronbach  $\alpha$ 's ranged from .65 to .74; mean  $\alpha = .70$ ). Students in all conditions completed the test around the same time for a given unit.

### 2.4.5 | Fidelity of implementation

Teachers across all conditions completed surveys at the end of each instructional unit to document amount of instructional time and the extent to which they implemented the cognitive science principles in their instruction. For the two control conditions, this assessed the possibility of leakage across conditions, and for the cognitive science condition, this assessed fidelity of implementation. The surveys emphasized the importance of honest reporting for the scientific goals of the study, and questions were worded to avoid the appearance of only one socially acceptable response (e.g., by providing contextually reasonable factors for nonimplementation). The methods for developing and validating the survey, along with the key findings, are presented in depth in Desimone and Hill (2017). Overall, these data showed that teachers generally implemented the conditions as designed. In addition, Cognitive Science condition teachers' classrooms were visited at least once during implementation in two of the school districts, often across multiple days, and fidelity of implementation in these visits was observed to be high.

## 2.5 | Statistical approach

### 2.5.1 | Modeling the effects of our intervention

To develop an effective statistical approach for this complex situation but avoid Type I errors from broad model search, the hierarchical model analytic approach was initially explored within the Cells unit data to determine an approach that produced good fitting

models and also stable results across many model variations. This approach was then applied to the analysis of the other three units. A two-level model (students nested within teachers) was selected because most of the schools had only one science teacher per grade (and adding school did not account for more variance), there were often too few classrooms per teacher to include both classroom and teacher in the model, and teacher was a larger source of variance than school or classroom once student predictors were included in the model.

Each unit was modeled separately for best-fit purposes first (i.e., determine significant covariates). Potential student characteristics included underrepresented status (1 = URM), gender (1 = female), disadvantaged status (1 = qualifies for free or reduced-price lunch), English learner status (1 = English language learner), disability status (1 = reported to have a disability), and prior achievement scores (mean of prior 2 years of math and reading state test z-scores).<sup>1</sup> Averaging prior achievement scores across grades allowed students who were missing a score on one of the four measures to be included. Further, although math and reading scores differently predict performances on specific questions on the end-of-unit test, there was little differential predictiveness for reading versus math for the summed end-of-unit test score.

At the second level of our nested models, we included dummy codes for treatment (cognitive science intervention—CogSci, or content training—Content), and the percentage of the teacher's students who were of a traditionally underrepresented ethnic status (Percent URM).

### 2.5.2 | Taught 1 year versus 2 years

To ensure appropriate comparisons of the cognitive science intervention and content training to the most relevant control condition data, the data were split and modeled by the year of implementation (in addition to being split by proportions of URM students within classrooms). Thus, the combined codes consist of dummy variables for six mutually exclusive categories: CogSci1 indicates that students received the cognitive science treatment when their teacher was implementing it for the first time. CogSci2 indicates that students received the cognitive science treatment when their teacher was implementing it for a second time. Content1 indicates that students were instructed by a teacher who received content training for the first year. Content2 indicates that students were instructed by a teacher who received content training for a second year. Control1 and Control2 indicate that students participated in the business-as-usual control when their teacher was participating for his/her first or second year, respectively.

### 2.5.3 | Imputations and modeling

Although few data points were missing relative to the size of the sample, data were missing-not-at-random, which necessitated 10 imputations to be generated using Mplus Version 6.12 (Muthén

<sup>1</sup>These categories are the ones defined by the federal government in the United States. Underrepresented refers to ethnicities underrepresented in science: African American, Latinx/Hispanic, Native American, and Pacific Islander. Disadvantaged refers to income levels that qualify the student for free or reduced lunch at school.

& Muthén, 2011) with the student's teacher as the clustering variable. In Cells for example, underrepresented status (<1% missing out of 9,611 students), disadvantaged status (4.5%), English learner status (<1%), disability status (6.3%), and previous achievement (5.9%) were all imputed using those same variables along with cohort, school, and age to impute missing cells. No imputations were conducted for missing end-of-unit test scores (the dependent measure). Imputations were then accessed by HLM 7.01 (Raudenbush, Bryk, & Congdon, 2004), which averaged results across imputations.

