JOURNAL OF RESEARCH IN SCIENCE TEACHING

Seeking the General Explanation: A Test of Inductive Activities for Learning and Transfer

Jonathan T. Shemwell,[1] Catherine C. Chase,[2] and Daniel L. Schwartz[3]

[1]*University of Maine, Orono, Maine*
[2]*Teachers College, Columbia University, New York, New York*
[3]*Stanford University, Stanford, California*

Received 29 April 2013; Accepted 12 October 2014

Abstract: Evaluating the relation between evidence and theory should be a central activity for science learners. Evaluation comprises both hypothetico-deductive analysis, where theory precedes evidence, and inductive synthesis, where theory emerges from evidence. There is mounting evidence that induction is an especially good way to help learners grasp the deep structure (i.e., underlying principles) of phenomena. However, compared to the clear falsification logic of hypothetico-deductive analysis, a major challenge for induction is structuring the process to be systematic and effective. To address this challenge, we draw on Sir Francis Bacon's original treatise on inductive science. In a pair of experiments, college students used a computer simulation to learn about Faraday's law. In the inductive conditions, students sought a general explanation for several cases organized according to Bacon's tenets. In contrast, other students used a more hypothetico-deductive approach of sequentially testing (and revising) their hypotheses using the simulation. The inductive activity led to superior learning of a target principle measured by in-task explanations and posttests of near transfer and mathematical understanding. The results provide two important pieces of information. The first is that inductive activities organized by Bacon's tenets help students find the deep structure of empirical phenomena. The second is that, without an inductive "push," students tend to treat instances separately and fail to search for their underlying commonalities. © 2014 Wiley Periodicals, Inc. J Res Sci Teach

Keywords: learning; scientific reasoning; induction; explanation; deep structure; transfer; Francis Bacon; contrasting cases; physics education; electro-magnetism; college physics; instructional methods

Scientific investigation is a high priority for science education, and this priority reflects broad agreement that students of all ages should learn how to engage scientifically with phenomena (National Research Council, 2000, 2012). It also reflects an appreciation that scientific inquiry is an effective way to learn core concepts (NRC, 2000, 2012). At the center of scientific inquiry is the coordination of theory and evidence, which the Framework for K-12 Science Education (NRC, 2012) refers to as "evaluating" (p. 45). In a pair of studies, we examine how inductive evaluation can support the learning of scientific principles.

Contract grant sponsor: National Science Foundation; Contract grant number: EHR-1020362; Contract grant sponsor: Institute of Education Sciences; Contract grant number: R305A140314.
Correspondence to: Jonathan Shemwell; E-mail: jonathan.shemwell@maine.edu
DOI 10.1002/tea.21185
Published online in Wiley Online Library (wileyonlinelibrary.com).

Inductive evaluation, or synthesis, involves generating hypotheses that fit the data in-hand. It differs from hypothetico-deductive evaluation, or analysis, where students test hypotheses by making predictions and collecting relevant data. Science educators have recognized the distinction between these two forms of evaluation, and they have naturally agreed that both are important and intertwined in science and science education (NRC, 2012). Historically, however, hypothetico-deductive analysis has been more prevalent in science instruction (Chen & Klahr, 1999; Karplus et al., 1977; Lawson, 2010). A common example is the Predict–Observe–Explain (POE) sequence of instruction. Students predict the outcome of a specific experiment; they observe the results; and, if analysis shows discrepancies between their predictions and observations, they need to explain the differences in the service of hypothesis revision. POE was originally designed as a formative assessment so that teachers could observe student thinking (White & Gunstone, 1992), but it is also an approach to instruction in its own right. For instance, if a student prediction is refuted, it can create a teachable moment when the instructor introduces or supports a new way of thinking (Chin & Brewer, 1993; Joshua & Dupin, 1987; McDermott, 1991).

The idealized form of hypothetico-deductive analysis, originating from Galileo, involves a critical test of two competing theories that shows one of them to be false (Medawar, 1979). The head-to-head comparison of two fully-formed theories is rarely obtained. More often, a single hypothesis is tested using various systems of controls to rule out alternative explanations. As a result, hypothetico-deductive practices of science are diverse and complex. They cannot be reduced to formulaic expressions of "the steps of the scientific method" which over-specify the processes and reasoning involved. Nevertheless, the tendency toward such over-specification has been a persistent problem in science education (Bauer, 1992; Grandy & Duschl, 2007; Windschitl et al., 2008).

If hypothetico-deductive analysis sometimes suffers from the problem of over-specification in educational practice, then inductive synthesis seems to suffer from the opposite problem of under-specification. While there is agreement among educators that induction should involve a systematic search for patterns in data (NRC, 2012), there is scant literature about how to foster systematic search. For example, Felder (1993) advocated inductive synthesis to teach Ohm's law by asking students to induce the law from experimental data. Unfortunately, he did not discuss ways of structuring the data or the students' tasks to maximize successful induction. More generally, Duschl (2008) pointed out that the value of finding patterns in data depends crucially upon how the evaluation is conducted, but left open the question of which approaches might be valuable. This lack of specification creates the danger of equating scientific induction with simply looking for patterns in data, messing about, or other forms of minimally guided discovery. Students can easily flounder in these situations (Kirschner, Sweller, & Clark, 2006; Klahr & Nigam, 2004; Mayer, 2004) or stay at the surface of phenomena rather than notice underlying patterns.

On the assumption that there should be more structure to inductive activities than simply looking for patterns, what sorts of structure might be useful? Philosopher and scientist Sir Francis Bacon grappled with this question in the early 17th century and wrote a series of aphorisms for the organization of data and the search for explanations. In the following, we sketch Bacon's proposals and show how they relate to the contemporary empirical literature on inductive learning. We then instantiate his proposal in a set of materials and activities designed to foster induction. Afterward, we describe two studies with university students who investigated electromagnetic flux using a computer simulation. Students assigned to our inductive treatment penetrated to the deep structure of the phenomenon during learning, which in turn enabled them to complete near transfer tasks and understand the relevant mathematical formula more deeply.

Bacon's Philosophy of Induction and Its Relation to Contemporary Learning Theory

In 1620, Bacon introduced his classic treatise on inductive science, *Novum Organum* (Bacon, 1620/2000). He wrote that the preliminary work of the scientist is to assemble data in the form of three tables. The first table should include a variety of phenomena that share a common dimension of interest or "essence." Bacon used the example of heat and tabulated many different things that gave off heat, such as the sun's rays, fresh horse dung, and quicklime sprinkled with water (Book 2, Aphorism XI). The second table should be of negative instances that do not share the dimension of interest but are otherwise similar. To exemplify, Bacon tabulated situations "where the nature of heat is absent but which are in other ways close to ones where it is present" (Book 2, Aphorism XII) such as the fact that the moon's rays, in contrast to those of the sun, do not appear to have a heating effect. The third table should list instances of gradation within the phenomena. Staying with the example of heat, Bacon included astronomers' traditional belief that Mars is the hottest planet, then Jupiter, then Venus, and so on down to Saturn (Book 2, Aphorism XIII). Once the tables of essence, negative instances, and gradation have been duly assembled, then the task of inductive synthesis is to formulate a general explanation. As expressed in Bacon's words, the general explanation must explain all three types of patterns by capturing the essence but also accounting for negative instances and gradation:

> After the presentation has been made, induction itself must get to work. After looking at each and every instance we have to find a nature which is always present when the given nature (in our present case: heat) is present, is always absent when the given nature is absent, always increases or decreases with the given nature, and is a special case of a more general nature (Book 2, Aphorism XV).

Bacon's proposal for a new philosophy of scientific method neatly combines several distinct learning processes that support induction and positive learning outcomes. First, the idea of analyzing instances that share a common essence is strikingly similar to the literature on analogical induction. Studies that focus on analogical induction ask participants, explicitly or implicitly, to find the commonality of multiple instances that are otherwise different at the surface. For instance, a crested rat and hawk moth caterpillar differ on many dimensions, but they do share the common structure of mimicking something more dangerous (a skunk and a poisonous snake). Analogical induction improves people's ability to understand a phenomenon in terms of its deep relational structure. A deep structure is an underlying principle or fundamental relationship that runs throughout a domain. Appreciating the deep structure of different instances is a characteristic of expertise. Physics experts recognize that spring and inclined plane problems both critically involve potential energy, whereas physics novices view them as different types of problems because one involves springs and the other planes (Chi, Feltovich, & Glaser, 1981). More functionally, finding deep structure supports transfer to new situations where surface features differ but the underlying structure is the same (Gentner, 1983; Gick & Holyoak, 1983).

Second, Bacon's table of negative instances maps onto present day researchers' use of contrasting cases to promote learning (Schwartz & Bransford, 1998). Here, learning involves discerning key features that differentiate the cases. For instance, Rittle-Johnson & Star (2007) asked students to notice pairwise differences among algebra solution strategies, which helped the students appreciate the unique strengths of each strategy. The ability to discern diagnostic properties of a situation is a general characteristic of expertise as shown by studies of radiologists, archeologists, and other professions (Biederman & Shiffrar, 1987; Goodwin, 1994). Contrasting cases also support transfer because they help learners recognize precise cues that call for one set of ideas versus another (Bransford, Franks, Vye, & Sherwood, 1989).

Third, Bacon's table of gradations prefigures research on the importance of systematic variation for helping people learn the major underlying dimensions of empirical phenomena (Gibson & Gibson, 1955; Gibson, 1979). For instance, O'Kuma, Maloney, & Hieggelke (2000) created physics exercises in which students ranked cases to help them appreciate the major dimensions of variation. Appreciating systematic variation is important for the meaningful quantification of phenomena, so students can map symbolic manipulations to changes in empirical magnitude.

While the three tables specify the necessary information for successful induction, Bacon recognized that they are not sufficient. People also need to seek a general explanation of the data. Bacon proposed this as a major goal of science—developing generalizable theory. Seeking a general explanation may also be valuable for learning, especially transfer. Transfer has been broadly linked to the process of explaining (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Ryoo & Linn, 2014; Siegler, 1995, 2002). Seeking a general explanation may be especially useful for transfer, compared for example, to explaining a discrepancy in a text or between a hypothesis and a result. For instance, in a laboratory categorization task, seeking a general explanation yielded better learning and transfer than self-explanation and other processing directives (Williams & Lomborozo, 2010). Presumably, this is because seeking a general explanation drives learners to find the deep structure that unifies positive instances and eliminates negative ones. The current research examines whether asking students to seek a general explanation is useful for discovering the deep structure of a scientific phenomenon and for transferring to new problems.

Purpose of the Investigation

The current investigation has two purposes. One is to exemplify how to structure a science learning activity to facilitate Bacon's style of induction. The other is to point out the importance of the inductive orientation to seek the general explanation. The investigation builds on earlier work (Schwartz, Chase, Oppezzo, & Chin, 2011) which had middle-school students invent a rule for describing a set of "Baconian cases" that exhibited systematic variation in the density of objects within containers. Students had to invent a "crowdedness index." The directive of "inventing an index" was a way of conveying the idea of producing a general explanation. The students largely found the deep structure and exhibited strong transfer to a new situation that shared the deep structure of an intensive ratio but differed on the surface (i.e., spring constant). The study did not isolate the value of seeking a general explanation, but it provided a hint as to its importance. In a second condition, students were told the density formula beforehand and used the exact same cases to practice the formula. These latter students simply applied the formula individually to each problem. They missed the deep structure of ratio that unified the instances, and without the ratio concept, there was nothing to transfer to the novel problems. Similarly, other authors have found that when exposed to multiple related instances, people will often treat each instance separately (Gentner, Loewenstein, & Thompson, 2003; Rittle-Johnson & Star, 2007). Thus, students may need explicit directives to seek a general explanation to reap the benefits of inductive activities.

Our primary instructional treatment asked students to synthesize a general explanation for a set of cases organized according to Bacon's tenets. As a comparison point, other students took a more hypothetico-deductive approach to the same cases by following the predict-observe-explain cycle for each case. We chose hypothetico-deductive analysis as the control activity for two reasons. First, in the study by Schwartz and colleagues (2011), the control students completed tell-and-practice activities for learning about density. Hearing a formula and then practicing its application does not naturally call for any sort of analysis or inquiry. Hypothetico-deductive activities, in contrast, are explicitly investigative. This creates an opportunity to see if students will naturally seek a general explanation during inquiry when not explicitly prompted to do so.

Second, we argue for the complementarity of inductive and deductive evaluation. Some learning outcomes may benefit more from inductive synthesis and some may benefit more from hypothetico-deductive analysis. Well-structured hypothetico-deductive analysis should be especially effective for confronting mistaken beliefs by focusing attention (Minda & Ross, 2004), creating dissonance and reflection, and triggering the search for revised understanding (Dega, Kriek, & Mogese, 2013; Khishfe & Abd-El-Khalick, 2002; Nussbaum & Novick, 1982; Strike & Posner, 1992). In contrast, inductive synthesis may be especially useful for initial learning where the goal is for students to learn the deep structure of a new phenomenon about which they have few prior beliefs. Hypothetico-deductive analysis works for scientists because they already have a strong theory generated from many prior instances, and a single case can test their theories. However, a strong theory built up from many instances is exactly what students are missing when learning a scientific theory for the first time. Therefore, inductive synthesis may be a useful approach when initial theory construction is the pedagogical goal. Given that hypothetico-deductive activities are prevalent in schools and educational research, we thought it would be useful to demonstrate the unique benefits of Bacon's well-structured induction while isolating the need to support students to seek a general explanation.

Overview of the Experiments

The study employed three empirical examples that incorporated essence, negative instances, and gradients that students could ideally use to produce a general explanation. Figure 1 shows simplified versions of the examples. They were chosen to help students understand that a changing magnetic field causes current to flow through a coil, as indicated by the changes in brightness of an attached light bulb (Faraday's law). The cases were not designed to help students learn about the

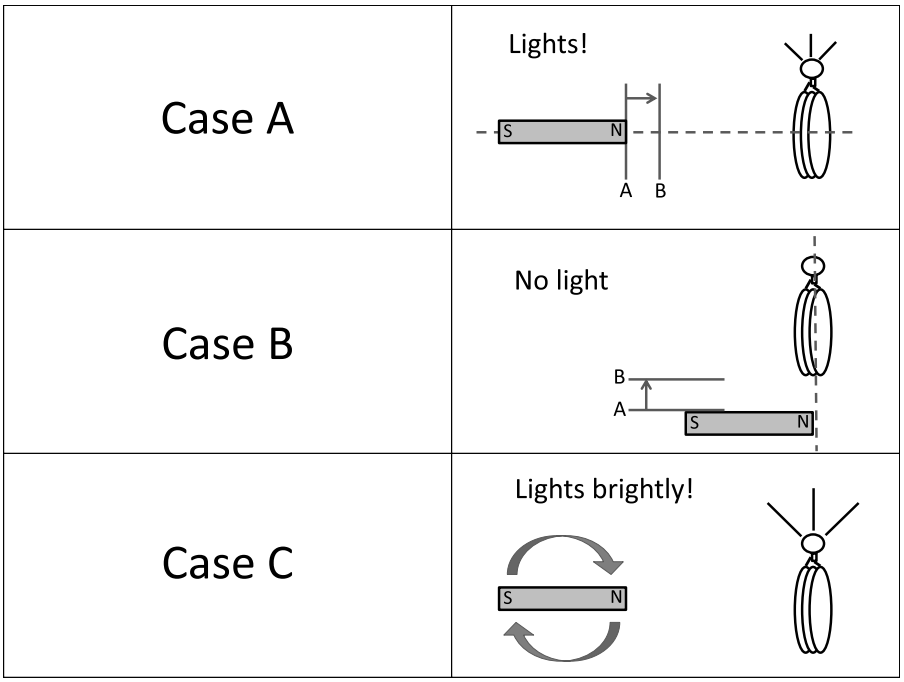


Figure 1. Contrasting cases for Faraday's law.

effect of the rate of field change. Rather, we wanted them to learn that only a change in the magnetic field perpendicular to the face of the coil will induce a current. Cases A and C exhibit a commonality—the light turns on when the field perpendicular to the coil is changed. Cases A and C also exhibit gradation because there is brighter light with a greater amount of change in perpendicular field. Case B provides a negative instance where the bulb does not light because the change in magnetic field is not perpendicular to the face of the coil. No single case provides sufficient information for inducing the principle, but taken together they may.

The main experimental manipulation was the method of evaluation used with the cases. Undergraduate students completed either hypothetico-deductive or inductive evaluation. For both, the students used a computer simulation for about 40 minutes (see Figure. 2; Wieman, Adams, & Perkins, 2008). Using the simulation, students recreated the three cases that are shown in Figure 1. In the inductive condition, students gathered data and tried to induce a single general explanation to account for all the cases. In the hypothetico-deductive condition, they predicted, tested, and explained the results for each case in turn.

In the first study, we confined the investigation to the straightforward question of whether asking students to seek a general explanation with well-organized cases would help them penetrate to the deep structure of the phenomenon and subsequently transfer. Reciprocally, we also wanted to determine whether students in the hypothetico-deductive condition would seek a general explanation without prompting. The second study was a replication of the first, while addressing secondary questions about task instructions, the optimal composition of the cases, and the effects on mathematical understanding.