### 3 | RESULTS

As indicated in Table 1, data were available for a variety of individual and contextual factors to be entered in the models (e.g., the student's gender, disability status, previous achievement scores, and the percentage of URM students within classrooms). Means are presented for each variable for each dataset (i.e., organized by classroom context, year of implementation, and unit). As described in Section 2, selection of HLM models for each dataset was accomplished by considering both the statistical significance of the variable in question and its contribution to the fit of the model. Given context factors and curriculum differences, the same student-level covariates were not predictive in each dataset. But to add clarity to the effects of the intervention across years and units within curriculum type, the same model was used for both years within RCTs, one model for higher-proportion URM classrooms and one for lower-proportion URM classrooms. Thus, there were four models: (a) higher-URM-proportion classrooms in the textbook-based curriculum, (b) lower-URM-proportion classrooms in the textbook-based curriculum, (c) higher-URM-proportion classrooms in the hands-on curriculum, and (d) lower-URM-proportion classrooms in the hands-on curriculum. The addition of nonsignificant variables to models did not impact patterns noticeably; the more important benefit of unique models per dataset is that the strength of covariate effects could vary across datasets. The same teacher-level model was used in all cases, including predictors for condition and the percent of URM in the class. Also note that combining the higher and lower proportion URM groups and both cohorts of students in a larger simple model produces similar overall main effects in both curricula: a small but consistent benefit for the Cognitive Science condition relative to the other conditions across units, that is, statistically significant in the second cohort of students and sometimes statistically significant in the first cohort of students. However, as shown below, the subgroups require different models and show variations in effects.

#### 3.1 | Higher-URM-proportion classrooms in the textbook-based curriculum

At the student level, the following model was compiled, with being female (negative or *ns*<sup>2</sup> predictor), previous achievement (positive

predictor), and being of a URM status (negative or *ns* predictor) predicting outcomes:

$$\text{Score}_{ij} = \beta_{0j} + \beta_{1j}^*(\text{Female}_{ij}) + \beta_{2j}^*(z \text{ Previous achievement}_{ij}) + \beta_{3j}^*(\text{URM}_{ij}) + r_{ij}$$

#### 3.2 | Lower-URM-proportion classrooms in the textbook-based curriculum

At the student level, being female (negative or *ns* predictor), being disadvantaged (negative or *ns* predictor), being an English language learner (positive or *ns* predictor), previous achievement (positive predictor), and being of a URM status (negative or *ns* predictor), were included within the model:

$$\text{Score}_{ij} = \beta_{0j} + \beta_{1j}^*(\text{Female}_{ij}) + \beta_{2j}^*(\text{Disadvantaged}_{ij}) + \beta_{3j}^*(\text{English learner}_{ij}) + \beta_{4j}^*(z \text{ Previous achievement}_{ij}) + \beta_{5j}^*(\text{URM}_{ij}) + r_{ij}$$

#### 3.3 | Higher-URM-proportion classrooms in the hands-on curriculum

At the student level, being female (negative or *ns* predictor), previous achievement (positive predictor), being of a URM status (negative or *ns* predictor), and having a disability (negative or *ns* predictor) were included within the model:

$$\text{Score}_{ij} = \beta_{0j} + \beta_{1j}^*(\text{Female}_{ij}) + \beta_{2j}^*(z \text{ Previous achievement}_{ij}) + \beta_{3j}^*(\text{URM}_{ij}) + \beta_{4j}^*(\text{Disability}_{ij}) + r_{ij}$$

#### 3.4 | Lower-URM-proportion classrooms in the hands-on curriculum

At the student level, being female (negative or *ns* predictor), being economically disadvantaged (negative or *ns* predictor), being an English language learner (negative or *ns* predictor), previous achievement (positive predictor), and being of a URM status (negative predictor) were included within the model:

$$\text{Score}_{ij} = \beta_{0j} + \beta_{1j}^*(\text{Female}_{ij}) + \beta_{2j}^*(\text{Disadvantaged}_{ij}) + \beta_{3j}^*(\text{English learner}_{ij}) + \beta_{4j}^*(z \text{ Previous achievement}_{ij}) + \beta_{5j}^*(\text{URM}_{ij}) + r_{ij}$$