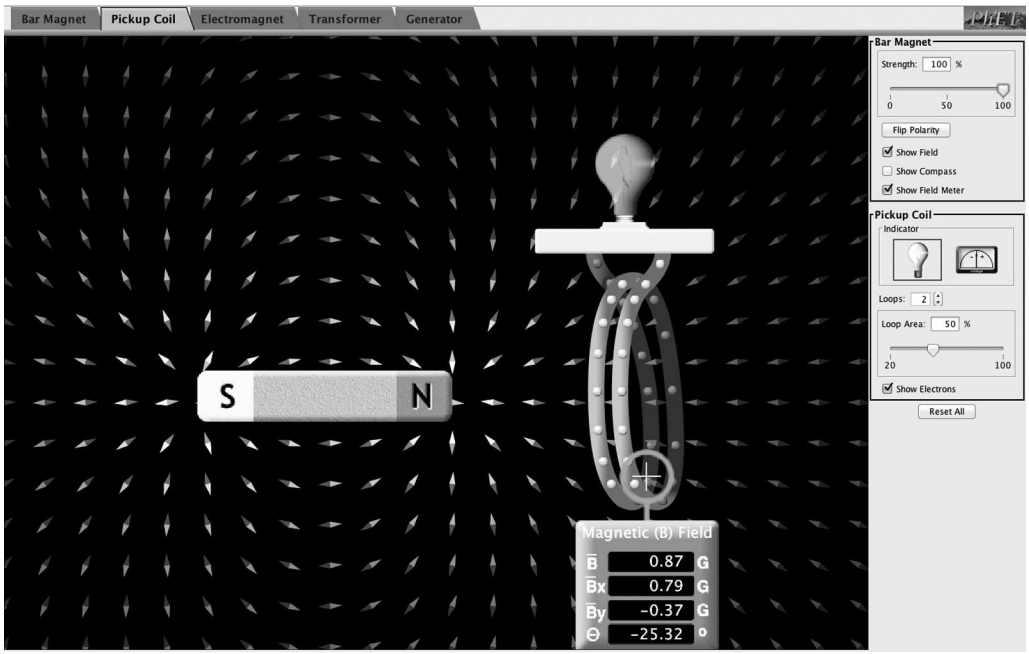


Figure 2. The simulation that students used to experiment with Faraday's law (Wieman, Adams, & Perkins, 2008). The magnet could be moved to change the magnetic field, which was represented by small compass needles. Changes in field direction and strength were indicated by rotation of the needles and changes in their brightness. Students in some conditions were asked to use the field meter, which indicated the strength and direction of the field at any location in numerical form. Image adapted from PhET Interactive Simulations, University of Colorado Boulder under the Creative Commons Attribution license.

We evaluated the effects of the treatments on learning in two ways. The first involved an analysis of students' written explanations during learning. Students had to write explanations of the cases when working with the simulation. We examined these to determine whether students' explanations identified the deep structure of the phenomenon or whether they focused on surface features. Surface features are properties of a problem that are idiosyncratic, compared to deep structures which are relations among properties that are common across different problems (e.g., Gentner & Markman, 1997; Ross, 1987). Compared to experts, novices in a domain often focus on surface features, which may interfere with their abilities to find the deep structure (Chase & Simon, 1973; Chi et al., 1981; Medin & Ross, 1989; Kaminski, Sloutsky, & Heckler, 2008).

Our expectation was that the students who were not told to seek a general explanation would more frequently identify surface features in their explanations. This is because they would be focusing on each case independently and would often rely on features unique to that case to formulate their explanations. For instance, the following explanation is true of case C, but does not handle case A: "when the magnetic field changes direction inside the coil, the light bulb turns on." Thus, we coded whether students gave explanations that relied on surface features.

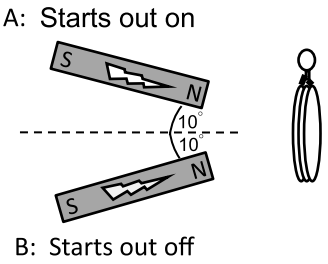
A second expectation was that students who were told to seek a general explanation would more frequently identify the target deep structure. This is because they were trying to find an explanation that worked for all the cases, which depends on the common deep structure. The target deep structure was operationalized as follows: "to induce a current (and light the bulb) there needs to be a change in the x-component (horizontal) of the magnetic field inside the coil." However, it is possible that students could produce general explanations that did not find the deep structure, for example, by being vague. For instance, "a change in the magnetic field in the coil turns on the light," is general to all the cases, but it is imprecise and less scientifically accurate than the statement, "a change in the *x-component* of the magnetic field turns on the light." Therefore, we only coded an explanation as having identified the deep structure when it was precise and accurate, meaning that it had to explicitly refer to the horizontal component of the magnetic field in some way.

The second way we evaluated student learning was by examining their performance on a posttest of novel paper-and-pencil problems that also involved the magnet and coil. Figure 3 provides an example. General explanations, by their nature, should generalize and improve transfer to novel problems, as shown in the literature on analogical learning (Brown & Kane, 1988; Gick & Holyoak, 1983), physics problem solving (Bassok & Holyoak, 1989; Chi et al., 1989), and explanation (Roscoe & Chi, 2007, 2008; Williams & Lombrozo, 2010).

Electric magnets A and B generate the same magnetic field as a regular bar magnet, but they can be switched on and off.

Electric magnet A switches off at the same instant that electric magnet B switches on. The field quickly fades from A and quickly builds up from B, so that the overall amount of field is held constant, but the direction of the field is changed.

Will the bulb light and if so, when?



Why? Your explanation should discuss what happens to the magnetic field inside the coil.

Answer: No, because if the field changes direction, the amount of field perpendicular to the coil stays the same.

Figure 3. An example posttest question. The top magnet phases out just as the bottom magnet phases in, so there is a change in the direction of the magnetic field from angling down to angling up. However, the component of the field perpendicular to the coil remains constant, so the bulb should not light.

All told, we predicted that the inductive students, who were instructed to form general explanations, would more frequently identify the deep structure during learning, and they would better solve transfer problems. In contrast, without the push to synthesize a general explanation, hypothetico-deductive students would not make effective use of the essence, negative instances, and gradients built into the set of cases. Instead, they would be more focused on the separate analysis of each case. As a result, they would be more concerned with surface features of the cases, less likely to identify the deep structure, and less likely to solve problems that require a degree of generalization.

Experiment 1

The experiment took place during the recitation sections of a calculus-based introductory physics course for undergraduate engineering majors. Students completed worksheets which directed them to use the computer simulation shown in Figure 2 to learn about electromagnetism (Wieman et al., 2008). The experiment crossed two factors. The main factor was the pedagogical emphasis: inductive (ID) or hypothetico-deductive (HD), shown in Figure 4.

In the HD condition, students made predictions for all three cases. They then used the simulation to test their predictions, recorded the results, and explained each case in turn. This gave them a chance to revise their explanations with each case. The HD condition was inded to approximate a typical set of HD cycles, except that it presented all three cases together to better parallel the ID condition, which necessarily received the cases presented together. Having all three cases together, if anything, should have tilted the HD students towards a general explanation. In the ID condition, students made no predictions. Instead, they completed an introductory worksheet, Figure S1, which exemplified the process of synthesizing a general explanation using the context of buoyancy. Then, they observed and recorded results for each case in the simulation. Finally, they wrote a single, general explanation for all cases.

The second factor varied in this study, which was ultimately of less importance, was the presence or absence of a measurement tool (MT) within the computer simulation. This tool was

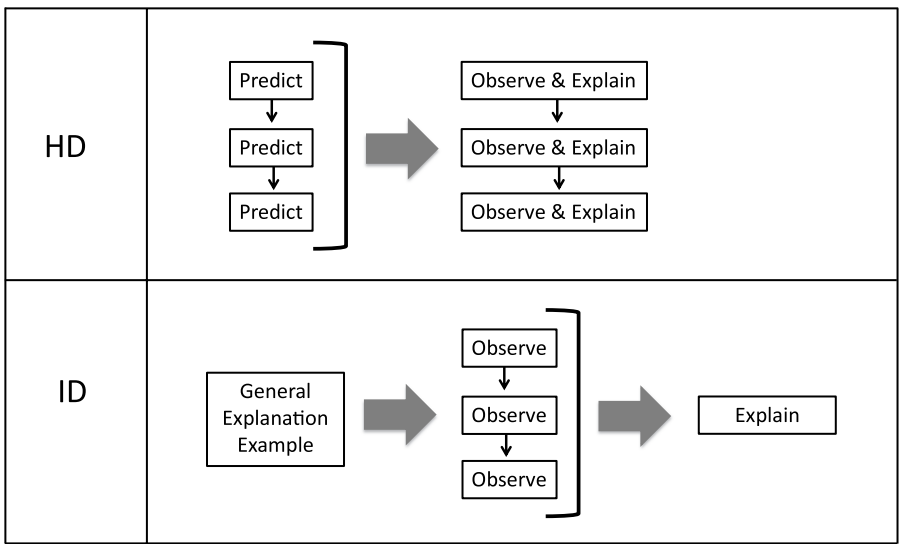


Figure 4. Schematic of the primary experimental manipulation. In the hypothetico-deductive (HD) condition, students predicted each case, then observed and explained each case in turn. In the inductive condition, students read an example general explanation, then observed each case in turn and produced a general explanation for the set of cases.

the magnetic field meter shown in Figure 2. It could be moved to any location to provide a numerical value for the magnetic field vector and its horizontal and vertical components. The MT factor was included to see whether numerical values for field intensity would help students detect field gradations relevant to the deep structure. We predicted that the tool would benefit the ID condition more than the HD condition because the relevant gradations would occur across the cases rather than within each case.