#### 3.5 | The impact of treatment on specific students in specific contexts

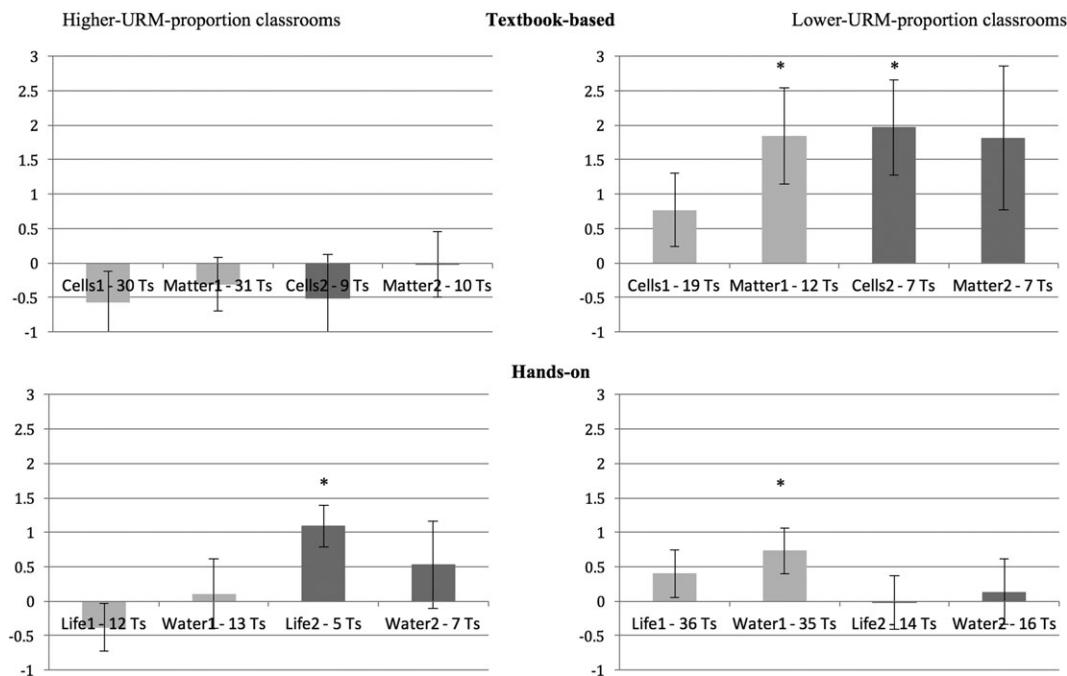
The observed statistical patterns varied considerably across curricula and classroom types, but were relatively similar across units within a curriculum. Table 3 presents all eight models. Figure 7 presents the estimated effect of being in a cognitive science condition classroom relative to being in a control condition classroom. This pattern was robust to variations on the models that did not include covariates. The dependent variable is a gamma coefficient, similar to a beta coefficient, but so named because it is at Level 2 of the model. A gamma of 2 can be interpreted as correctly answering two more

<sup>2</sup>*ns* = nonsignificant variable in some models.

**TABLE 3** Models of student performance across years and curricula

	Higher-URM-proportion classes ICC = .23			Lower-URM-proportion classes ICC = .22			Higher-URM-proportion classes ICC = .19			Lower-URM-proportion classes ICC = .21			
	Coeff.	d	p	Coeff.	d	p	Coeff.	d	p	Coeff.	d	p	
Textbook-based Cells 1st year implemented	Intercept	8.62	-	<.001	11.09	-	<.001	8.28	-	<.001	8.99	-	<.001
	CogSci1	-0.57	-0.17	.211	0.77	0.21	.152	-0.31	-0.09	.430	1.85	0.50	.012
	Content1	-0.60	-0.18	.171	-0.31	-0.08	.590	-0.32	-0.09	.454	0.14	0.04	.818
	Percent URM	-	-	-	-0.03	-	.020	-	-	-	-0.01	-	.530
	Female	-0.14	-0.04	.242	-0.32	-0.09	.017	-0.11	-0.03	.297	-0.22	-0.06	.193
	Disadvantaged	-	-	-	-0.26	-0.07	.081	-	-	-	0.37	0.10	.273
	English learners	-	-	-	0.30	0.08	.035	-	-	-	0.26	0.07	.307
	z previous achievement	1.78	-	<.001	2.34	-	<.001	1.85	-	<.001	2.15	-	<.001
	URM	-0.80	-0.24	<.001	-0.15	-0.04	.278	-0.68	-0.20	<.001	-0.59	-0.16	<.001
	Matter	-	-	-	-	-	-	-	-	-	-	-	-
2nd year implemented	Intercept	8.18	-	<.001	11.29	-	<.001	9.17	-	<.001	10.42	-	<.001
	CogSci2	-0.51	-0.15	.418	1.97	0.52	.010	-0.02	-0.01	.967	1.82	0.48	.097
	Content2	-0.09	-0.03	.891	-1.31	-0.34	.011	-0.19	-0.05	.713	-0.58	-0.15	.568
	Percent URM	-	-	-	-0.03	-	.017	-	-	-	.005	-	.851
	Female	-0.26	-0.07	.031	-0.26	-0.07	.223	-0.43	-0.12	<.001	-0.54	-0.14	<.001
	Disadvantaged	-	-	-	-0.30	-0.08	.128	-	-	-	-0.38	-0.10	.017
	English learners	-	-	-	0.44	0.12	.239	-	-	-	0.44	0.12	.159
	z previous achievement	1.88	-	<.001	2.28	-	<.001	1.95	-	<.001	2.36	-	<.001
	URM	-0.27	-0.15	.273	-0.50	-0.15	.019	-0.87	-0.15	.004	-0.59	-0.15	<.001
	Intercept	-	-	-	-	-	-	-	-	-	-	-	-
Hands-on Life 1st year implemented	Intercept	10.78	-	<.001	11.89	-	<.001	15.78	-	<.001	12.23	-	<.001
	CogSci1	-0.38	-0.13	.288	0.40	0.12	.251	0.11	0.04	.826	0.73	0.22	.030
	Content1	-0.66	-0.22	.067	0.22	0.06	.461	-1.03	-0.36	.046	0.19	0.06	.606
	Percent URM	-0.02	-	.416	-0.04	-	<.001	-0.07	-	.084	-0.05	-	<.001
	Female	-0.06	-0.02	.550	-0.25	-0.07	.006	-0.36	-0.12	<.001	-0.45	-0.14	<.001
	Disadvantaged	-	-	-	-0.19	-0.06	.032	-	-	-	-0.30	-0.09	<.001
	English learners	-	-	-	0.06	0.02	.665	-	-	-	0.06	0.02	.637
	z previous achievement	2.10	-	<.001	2.24	-	<.001	1.49	-	<.001	1.75	-	<.001
	URM	-0.24	-0.08	.195	-0.24	-0.07	.009	-0.42	-0.14	.056	-0.48	-0.15	<.001
	Disability	0.03	-0.01	.897	-	-	-	-0.33	-0.11	.145	-	-	-
Water	-	-	-	-	-	-	-	-	-	-	-	-	
2nd year implemented	Intercept	16.65	-	.001	12.43	-	<.001	16.49	-	.008	12.08	-	<.001
	CogSci2	1.09	0.36	.022	-0.02	-0.01	.965	0.53	0.18	.412	0.14	0.04	.774
	Content2	-	-	-	-0.31	-0.09	.424	-	-	-	-0.04	-0.01	.915
	Percent URM	-0.08	-	.053	-0.04	-	<.001	-0.09	-	.163	-0.04	-	<.001
	Female	0.12	0.04	.460	-0.07	-0.02	.439	-0.13	-0.03	.371	-0.21	-0.07	.025
	Disadvantaged	-	-	-	-0.15	-0.04	.107	-	-	-	-0.30	-0.09	.007
	English learners	-	-	-	-0.41	-0.12	.012	-	-	-	-0.45	-0.14	<.001
	z previous achievement	1.96	-	<.001	2.32	-	<.001	1.52	-	<.001	1.63	-	<.001
	URM	-1.20	-0.40	<.001	-0.24	-0.07	.013	-0.78	-0.27	.011	-0.32	-0.10	.007
	Disability	0.19	0.06	.543	-	-	-	-0.67	-0.23	.016	-	-	-

Note. Model coefficients (in number of correct items), Cohen's *d* (for dichotomous predictors), and *p* values. URM = underrepresented minority, - = predictors not included in model.



**FIGURE 7** The cognitive intervention effect on posttest relative to control for each unit, time point, and subgroup (with SE bars and # of teachers included in analyses). The y axis is the gamma coefficient from the Hierarchical Linear Modeling model contrasting the effects against the control condition, and the units are thus the difference in mean number of items correct on the posttest. URM = underrepresented minority

questions correctly on the 18-item posttest, controlling for all other covariates. Taking into account the standard deviations of the posttest scores within each implementation and classroom context subgroup (see Table 1), the largest condition effects in the textbook curriculum correspond to an effect size of approximately 0.5, and the largest condition effects in the hands-on curriculum corresponds to an effect size of approximately 0.4 (see Table 3 for all effect sizes).

When added to the textbook-based curriculum, the cognitive intervention predicted better scores than control for students in lower-URM-proportion classrooms, particularly when the teacher was implementing for the second time.<sup>3</sup> However, the cognitive intervention did not predict higher scores relative to control in higher-URM-proportion classrooms in either implementation year for either unit of the textbook-based curriculum—effects were either zero or nonsignificantly negative ( $p > .2$ ).

Patterns in the hands-on curriculum were more complex. In the lower-URM-proportion classrooms, the intervention predicted higher scores in the first year but not in the second. This effect was statistically significant only in the Water unit but close to significant and in the same direction for the Life unit. The intervention did predict higher scores in higher-URM-proportion classrooms, but only in the second year. This time, the effect was statistically significant only for the Life unit, but again, the pattern was similar for the other unit.

As shown in Table 3, scores were not positively predicted by the teacher's participation in the active control condition (content training). In fact, in several cases, scores were negatively predicted (at least marginally) by this training. Although this small negative effect

was found in three of the eight models, the most commonly observed effect was a null effect for content training. Overall, this pattern of results suggests (a) that the benefits of the cognitive science intervention could not be explained as a Hawthorne effect and (b) that the second-year cognitive science benefits could not be explained by a selection artifact of teachers who remained for two consecutive years being more willing or better able to take up new approaches.

### 3.6 | Individual-level predictors across models

As can be seen in Table 3, only previous achievement positively predicted scores consistently across our eight models: Higher  $z$ -scored previous achievement predicted higher scores for both groups of students in both types of classrooms. By contrast, many factors differed across models, highlighting the importance of building different models across datasets. One of the most salient differences between our models is the directional difference in predictions when students are English language learners. Being an English learner within a lower-URM-proportion classroom in the textbook-based curriculum predicts higher scores (particularly in the taught once group for Cells but the nonsignificant trend appears in the other three models as well). In contrast, being an English language learner within a lower-URM-proportion classroom in the hands-on curriculum predicts lower scores (particularly in the taught twice group of both units). But there were also a number other differences across contexts in which predictors mattered. For example, whereas being English learners or disadvantaged predicted scores in lower-URM-proportion classrooms, they did not in higher-URM-proportion classrooms across curricula and years.