Methods

Participants. Participants were 103 undergraduate engineering students from an introductory physics course in electricity and magnetism at a highly selective, private university. We collected students' midterm examination scores, given one week prior to our study, to ensure that random assignment had not favored one condition over another. The study was timed so that its central topic, Faraday's law, fit into the natural sequence of the course. Just prior to the experiment, students attended a lecture on topics that led up to Faraday's law. One of these topics was magnetic flux, which is a way of quantifying how much magnetic field passes perpendicular to a surface. The study occurred within one of the course's recitation periods. There were 11 recitation sections ranging in size from 6 to 13 students. Sections were led by one of six graduate teaching assistants (TAs) who played a minimal role in the instruction; they simply introduced the activity. A member of the research team was present in each recitation section and provided clarification to the students when needed, typically regarding how to use the measurement tool or position the magnet in the simulation for any given case.

Because the posttest took longer than expected, 23 out of the 103 students left class before they had finished the posttest, stating that they had to make another class or appointment. These students did not complete at least one of the six questions on the posttest. This left complete data for 80 students. We address the threat of attrition in the Results section.

Design and Procedures. The study used a 2×2 between-subjects design that crossed pedagogy (HD vs. ID) with the measurement tool (MT vs. NoMT). Sections were randomly assigned to one of the four conditions. The number of students in each condition varied due to section enrollment: HD-MT (3 sections, 20 students), ID-MT (3 sections, 20 students), HD-NoMT (2 sections, 15 students), and ID-NoMT (3 sections, 25 students).

The instructional period was 40 minutes, with an additional 10 minutes allotted to the posttest. Students used worksheets which directed them to interact with a computer simulation to learn about Faraday's law. The worksheets also contained the instructions that differentiated the four conditions. Students worked in groups of two to three, sharing one laptop computer for the simulation, and filling in their worksheets individually. Working in groups was meant to facilitate interaction that would promote learning and was consistent with the style of instruction in the course recitation sections. At the end of the recitation period, students completed the posttest individually.

In the HD conditions, students processed the three cases shown in Figure 5 sequentially. Students first made predictions for all three cases, then they observed what happened using the simulation, and finally, they explained why. For the predict phase, cases were depicted in a table on a single sheet of paper as in Figure 5, and for each case, students were asked to write down whether the bulb would light and to draw the magnetic field before and after repositioning the magnet. During the observe and explain phases, students used the simulation to test their predictions by recreating each of the three cases. For each case, there were spaces, shown in Figure 5, for students to record "what the light did," describe the light's brightness, and draw the magnetic field they observed for the initial and final magnet positions as depicted in the case. Finally, there was a space

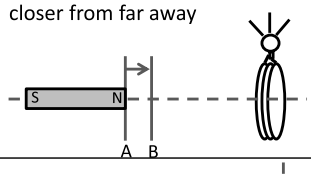
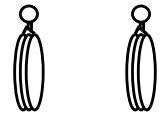
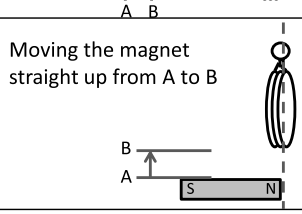
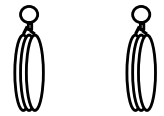
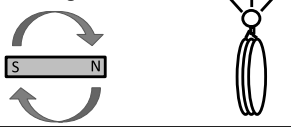
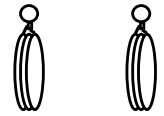
Case	Picture	Record what the light did	Describe the brightness of the light (e.g. dim, bright, very bright)	Carefully record the changes that happened to the magnetic field inside the coil. You may use any representation that you like, such as arrows.	Explain the change in the magnetic field that caused the bulb to light.
A	Moving the magnet slightly closer from far away 			before moving after moving 	
B	Moving the magnet straight up from A to B 			before moving after moving 	
C	Flipping the polarity of the magnet 			before moving after moving 	

Figure 5. Worksheet for the HD condition. Having recorded their predictions on the previous page, students produced each case in the simulation, recorded what happened, and then explained the change in magnetic field that caused the bulb to light. The ID worksheet was identical except that it told students what the light should do in each case, and it omitted the explanation column at the far right, replacing it with a space at the bottom of the sheet on which students wrote a single, general explanation.

at the far right of each case to "explain the change in the magnetic field that caused the bulb to light."

In the ID conditions, students conducted simulations for the same cases as HD using a very similar worksheet. There were four main differences as follows: (1) ID students did not make predictions. They simply observed and recorded what happened for each case, (2) Instead of explaining each case in turn in the right-hand column of the sheet, students were asked to write a single, general explanation for all cases at the bottom of the sheet, (3) To take the place of prediction, and to help ID students understand the idea of a single, general explanation, a cover sheet gave them an example general explanation in the context of buoyancy (see Figure S1). This example presented three contrasting cases in which objects of different volumes and masses either sank or floated, and it provided one model explanation for all of the cases in terms of density, and (4) The ID worksheet told students the results of each case—whether the bulb would light, and how brightly. All told, the ID activity was meant to simulate situations where scientists have the results in hand and are seeking an explanation for them.

Students in the MT conditions used the magnetic field meter to measure and record numerical values for the overall magnetic field intensity at the center of the coil (B) and the x- and y-components of this intensity (Bx and By), for each case. Students in the NoMT conditions were told not to use the measurement tool.

Table 1
Coding definitions and example student explanations on worksheets and tests

	Indicates Deep Structure	Does not Indicate Deep Structure	
		Surface Focus	Non-Surface Focus
Cause of Induced Voltage that Lights the Bulb	A change to the perpendicular component of the magnetic field	A change to an observable feature such as strength or direction of the magnetic field	All other responses (often an under-specified change to the magnetic field)
Example Student Responses	*The induced current is dependent on the speed at which the x-component of the magnetic field changes*	*The greater the change in the direction or slope of magnetic field lines, the greater the brightness*	*As long as the magnetic field is changing, the light lights up*

Measures and Coding. Dependent measures comprised worksheet and posttest explanations. On the worksheet, students were asked to explain the results of the simulated cases (see the right-hand column in Figure. 5). On the posttest, students were asked to make predictions about novel scenarios and explain them.

We scored student worksheet explanations as 1 or 0 depending on whether students explained magnetic induction in terms of the target deep structure (see Table 1). We further coded those explanations that did not reference the deep structure (e.g. non-deep explanations) for dependence on surface features—observable features particular to the case at hand.

The posttest consisted of six items that asked students to make predictions about situations that could not be recreated in the simulation, as in the item shown in Figure 3. These were considered near transfer items because students had to apply the target deep structure in situations to which they had not been exposed during instruction. Like the worksheet explanations, posttest explanations received a score of 1 if the prediction was linked to the deep structure; otherwise they received a score of 0. Non-deep posttest explanations were also scored for surface feature inclusion, which was defined as describing a change to an observable feature in the simulation related to lighting the bulb. Table 1 gives an example. The coding procedure began with two coders independently coding a random sample of 20% of the worksheet and posttest data. Agreement was 90% for worksheets and averaged 88% for posttest items (minimum 84%). Once inter-rater agreement was established, one of the two researchers coded the remainder of the material.

Posttest items were designed expressly for the study and not pilot tested. After coding student responses, we found that two of the items did not discriminate understanding of the deep structure as intended, reducing the internal consistency statistic (Alpha) for the test. We discarded these two items. The remaining four posttest questions, which included Figure 3, were fairly reliable, $\alpha = 0.70$. These items are shown in Figure S2 (Items 1–4).

Results

To check the equivalence of student prior achievement across conditions, we compared students' scores on the midterm exam taken before the study began. An ANOVA crossed the two factors (Pedagogy: HD vs. ID, Measurement Tool: MT vs. No-MT) with midterm scores as the dependent measure. There were no differences in midterm scores according to pedagogy, $F(1,73) = 0.17$, $p = 0.68$. There was a marginal difference by measurement tool, $F(1,73) = 3.30$,

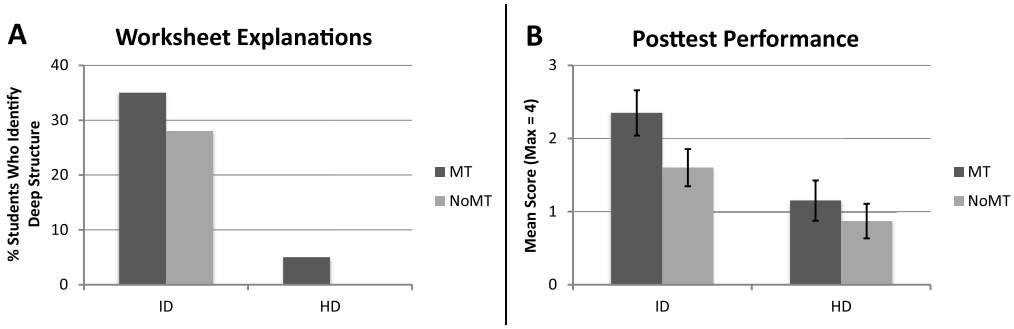


Figure 6. (**A**) Percentage of students who identified the deep structure on their worksheets while working with the simulation. (**B**) Mean scores on a posttest that assessed transfer of the deep structure to novel situations.

$p = 0.07$, however, given that this was not the main comparison in our study, we were not concerned with this near difference. There was no interaction of the two factors, $F(1,73) = 0.02$, $p = 0.90$.