<sup>3</sup>The marginal gamma coefficient found for the CogSci2 condition in Year 2 of Matter is likely due to low power. It included only 7 teachers and 401 students.

## 4 | DISCUSSION

Overall, the intervention based on four principles taken from cognitive science produced learning benefits for both textbook science in a large urban school district and hands-on science in various urban districts in the southwest. The size of the effects varied, but combination of modifications based on the four cognitive science principles sometimes enhanced, and never hurt, learning. This evidence adds to the knowledge gained from small-scale studies in the cognitive science of education, often focused on a single principle, which have indicated the importance of the principles taken alone and implemented by experimenters. We can now see some evidence of effectiveness in realistic classroom contexts as implemented by practicing teachers over larger units of curriculum than previously studied. That is, there is now evidence of significant cumulative impact of these four principles as a set. However, it is possible that some of the four principles did not add value in this observed effect. Separate studies and/or separate analyses could address the effectiveness of single principles, and we have analyzed this data set more intensively to show the impact of the visualization exercises on student's ability to extract information from Diagrams (Cromley et al., 2016). There could also be evaluation of packages that added fifth or sixth principles, or of entirely different packages of modifications. Such efforts should be the focus of the next wave of research on the use of cognitive science in modifying curriculum in real-world settings.

By contrast, an active control group using short-term enhancement of content knowledge did not show the same pattern compared with business-as-usual, indicating that the effects of the enhanced curriculum were not due to Hawthorne effects (i.e., benefits from any kind of perceived support/change) or to characteristics of teachers who participate in curriculum modifications. Additionally, data from the active control group suggest that providing teachers with a modest amount of increased background knowledge is not an effective strategy for improving student performance. We suspect that the amount of content training was inadequate to develop teachers' content knowledge in a significant way. Further, teachers were likely uncertain as to how to integrate this new knowledge into their classroom practices. Perhaps content training coupled with pedagogical training could have improved their ability to apply the content that they did learn. In any case, the current results do not support districts implementing content-deepening reform efforts if the amount of such content deepening is limited to 20 hr or less. This point does have high policy relevance, as districts attempting to implement their own content deepening for the common case of underprepared middle school science teachers will likely only have resources for 20 or fewer hours of professional development.

Despite the general support for using cognitive science principles, however, there were important policy-relevant differences in the evidentiary support for which kinds of situations are likely to benefit from such curricular interventions. We discuss these differences in the light of the contextual factors mentioned in Section 1.

### 4.1.1 | Immediate versus delayed effects

The effects of the intervention were not generally larger or more likely to be significant in the second year of implementation. Thus, there was some indication that our expectations that teachers might need time to become comfortable with the intervention were not correct. For low-proportion URM classrooms in the textbook curriculum implemented in the large urban school district, benefits were generally found in the first year of implementation, with only modest professional development support. In addition, one of the two significant effects in the data from the hands-on approach in the southwest state came in the first year of implementation. We suspect that the relatively rapid success of the intervention partly rests on its design with teacher resources and materials that made it directly implementable. In surveys, teachers reported implementing the intervention as described to them (which also matched our informal classroom observations) and they reported valuing the well-structured organization of the materials (Desimone & Hill, 2017).

### 4.1.2 | Type of curriculum and learning context

For textbook science as implemented in a large urban district in the Northeast, consistent evidence of benefits appears only in classrooms with lower proportions of underrepresented students (below 80%). For hands-on science as implemented in several Southwestern districts, evidence of benefits appears in classrooms with both higher and lower proportions of underrepresented minorities. Although there is some variation in statistical significance of the condition effect across units within implementation year and learner group for hands-on science (see Figure 7), note that the gamma estimates for the condition effects are generally remarkably stable in size across units.

Evidence for percent URM as a contextual factor rather than an individual factor in the textbook curriculum in the large urban district comes from the fact that percent URM in the classroom had a clear negative effect on student learning outcomes, even within subgroups and even when controlling for URM status at the individual level (i.e., non-URM students performed at lower levels in higher proportion URM classrooms than in lower URM classrooms). Further, URM students in general did not benefit less from the cognitive science condition than did non-URM students—it was the context and not the student level that moderated the effect.

Why did proportion of URM students matter in the large urban school district, although not in the parallel study? One key reason may be that poverty levels were quite high in the large urban district used for the textbook science interventions (see data on free and reduced lunch in Table 1). There may be specific classroom conditions faced by students and teachers in these high URM and high poverty contexts. Desimone and Hill's (2017) report of data from teachers in this study suggests some of the debilitating factors within these higher-proportion classrooms, and we also have some good ideas from prior research. These factors include disruptive interruptions and behavioral issues (Abel & Sewell, 1999; Ingersoll, 2004), resources available to these classrooms and to these students outside of school (Songer, Lee, & Kam, 2002), and teachers being new and/or less prepared (Ingersoll, 2004; Shen, 1997). These are organizational, administrative issues that likely need to be addressed in conjunction with

interventions targeting learning processes in order to see improvements (Cochran-Smith, 2005; Haberman, 1991; Quartz & TEP Research Group, 2003). That is, it is unlikely that any pedagogy-based intervention can provide an impact great enough to overcome these larger problems. The needs of some students might be met with only the principles of cognitive science, but various other problems likely need to be addressed through other kinds of interventions in conjunction with principles of cognitive science.

## 4.2 | Implications for applied cognitive research

In applying basic science to practical situations, there is often a question of translation or operationalization. In this work with curriculum materials, there were many possible ways principles could be implemented in the materials. We used extensive pilot testing to find implementation variations that matched the particular curriculum content and were feasibly implemented by teachers; the Appendix contains examples of the materials to provide worked examples of our approach. Because the exact details varied substantially across unit subsections (e.g., different kinds of visualization exercises and contrasting case organization), it is unlikely that our results depend upon one particular implementation approach.

Although large field trials to evaluate effectiveness are common (Sternberg et al., 2006), there are also concerns about RCTs focused only on general effectiveness questions (Burtless, 2002; Cook & Payne, 2002). From a policy perspective, there is the question of whether the interventions work for a wide range of curricula, students, and contexts. For example, an intervention might work especially well or especially poorly for at-risk students or for curricular materials with high complexity. Analysis of the effectiveness of interventions by subgroup is particularly important because state, district, school, and teacher performances are often calculated separately for various subgroups. There are many schools in the United States, for example, in which performance overall is deemed to be "failing" because of the relatively weak performance of only one subgroup. Individuals making education policy decisions need to make these decisions based on the best available evidence for their specific context: choosing interventions that have low efficacy for their type of curriculum or context is not sensible.

The current results used a new method for analyzing context variation of effects; rather than simply testing main effects and interactions in a single model, separate models were created for key policy-relevant variables. We argue that such a new approach is required for complex HLM models applied to complex real-world data in which (a) learner characteristics are far from evenly distributed across teachers; (b) covariates have differential value across learning contexts; and (c) disruptive factors are highly variable across learning contexts. To use the traditional single-model approach both misrepresents the data and misses important patterns/phenomena of relevance to theory and practice. The main reason for using the single-model approach is to avoid Type I errors. Looking for replications (as done here) across units addresses this concern. Further, all research methods should carefully consider the tradeoffs between Type I and Type II statistical errors; indeed, traditional interaction analyses are

underpowered because high variability in any of the cells will render the effect nonsignificant.

We were unable to conduct more than two large RCTs, as would be required to examine questions such as whether the cognitive science modifications work differently for a textbook curriculum in a rural setting or in a prosperous suburban school district. The sheer scope of such work is daunting, but the present pattern of results shows that efforts in this direction are necessary. The key strategy should be to aim to identify general principles of facilitators and barriers to successful implementation, given this evidence that the principles package can sometimes work, and never seems to hurt. Identifying and defining these factors, and quantifying their impact, would afford us the opportunity to prioritize and to direct our reform efforts more effectively. For example, studies of this size could block random assignment based on the percentage of underrepresented students, match students in higher- and lower-URM classrooms, and collect multiple measures of contextual factors (class-based, school-based, and district-based) to investigate further what types of obstacles teachers/students face and whether those challenges are largely responsible for null effects. Such research would provide administrators with clearly executable suggestions for improvement, and potentially quantify the distribution of responsibility when students are not performing at desired levels.