Figure 6a indicates that more ID students identified the deep structure in their worksheet explanations. In the ID condition, 14 of 45 students identified the deep structure compared to 1 of 35 students in the HD condition, χ^2 (1, $N = 80$) = 10.32, $p = 0.001$. The posttest shows a similar pattern. ID students nearly doubled the HD students on the posttest score, as shown in Figure 6b. A factorial ANOVA crossed pedagogy by measurement tool on posttest score. There was a main effect of pedagogy, $F(1,76) = 11.57$, $p = 0.001$, with no main effect or interaction involving the measurement tool, $p > 0.05$. When comparing the HD and ID conditions, the effect size was $d = 0.73$. Thus, ID students were more likely to find the deep structure during learning and transfer their knowledge of the deep structure to novel situations at posttest.

We explored the relationship between finding the deep structure during the worksheet activity and subsequent posttest accuracy. Because only one student in the HD condition wrote a deep explanation on the worksheet, the analysis considered only the ID condition. ID students who explained the deep structure scored roughly twice as high on the posttest as ID students who did not, $M_{Deep} = 2.93$ (SE = 0.37), $M_{NonDeep} = 1.48$ (SE = 0.22), F (1,43) = 12.61, $p = 0.001$. Our preferred interpretation is that finding the deep structure during the learning activity led students to recognize the deep structure of the novel materials presented at posttest. An alternative interpretation is that the students who knew enough to figure out the deep structure were also the ones who could solve the new problems. However, the mid-semester exam scores were the same for those who did and did not find the deep structure on the worksheet, $t(42) = 0.02$, $p > 0.05$, which makes it unlikely that pre-existing knowledge or ability can explain the linkage between worksheet and posttest scores.

We had predicted that HD analysis would incline students to handle each case independently. One way to detect this behavior is to examine whether their worksheet explanations hinged on surface information that was specific to a single case rather than all cases. To find out, we confined the analysis to those who did not identify the deep structure. Sixty-seven percent of HD students wrote surface-level explanations. In contrast, only 29% of ID students wrote surface-level explanations, χ^2 (1, $N = 65$) = 9.67, $p = 0.002$. Most of the non-deep, non-surface explanations of ID students were vague generalities such as "any change in magnetic field" lights the bulb rather than specifics to any single case. An attenuated version of this worksheet pattern occurred on posttest questions. Looking at the non-deep explanations on all items of the posttest, the HD group wrote a

descriptively higher percentage of surface-level explanations than the ID group, HD = 54.6% (SE = 6.2%), ID = 48.2% (SE = 6.5%). However, this difference was not statistically significant, $t(1, 68) = 0.72, p > 0.05$. The prevalence of surface-level explanations for the HD condition supports the contention that, without explicit encouragement to treat the cases jointly, students tended to treat each case separately and inadvertently focused on their unique properties.

The effects of the measurement tool, while much more moderate than those of pedagogy and not statistically significant, were in the expected direction. There was no main effect of measurement tool on worksheet explanations. Twenty percent of students in the MT condition identified the deep structure in their explanations compared to 18% in the NoMT condition. The measurable difference in worksheet explanations occurred within the ID condition, where a slightly higher percentage of students indicated the deep structure when using the measurement tool, 36% MT vs. 29% NoMT. This difference was not statistically significant, χ^2 (1, $N = 80) = 0.08$, $p = 0.78$. The advantage for MT on the posttest was slightly more pronounced, $M_{MT} = 0.44$, $M_{NoMT} = 0.33$, $F(1,76) = 3.30$, $p = 0.07$, but still not significant. More importantly, there was no interaction between pedagogy and measurement tool, F $(1,76) = 0.67$, $p = 0.41$. Thus, the prediction that the measurement tool would benefit the ID condition more than the HD condition did not turn out.

While all students completed worksheets, 23 students did not complete one or more of the posttest items because they had to leave the session before it ended. These students' worksheet and (partial) posttest results were not included in the preceding analyses. The rate of attrition differed by condition: 15% of HD students and 27% of ID students left early. This raises the question of whether the ID condition performed better simply because less knowledgeable students left early. Evidence indicates this was not the case. The 23 students who departed during the posttest did not have different worksheet outcomes than those who completed the test. Of the HD students who left before completing the posttest, none of them identified the deep structure on the worksheet, whereas 29% of the ID students who left did, which is nearly identical to the 31% of ID students who remained. With respect to the posttest, 20 of the 23 students who left early completed the first two questions on the test. Among those 20, the results were in the same direction as the full sample, with students in the ID condition outperforming students in the HD condition, $F(1,18) = 19.02$, $p < 0.001$.

As a final check, we examined differences in prior midterm scores between students with complete and incomplete data. The scores did not differ, $M_{Incomplete} = 26.8$ (SE = 1.5), $M_{Complete} = 27.0$, (SE = 1.0), $t(1,97) = 0.09$, $p = 0.93$. Given these multiple checks, attrition is an unlikely explanation for the treatment effects.

Discussion

To learn about Faraday's law, students received three cases and collected simulated data for each. One condition was told to find a general explanation for all the results (ID condition), whereas the other condition was told to predict-observe-explain for each case in turn (HD condition). Only one HD student spontaneously identified the deep structure. Instead, most HD students focused their explanations on surface features unique to each specific case. In contrast, ID students generated explanations that identified the deep structure more frequently. These differences in how students engaged with the three cases had strong implications for their learning. On a posttest of novel problems involving the same apparatus (a moving magnet, a coil, and a light bulb), ID students were again more likely to identify the deep structure. Moreover, identifying the deep structure during the learning activity was significantly associated with how well students performed on the posttest.

Our interpretation of the current results is that inductive synthesis, with its goal of producing a general explanation from carefully arranged cases, helped students to find the deep structure that

represents the unifying principle. However, it is also important to note that many students in the ID condition never found the deep structure (they simply wrote vague explanations), a concern we address in the General Discussion. Nevertheless, given that the ID condition outperformed the HD condition so dramatically, the current results show the value of structured induction activities.

There are alternative interpretations of the results. One is that the HD worksheet did not show the outcome of the light bulb as shown on the ID worksheet. For the HD condition, showing the outcome would have ruined the prediction task. This difference may have been important. The HD students may not have produced the cases accurately enough, so that the simulation yielded poor data. Making the simulated magnet behave as shown in the worksheet was prone to manipulation errors (e.g., sliding it a little bit on the x-axis, when it is supposed to only move on the y-axis). The next experiment addresses this concern. After making each prediction and conducting its associated experiment, HD students received the same outcome information as ID students before moving on to give their explanations.

A second concern is that the ID students received an initial activity that was intended to help them understand the nature of a general explanation (Figure S1). The example demonstrated that a general explanation accounts for all cases, and it also showed a model explanation. The HD students did not receive this instruction, which means they may not have understood the components of a high-quality explanation, or they or may have placed little emphasis on the importance of their explanations. Therefore, in the next study, HD students received a preliminary worksheet similar to the ID worksheet where they were shown a model explanation using the same example of buoyancy.

The effects of the measurement tool were inconclusive. Descriptively, the tool helped the ID condition more than the HD condition, but the effect was insignificant and far less than the effect of pedagogy. Perhaps it helped a few students detect the gradient of x-component changes more precisely. Regardless, given that measurement is a regular feature of scientific investigations, we decided to provide the tool for all students in the next study.

Experiment 2

Experiment 2 served as a replication of Experiment 1 with adjustments to remove potential confounds, as described below. The experiment also incorporated new conditions and new measures. It took place one year later at the same course at the same institution.

One new condition, low-contrast induction (LCID), evaluated Bacon's proposal that negative instances should have an outcome that "*is always absent when the given nature is absent"* (Bacon, 2000, Book 2, Aphorism XV). Bacon recognized that negative instances are crucial for induction because they help to rule out plausible false explanations which are otherwise consistent with the data. The negative case (case B, Figure. 1) in the ID condition implements Bacon's proposal, because the magnetic field changes, but light does not come on. It was intended to help students rule out surface-level explanations consistent with cases A and C, such as the explanation that a change in the strength or direction of the magnetic field would induce a voltage. In the LCID condition, we replaced the "absent," high-contrast case with a low-contrast one where the bulb lights dimly because the magnetic field is in between vertical and horizontal (see Figure. 7). Otherwise, the LCID condition had the same instructions as the ID condition. We predicted that students in the LCID condition would do worse than the ID condition, because the low contrast would make it harder to notice the significance of the x-component of the field.