## ORCID

Christian D. Schunn  <http://orcid.org/0000-0003-3589-297X>

Nora S. Newcombe  <http://orcid.org/0000-0002-7044-6046>

Jennifer G. Cromley  <http://orcid.org/0000-0002-6479-9080>

Christine Massey  <http://orcid.org/0000-0001-8779-9676>

## REFERENCES

- Abdal-Haqq, I. (1996). Making time for teacher professional development. In *ERIC Digest* (#95-4). Washington, D. C.: ERIC Clearinghouse on Teaching and Teacher Education.
- Abel, M. H., & Sewell, J. (1999). Stress and burnout in rural and urban secondary school teachers. *The Journal of Educational Research*, *92*(5), 287-293. <https://doi.org/10.1080/00220679909597608>
- Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, *69*(2), 334-345.
- Alberts, B. (1999). Science and the world's future. Talk presented at the 136th Annual Meeting of the National Academy of Sciences, Washington, DC. Retrieved from: [biochemistry.ucsf.edu/labs/alberts/Editorials/NAS1999.pdf](http://biochemistry.ucsf.edu/labs/alberts/Editorials/NAS1999.pdf).
- Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist*, *48*(2), 87-113. <https://doi.org/10.1080/00461520.2013.775712>
- Bartholomé, T., & Bromme, R. (2009). Coherence formation when learning from text and pictures: What kind of support for whom? *Journal of Educational Psychology*, *101*(2), 282-293. <https://doi.org/10.1037/a0014312>
- Barton, P. E. & Coley, R. J. (2010). The black-white achievement gap: When progress stopped. Policy information report submitted to Educational Testing Service. Retrieved from [www.ets.org/research/pic](http://www.ets.org/research/pic)
- Beede, D., Julian, T., Khan, B., Lehrman, R., McKittrick, G., Langdon, D., & Doms, M. (2011). Education supports racial and ethnic equality in STEM. Issue brief 05-11 submitted to the *Economics and Statistics Administration: U.S. Department of Commerce*. Retrieved from <http://eric.ed.gov/ERICWebPortal/detail?accno=ED523768>

- Berthold, K., & Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. *Journal of Educational Psychology, 101*(1), 70–78. <https://doi.org/10.1037/a0013247>
- Bishop, B. A., & Anderson, C. W. (1990). Student conceptions of natural selection and its role in evolution. *Journal of Research in Science Teaching, 27*(5), 415–427.
- Burtless, G. (2002). Randomized field trials for policy evaluation: Why not in education? In F. Mosteller, & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 179–197). Washington, D.C.: Brookings Institution Press.
- Byars-Winston, A., Estrada, Y., & Howard, C. (2008). Increasing STEM retention for underrepresented students: Factors that matter. Research brief submitted to University of Wisconsin-Madison: The center on education and work. Retrieved from [www.cew.wisc.edu/docs/.../CEW\\_InSTEMRetention\\_UWMadison.pdf](http://www.cew.wisc.edu/docs/.../CEW_InSTEMRetention_UWMadison.pdf)
- Cambone, J. (1995). Time for teachers in school restructuring. *Teachers College Record, 96*(3), 512–543.
- Carey, S. (1991). Knowledge acquisition: Enrichment or conceptual change? In S. Carey, & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 257–291). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Carey, S. (2009). *The origin of concepts*. New York, NY: Oxford University Press.
- Chi, M. T. H. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. Glaser (Ed.), *Cognitive models of science: Minnesota studies in the philosophy of science* (pp. 129–186). Minneapolis, MN: University of Minnesota Press.
- Cochran-Smith, M. (2005). The new teacher education: For better or for worse? *Educational Researcher, 34*(7), 3–17. <https://doi.org/10.3102/0013189X034007003>
- Condrón, D. J., & Roscigno, V. J. (2003). Disparities within: Unequal spending and achievement in an urban school district. *Sociology of Education, 76*(1), 18–36.
- Cook, D. L. (1967). The impact of the Hawthorne effect in experimental designs in educational research. Final report (Office of Education report p-1757). Retrieved from ERIC website: [http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?\\_nfpb=true&\\_ERICExtSearch\\_SearchValue\\_0=ED021308&ERICExtSearch\\_SearchType\\_0=no&accno=ED021308](http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_ERICExtSearch_SearchValue_0=ED021308&ERICExtSearch_SearchType_0=no&accno=ED021308)
- Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational research. In F. Mosteller, & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 150–178). Washington, D.C.: Brookings Institution Press.
- Corcoran, T. C. (1995). *Transforming professional development for teachers: A guide for state policymakers*. Washington, D. C: National Governors' Association. ISBN: 55877-236-7.
- Cromley, J. G., Weisberg, S. M., Dai, T., Newcombe, N. S., Schunn, C. D., Massey, C., & Merlino, F. J. (2016). Improving middle school science learning using diagrammatic reasoning. *Science Education, 100*, 1184–1213.
- Desimone, L. M., & Hill, K. L. (2017). Inside the black box: Examining mediators and moderators of a middle school science intervention. *Educational Evaluation and Policy Analysis, 39*(3), 511–536.
- Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*(2), 103–112. [https://doi.org/10.1207/S15326985EP3602\\_5](https://doi.org/10.1207/S15326985EP3602_5)
- Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915–945.
- Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95*(2), 393–408. <https://doi.org/10.1037/0022-0663.95.2.393>
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–38. Retrieved from [doi:https://doi.org/10.1016/0010-0285\(83\)90002-6](https://doi.org/10.1016/0010-0285(83)90002-6)
- Greene, B. (2008, June 1). Put a little science in your life. *The New York Times*. Retrieved from <http://www.nytimes.com/2008/06/01/opinion/01greene.html?pagewanted=all&r=0>
- Haberman, M. (1991). The pedagogy of poverty versus good teaching. *Phi Delta Kappan, 73*, 290–294. Retrieved from [www.ithaca.edu/compass/pdf/pedagogy.pdf](http://www.ithaca.edu/compass/pdf/pedagogy.pdf)
- Hegarty, M., Kriz, S., & Cate, C. (2003). The roles of mental animations and external animations in understanding mechanical systems. *Cognition & Instruction, 21*(4), 325–360. [https://doi.org/10.1207/s1532690xci2104\\_1](https://doi.org/10.1207/s1532690xci2104_1)
- Henry, G. T., Bastian, K. C., & Smith, A. A. (2012). Scholarships to recruit the “best and brightest” into teaching: Who is recruited, where do they teach, how effective are they, and how long do they stay? *Educational Researcher, 41*, 83–92. <https://doi.org/10.3102/0013189X12437202>
- Holt, Rinehart, & Winston (2007). *Holt science & technology: Cells, heredity, and classification*. Austin, TX: Holt, Rinehart, & Winston. ISBN: 0-03-049958-5.
- Ingersoll, R. M. (2004). *Why do high-poverty schools have difficulty staffing their classrooms with qualified teachers? (Report prepared for renewing our schools, securing our future – A national task force on public education)*. Washington, DC: The Center for American Progress and the Institute for America's Future.
- Jones, S. R. G. (1992). Was there a Hawthorne effect? *American Journal of Sociology, 98*(3), 451–468.
- Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37–62.
- Marincola, E. (2006). Why is public science education important? *Journal of Translational Medicine, 4*(7). <https://doi.org/10.1186/1479-5876-4-7>
- Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59*(1), 14–18.
- Muthén, L. K., & Muthén, B. O. (2011). *Mplus user's guide* (Sixth ed.). Los Angeles, CA: Muthén & Muthén.
- Porter, A., Polikoff, M. S., Barghaus, K. M., & Yang, R. (2013). Constructing aligned assessments using automated test construction. *Educational Researcher, 42*(8), 415–423.
- Provasnik, S., Kastberg, D., Ferraro, D., Lemanski, N., Roey, S., & Jenkins, F. (2012). Highlights from TIMSS 2011: Mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context (NCES 2013-009 Revised). Washington, D.C.: National Center for educational statistics, Institute of Education Sciences, U.S. Department of Education.
- Quartz, K., & TEP Research Group (2003). Too angry to leave: Supporting new teachers' commitment to transform urban schools. *Journal of Teacher Education, 54*(2), 99–111. <https://doi.org/10.1177/0022487102250284>
- Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). *HLM 6 for Windows [Computer software]*. Skokie, IL: Scientific Software International, Inc.
- Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science, 16*(4), 183–186.
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463.
- Shen, J. (1997). Teacher retention and attrition in public schools: Evidence from SASS91. *The Journal of Educational Research, 91*(2), 81–88.
- Songer, N. B., Lee, H., & Kam, R. (2002). Technology-rich inquiry science in urban classrooms: What are the barriers to inquiry pedagogy? *Journal of Research in Science Teaching, 39*(2), 128–150. <https://doi.org/10.1002/tea.10013>