The other new condition, compare and contrast (CC), evaluated the importance of telling students that they should seek a general explanation. Students in this condition received the same cases as the ID condition and addressed the cases as a group. However, they did not have an explicit

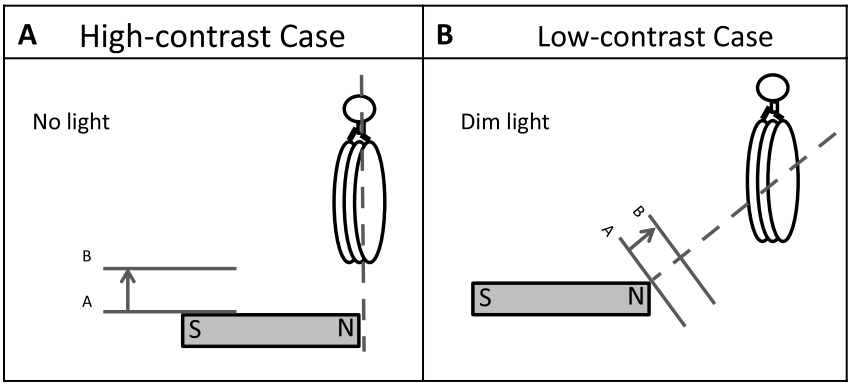


Figure 7. (**A**) ID, HD, and CC conditions received the negative (i.e., high-contrast) case in which the magnetic field changed in the vertical direction without lighting the bulb. (**B**) In the LCID condition, the negative case was replaced by a low-contrast case in which the magnetic field changed at an angle to the coil, producing dim light.

prompt to give a general explanation. Instead, their task was to explain the similarities and differences between the cases. This condition also mimics instruction from studies of analogical induction, where learners are typically asked to compare and contrast a set of analogous examples (Catrambone & Holyoak, 1989). We predicted that these students would do more poorly than the ID students, because the "compare and contrast" directive would be insufficient inducement to seek a general explanation.

A final change from Experiment 1 was that the posttest was modified to include measures of quantitative understanding, while still measuring qualitative understanding. Since the study context was a university physics course that largely focused on mathematical explanations, we wanted to determine whether inductive activities would have a negative, neutral, or positive effect on mathematical understanding of the physics.

Methods

Participants. Participants were 316 students from the following year of the same course as Experiment 1. As before, the study was conducted in students' recitation sections with minimal TA involvement. The study was again timed so that the topic of Faraday's law fit into the natural sequence of the course. There were 26 sections taught by 14 different TAs. Each TA taught two sections individually, except that two of the sections were co-taught by a pair of TAs. Each section was randomly assigned to condition under the constraint that no TA taught 2 sections from the same condition.

Design and Procedures. The study design had four between-subjects conditions with a planned comparison of the ID condition against each of the other three conditions. Table 2 summarizes the condition differences.

There were two key changes to procedures in Experiment 2 that were designed to address the two potential confounds discussed for Experiment 1. The first was that students in all conditions were exposed to a model explanation involving buoyancy, not just the ID condition. The model explanation was identical for all students, but in accordance with condition differences, there were different task directives for processing it. For the inductive conditions, ID and LCID, the directive

Table 2
Summary of condition differences

Condition	Introductory Example of a Model Explanation	Worksheet Contrasts	Task Prompt
ID	Showing several cases	High	General explanation
LCID	Showing several cases	Low	General explanation
CC	Showing a single case	High	Similarities & differences
HD	Showing a single case	High	Predict, observe, explain

was the same as for Experiment 1. It asked students to find a general explanation for a set of buoyancy cases; then it gave students the model explanation that buoyancy depends on mass per unit volume (see Figure S1). The two non-inductive conditions, CC and HD, were shown a single buoyancy case and the model explanation of buoyancy. Their task was to explain why this explanation was better than an explanation based on a single feature of the case, mass. Then they were told that the model explanation was better because it was more precise (see Figure S3). Thus, while all conditions were shown the model explanation, only the inductive conditions were encouraged to seek this explanation using a set of cases.

A second departure from Experiment 1 is that the HD condition received the results of each case to address the possibility that the HD condition's lower performance could have been caused by inaccurate data collected in the simulation. After completing their observations of each case, students in the HD condition were directed to a website containing the same outcome data that the ID, LCID, and CC students saw.

Measures and Coding. As in Experiment 1, the measures were the worksheet explanations and posttests. The posttest in Experiment 2 contained four items. Two were qualitative prediction tasks from Experiment 1: rotating magnet prediction and switching magnet prediction (see Figure S2, Items 1 and 2). Two new items assessed quantitative understanding. One was a physics word problem that measured application of a formula included in Faraday's law (see Figure S2, Item 5). It asked students to calculate a magnetic flux, which by definition requires calculating the dot product of the magnetic field vector, B, and the vector describing the surface that the field passes through, dS. This "formula application" problem was included to see if the inductive activities put students at risk for computation tasks. Our prediction was that they would not; understanding the physical phenomena should not interfere with standard word problems.

The other quantitative item measured insight into the mathematical structure of the same formula as the application problem (see Figure S2, Item 6). In this less prototypical item, students had to explain, "Why does Faraday's law take the dot product of B and dS?" The answer is that Faraday's law uses the dot product to quantify the extent to which the magnetic field vector, B, is perpendicular to the surface it passes through. We included this "formula insight" item to see whether inductive activities helped students understand why a formula takes the structure it does (Schwartz & Martin, 2004). We predicted that most students would be able to manipulate the formula to achieve the correct answer, but the ID students would better understand the formula's conceptual underpinnings.

The worksheet explanations and two qualitative prediction items were coded per Experiment 1. The formula application item was scored as 1 or 0 depending on whether students used the dot product to calculate the magnetic flux. The formula insight item was also scored as 1 or 0. A score of 1 was given if a student's explanation of the formula invoked the deep structure (i.e., the dot product is used because it takes the component of the magnetic field that is perpendicular to the

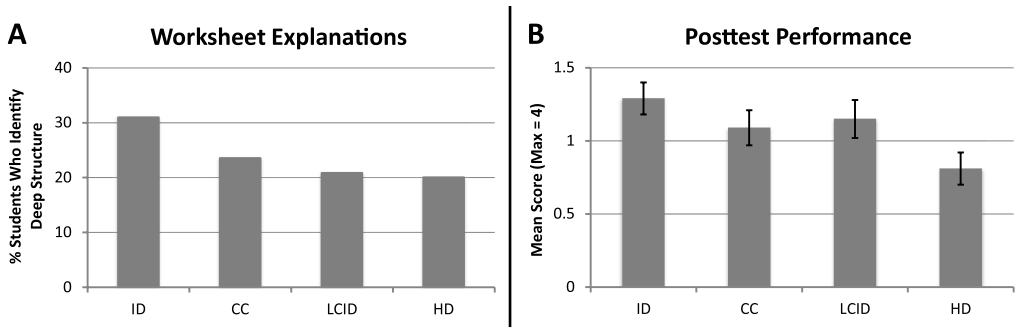


Figure 8. (**A**) Percentage of students who identified the deep structure on their worksheets while working with the simulation. (**B**) Mean scores on a posttest that assessed transfer of the deep structure in both quantitative and qualitative situations.

surface. As in Experiment 1, acceptable coder agreement (91% for worksheets and 84–100% for posttest questions) was established on a random 20% sample of the data. The remaining data were coded by one of two researchers.

Results

Prior achievement was equivalent across conditions based on the students' preceding midterm exam scores, $F(3,307) = 0.67$, $p = 0.57$; $M_{ID} = 60.9$ (SE = 1.8), $M_{HD} = 61.6$ (SE = 1.9), $M_{CC} = 58.2$ (SE = 1.8), $M_{LCID} = 59.8$ (SE = 1.9).

In the remaining analyses, we begin with an omnibus test of condition effects. We then apply planned orthogonal contrasts that compare the ID condition to each of the other three conditions.

As Figure 8a indicates, more student explanations identified the deep structure on the worksheet in the ID condition compared to the other conditions. However, this difference was not significant, χ^2 (3, $N = 316$) = 3.25, $p = 0.35$. The planned comparison of ID to HD was not statistically significant, χ^2 (1, $N = 159$) = 2.48, $p = 0.12$, nor did ID differ statistically from CC or LCID. While the ID students in this experiment performed similarly to the ID students in Experiment 1, the HD students performed much better in Experiment 2. Providing HD students with a model explanation and/or outcome data may have caused their improvement.

The ID students scored higher than the other conditions on the transfer posttest. As shown in Figure 8b, the ID students had the highest scores, followed by the LCID and CC students, and then the HD students, who had the lowest scores. A one-way ANOVA found a significant effect of condition on posttest performance, $F(3,312) = 2.71$, $p = 0.045$. Planned comparisons revealed that the ID group's posttest score was significantly greater than the HD group's, $p = 0.006$, effect size $d = 0.47$. These findings replicate the relative advantage of ID over HD found in Experiment 1. There were no significant differences between ID and LCID, $p = 0.19$, or between ID and CC, $p = 0.28$. To check if the LCID and CC conditions significantly outperformed the HD condition, we used a Tukey's *post hoc* comparison. This non-conservative test maximizes the chances of detecting any differences. The contrast of HD with LCID was not significant, $p = 0.22$, nor was the contrast of HD with CC, $p = 0.35$. By inference, the posttest scores of the LCID and CC conditions fell between the ID and HD conditions, as they were not significantly different from either. In sum, the ID students were best at generalizing to novel situations, while HD students fared the worst.

Given the umbrella protection of the significant ANOVA, we disaggregated the posttest to explore whether the results would hold up across the items. Individual Chi-square tests compared

Table 3
Percentage of students with correct answers on each posttest item, organized by condition

Posttest Item[a]	ID	HD	CC	LCID
Prediction (Item 1)	27.4	14.7*	22.4	23.6
Prediction (Item 2)	19.1	8.0*	17.7	16.7
Formula Insight (Item 5)	47.6	32.0*	38.8	34.7
Formula Application (Item 6)	34.5	26.7	30.6	40.3

[a]aAll items can be viewed in Figure S3.
**p* ≤ 0.05, comparison against ID condition.

the ID condition to each of the other three conditions for each posttest question. Table 3 shows the results of these tests along with the percent of students who correctly answered each question in each condition. The statistical advantage of ID over HD holds for the two prediction questions and the formula insight question. But, as anticipated, there was no difference between ID and HD conditions on the formula application item, which assessed procedural rather than conceptual understanding. Again, the LCID and CC conditions tended to fall between the ID and HD conditions.

As in Experiment 1, we examined the relation between finding the deep structure on the worksheet and performance on the posttest. Unlike Experiment 1, we were able to include all four conditions in this analysis because sufficient numbers of students identified the deep structure in each condition. We separated students who identified the deep structure in their worksheet explanations from those who did not. We then compared their posttest scores crossed by condition. Overall, students who identified the deep structure scored twice as well as those who did not on the combined posttest prediction and insight problems, $M_{Deep} = 1.75$, (SE = 0.12), $M_{NotDeep} = 0.88$, (SE = 0.06), $F(1,308) = 42.0$, $p < 0.001$, $d = 0.74$ (see Table 4). There was no interaction with condition, indicating that an active ingredient for transfer is finding the deep structure, regardless of condition.

Prior achievement did not explain the relationship between identifying the deep structure in the worksheet explanations and posttest performance. An ANCOVA on posttest performance crossed the factor of deep and non-deep worksheet explanations with course midterm scores as a covariate. When controlling for prior achievement, the effect of identifying the deep structure on posttest performance remained unchanged $F(1,302) = 39.9, p < 0.001, d = 0.74$.

HD students' explanations again tended to focus on the surface features of the cases. Confining our analysis to those students who wrote non-deep explanations on the worksheet, there was a significant difference in the percent of students who wrote surface-level explanations by condition, χ^2 (3, $N = 240$) = 37.98, $p < 0.001$. This was driven by the HD condition, where many students wrote surface-level explanations compared to other conditions: HD = 51.7%, ID = 13.8%, LCID = 7.0%, and CC = 21.5%. Planned comparisons against the ID condition revealed that more HD students wrote surface-level explanations than ID students, χ^2 (1, $N = 118$) = 19.12, $p < 0.001$, but the comparison of ID to CC and LCID did not yield significant differences. Of those students who never found the deep structure across conditions, the HD students were most likely to focus on features unique to a given problem.

A similar pattern occurred for the posttest problems. Again, we confined the analysis to the non-deep explanations on the posttest. A one-way ANOVA found marginal condition differences in the number of surface-level explanations across posttest problems, $F(3,290) = 2.46$, $p = 0.06$. The percentages of surface-level explanations out of all non-deep explanations for the three

Table 4
Mean posttest scores[a] (standard errors) of students who wrote deep worksheet explanations versus those who did not, organized by condition

Explanations:	ID	HD	LCID	CC
Deep Structure	1.96 (.20)	1.13 (.26)	2.13 (.26)	1.75 (.22)
Non-deep	0.98 (.13)	0.73 (.13)	0.90 (.13)	0.89 (.12)

[a]Means exclude the formula application problem, which was not designed to index deep understanding.

explanation problems were: ID = 44.7% (SE = 4.2%), HD = 57.2% (SE = 4.3%), LCID = 43.4% (SE = 4.7%), and CC = 53.6% (SE = 3.9%). Planned comparisons of data from students who did not identify the deep structure revealed that the ID condition had fewer surface-level explanations than the HD condition, $p = 0.04$, but the comparisons of ID to LCID and CC were not significant. HD students, who were not explicitly encouraged to consider the cases together, tended to analyze each case separately, honing in on the surface features specific to each case.

Discussion

Experiment 2 largely replicated Experiment 1 while extending the findings. First, students engaged in inductive synthesis (ID) were better at finding the deep structure among the three cases on the worksheets than students who engaged in hypothetico-deductive analysis (HD), though this difference was smaller than in Experiment 1 and not statistically significant. We speculate that the HD condition's improved worksheet performance in Experiment 2 was caused by the addition of the model explanation in the introductory materials and/or the newly provided outcome data.

A more critical replication from Experiment 1 is that students in the inductive synthesis condition (ID) were again better at transferring their learning experiences to solve novel problems compared to those in the hypothetico-deductive analysis condition (HD). The data closely replicated Experiment 1. In Experiment 1, the HD condition performed at 53% of the level of the ID students on the posttest, and in the Experiment 2, they were at 49%. Also consistent with Experiment 1 is the finding that HD students were more likely to base their explanations on surface features of a single problem for both worksheet and posttest problems. We conclude from these results that Bacon's inductive synthesis leads students to uncover the generalizable deep structure that supports transfer to new problems that share the same deep structure. In contrast, hypothetico-deductive analysis may be better used to focus students on the relation of a single instance to a hypothesis.

On the worksheet, the HD and ID students did not statistically differ in the frequency of explanations that identified the deep structure. Yet on the posttest, the HD students still did substantially worse. While statistical power is implicated in the lack of effect on the worksheets, we can still ask, why did HD students who found the deep structure on the worksheet still fail to handle the novel problems on the posttest? One possibility is that finding the deep structure is necessary, but not sufficient, for transfer. Perhaps the students in the HD condition never realized that one value of an explanation based on deep structure is that it generalizes. For example, HD students often found the deep structure for one of the worksheet cases, but did not use it for all the cases on the worksheet. Similarly, they may not have thought to use the deep structure to explain the new problems on the posttest. It is potentially informative that, among students who found the deep structure on the worksheets, those in the two conditions encouraged to generalize (ID and LCID) were more successful on the posttest than those who were not encouraged to generalize (HD and CC). While the difference is not statistically significant, it is possible that the push to

generalize from multiple situations helped students understand that a good explanation is supposed to generalize to new situations they might encounter.

Experiment 2 added measures of quantitative understanding. One measure examined whether students could apply a relevant formula to a situation. The ID and HD conditions performed similarly. The second measure examined whether students could explain why the formula has the structure that it does. For this formula insight problem, the ID condition outperformed the HD condition. This provides two pieces of information. The first is the familiar observation that knowing how to use an equation is not the same thing as understanding the phenomenon it describes. The second is that inductive synthesis does not displace students' learning how to use a mathematical procedure, but rather, this activity can help students understand the structure of a quantitative expression more deeply.

The low-contrast inductive condition (LCID) and compare and contrast condition (CC) were included to explore possible variations to inductive activities. The LCID condition replaced the negative instance of a non-lighting bulb with the low-contrast case of the bulb sliding at an angle and producing less light. The CC condition replaced the instructions to seek a general explanation with instructions to note the similarities and differences of the cases. Each is a degradation of Bacon's proposal for optimal conditions for induction. Descriptively, LCID and CC led to similar results as one another, and fell below the ID condition but above the HD condition. However, statistical tests indicated that they were not significantly different from either the ID or HD conditions. This makes it difficult to isolate their respective effects with confidence. The most prudent conclusion is that LCID and CC are not optimal conditions for helping students learn a general explanation, but why they are not optimal requires further study.

General Discussion

College students in a calculus-based physics course for non-majors analyzed the outcomes of three simulated experiments related to Faraday's law. In an inductive synthesis condition (ID), students collected all the data and then tried to generate a single, general explanation to handle the experimental results. In a hypothetico-deductive analysis condition (HD), students predicted the results of each experiment in turn and generated an explanation for each result. In both experiments, the ID condition showed strong advantages for learning. ID students were more likely to find the deep structure of magnetic field changes that cause a current to flow in a neighboring coil (i.e., changes to the magnetic field vector perpendicular to the coil), particularly in Experiment 1. This success during learning had a positive effect on near transfer problems at posttest. The second study also found that the ID students developed a better grasp of the mathematical formula. They were better able to explain one of its key components, the purpose of the dot product, even though the dot product was not broached in the lesson. In contrast, the HD students who completed a more traditional sequence of predicting, testing and explaining tended to focus on features unique to an instance to explain a given result. As a result, they did not identify the deep structure that spans multiple cases, and therefore, could not handle new problems that varied on the surface.

Without encouragement to find a general explanation, otherwise well-schooled HD students did not naturally look for one. The second experiment included a pair of conditions to determine why this was the case. One possibility was that the sequential presentation of the cases in the HD condition drove students down the path of treating each case separately. If true, then students in the compare and contrast condition (CC) should have done better than the HD students, because they were not led to view each case on its own. A different possibility is that students simply do not think to find a general explanation without explicit prompting, despite the fact that generalizing is central to most real-world scientific inquiry. If true, then the low-contrast inductive students

(LCID) should have done better than the HD students, because they received the relevant prompt. The results showed that the CC and LCID conditions did better than the HD condition but worse than the ID condition, and they were not significantly different from either. This middle-ground result suggests that both possibilities are in play—many students do not think to find the general explanation, and the HD inquiry pulls students to focus on instances. The latter point is highlighted by the disproportionate use of surface features in the HD student explanations compared to all others.