- Spencer, A. M. (2009). School attendance patterns, unmet educational needs, and truancy: A chronological perspective. *Remedial and Special Education, 30*(5), 309–319.
- Sternberg, R. J., Birney, D., Kirlik, A., Stemler, S., Jarvin, L., & Grigorenko, E. L. (2006). From molehill to mountain: The process of scaling up educational interventions (Firsthand experience upscaling the theory of successful intelligence). In M. A. Constan, & R. J. Sternberg (Eds.), *Translating theory and research into educational practice: Developments in content domains, large scale reform, and intellectual capacity* (pp. 205–222). Mahwah, N.J.: Lawrence Erlbaum Associates, Inc.
- Symons, C. W., Cinelli, B., James, T. C., & Groff, P. (1997). Bridging student health risks and academic achievement through comprehensive school health programs. *Journal of School Health, 67*(6), 220–227.
- Tretter, R. R., Jones, M. G., & Minogue, J. (2006). Accuracy of scale conceptions in science: Mental maneuverings across many orders of spatial magnitude. *Journal of Research in Science Teaching, 43*(10), 1061–1085.

- Weiner, L. (2003). Why is classroom management so vexing to urban teachers? *Theory Into Practice, 42*(4), 305–312. [https://doi.org/10.1207/s15430421tip4204\\_7](https://doi.org/10.1207/s15430421tip4204_7)

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

**How to cite this article:** Schunn CD, Newcombe NS, Alfieri L, Cromley JG, Massey C, Merlino JF. Using principles of cognitive science to improve science learning in middle school: What works when and for whom? *Appl Cognit Psychol.* 2018;32:225–240. <https://doi.org/10.1002/acp.3398>