Our primary proposal is that the psychological cause of the difference between the ID and HD conditions involved what they were searching for—a general explanation versus an explanation of a specific result. An alternative possibility is that the conditions exerted unequal cognitive load (Paas, Renkl, & Sweller, 2004; Sweller, 1988). On the one hand, trying to find an explanation that handles three cases simultaneously seems more taxing of working memory than handling one case at a time. On the other hand, trying to remember the results of an earlier experiment and integrating it with a current experiment may be even more demanding (Bruner, Goodnow, & Austin, 1967). It is informative to note that the HD students focused on surface features, which directly implicates a poor search strategy rather than a working memory burden. Nevertheless, in-task self-reports of cognitive load could help clarify whether working memory demands are causing the treatment differences.

Notwithstanding the positive effects of inductive synthesis on learning, there are two limitations of this investigation that we wish to highlight. First, our studies have made headway on, but have not conclusively addressed, a pair of important questions about the design of inductive activities. One question asks how to structure data for synthesis. In the second study, we predicted that the inclusion of the negative case would be especially important for successful generalization. The lower scores for LCID, which omitted the negative case, accorded with our prediction, but mean differences did not run to statistical reliability. Future studies could concentrate statistical power to better address this question and similar questions about optimal combinations of data. The second question asks how to help students adopt the goal of producing a general explanation. Our ID condition simply asked students to find an explanation that fit all of the data. Higher test scores for ID compared to CC suggest that this approach helped students adopt the goal of seeking the general explanation, as compared to instructions to compare and contrast the cases. However, these learning differences provide only indirect evidence of goal adoption, and here again, they were not statistically significant. A study providing more direct evidence of students' goals as they process data under different instructional conditions, for instance using think-aloud protocols, would settle the issue more conclusively.

The second limitation (or more precisely, concern) is that the inductive condition, while doing the best overall, yielded relatively low performance. One might dismiss the posttest performance because it had difficult items. But the worksheets are where the learning occurred, and in both experiments, only a third of the ID students found the deep structure, according to their worksheet explanations. There are three considerations that forestall grave concerns with this poor performance, while also pointing to future research needs.

The first involves the cases we selected. Francis Bacon rejected the idea of induction as searching randomly for patterns in data, which he warned would result in "fancies and guesses and ill-defined notions and axioms that have to be adjusted daily" (Bacon, 2000, Book 2, Aphorism XVIII). Instead, he set forth a clear theory of induction using structured comparisons. While this theory is simple to follow in principle, it is not straightforward to apply when designing inductive activities. In the present study, we did not follow Bacon's direction to assemble multiple cases of both positive and negative instances (Bacon's tables). We had two positive instances where the magnetic field changed in some way that was perpendicular to the coil (the bulb lit), but we only

had one negative instance where the field changed parallel to the coil, and the bulb stayed dark. A problem with a single negative instance is that it can be viewed as an anomaly (Chin & Brewer, 1993). For example, in Experiment 1 the negative case showed the vertical movement of the magnet. Because there were not two negative cases, students could use the surface feature of vertical movement to say why the bulb did not light, rather than the lack of a change to the horizontal component of the magnetic field. We suspect students would have done better had we included two negative instances with differing magnet movements, so that students could separate the surface from the deep structure that caused the bulb to light and not light.

The second consideration involves the restricted nature and timeline of the experiment. To isolate the effects of induction and to minimize class time, students did not receive any follow-up instruction such as a lecture or practice problems. Ideally, induction would not sit by itself without follow-up of some sort. Prior research indicates that even if one does not achieve the correct explanation, inductive tasks can create a time for telling (Schwartz & Bransford, 1998; Schwartz & Martin, 2004). The failure to find the deep structure does not mean students did not learn (Kapur, 2008). In the process of searching for a general explanation, students notice the deep features of the phenomena, and this readies them to appreciate the explanation that accounts for those features. So, while students engaged in induction earned modest scores on our posttest, the experience could have produced unmeasured but lasting benefits, helping them to learn related material in the future.

The third consideration is that the students in the current studies probably did not have much experience inducing general explanations. Typically, students are told general principles, which means the students in these studies may have never explicitly experienced the task of generalizing. A more disconcerting prospect is that they may have never learned that generalization is a major goal of scientific theory. In Experiment 2, all the conditions did well on standard formula application problems as might be found in many physics tests. These students knew how to do what they had been taught, so it seems reasonable they may not have been taught to seek general explanations.

Conclusion

In two studies, college students engaged in an inductive activity built on Bacon's tenets for scientific generalization. The studies demonstrate the promise of well-designed inductive instruction for learning topical content as part of developing generalizable theory from data. Several specific contributions are worth reviewing. First, we provided a model of an inductive activity that lends needed specificity to the strategy of having students search for and explain patterns in data. We worked from Bacon's original aphorisms to describe the selection of empirical cases to support induction and to provide the goal of finding a general explanation that can handle all the cases. Second, we found that this inductive activity led to greater student success at discovering the deep structure of the phenomenon, which in turn correlated with subsequent performance on novel but related problems. Third, we demonstrated a simple way to determine if students understand the purpose of an equation. While most of these students could manipulate the relevant equation to compute an answer, students who completed the inductive activities could more frequently describe why the formula included a specific operation within it (i.e., "why does the formula take the dot product?"). Finally, we found that college students needed support to seek general explanations, otherwise, they tended to explain each case independently rather than together.

One broad direction for future research would be to determine whether and how the Baconian model generalizes to other topics within physics and beyond. In a similar vein, it would be useful to further explore the transferability of inductively learned concepts. In the current studies, we used near transfer tasks. The posttest questions were novel situations, but they had many features

in common with the learning tasks, so students knew what knowledge they should apply. Far transfer tasks, where the problem does not resemble the conditions of initial learning, depend on the spontaneous recognition that what one has learned is relevant. It would be worthwhile to determine whether inductive learning can facilitate far transfer.

A second general direction for future research involves situating inductive activities within the fuller range of inquiry. For instance, how can students learn to select and organize data themselves for the purpose of inductive synthesis? How might teachers best integrate inductive synthesis with hypothetico-deductive analysis? As one example, induction could help students generate initial hypotheses, which they could then test using a hypothetico-deductive approach. While there are many effective activities designed to support hypothetico-deductive inquiry, there is always the question of where the hypotheses come from, especially when students have low prior knowledge. Inductive activities that support a systematic search for a hypothesis seems like a useful approach, at least compared to telling them a hypothesis, asking them to conjure a hypothesis from wispy beliefs, or having them mess about until they hit upon something.

Perhaps the most pressing issue involves the evidence that students do not spontaneously search for a general explanation across a set of manifestly related instances. In Experiment 1, and even in Experiment 2 that included an example of a general explanation, we found that the hypothetico-deductive task of generating and testing hypotheses rarely led students to seek the general explanation for all the cases. Instead, students were content to use features unique to each case at hand to explain a result. This is understandable. The search for specific situated solutions is characteristic of human cognition. But as Bacon argued, seeking a general explanation is a defining characteristic of disciplined scientific investigation and separates it from everyday reasoning. How can we help students develop an inductive scientific disposition? This is an essential question for science education given current goals for students to learn the attitudes and practices of scientific inquiry (NRC, 2012).

Acknowledgment

This work was supported by a Stanford Interdisciplinary Graduate Fellowship to the first author, and additional funding from the National Science Foundation (EHR-1020362) and the Institute of Education Sciences (R305A140314). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the granting agencies.

Notes

[1] The worksheets were completed individually but students worked in small groups of two or three to do the activities, a caveat for evaluating statistical comparisons of the worksheets.

[2] The worksheets were completed individually but students worked in groups of 2 or 3 for the activities, a caveat for evaluating statistical comparisons of the worksheets.

[3] We excluded the formula application problem since it was not designed to index deep understanding of the physics.

[4] Students in other conditions tended to give vague responses that did not focus on specific features.

References

Bacon, F. (1620/2000). In L. Jardine & M. Silverthorne (Eds.), The new organon. Cambridge: Cambridge University Press.

Bauer, H. H. (1992). Scientific literacy and the myth of the scientific method. Chicago: University of Illinois Press.

Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 153–166.

Biederman, I., & Shiffrar, M. M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. Journal of Experimental Psychology: Learning, Memory, & Cognition, 13(4), 640–645.

Bransford, J. D., Franks, J., Vye, N. J., & Sherwood, R. D. (1989). New approaches to instruction: Because wisdom can't be told. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning (pp. 470–497). New York: Cambridge University Press.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. Cognitive Psychology, 20, 493–523.

Bruner, J., Goodnow, J. J., & Austin, G. A. (1967). A study of thinking. New York: Science Editions.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(6), 1147.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. Cognitive psychology, 4(1), 55–81.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. Child Development, 70(5), 1098–1120.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science, 13, 145–182.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices*. Cognitive science, 5(2), 121–152.

Chin, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. Review of Educational Research, 63(1), 1–49.

Dega, B. G., Kriek, J., & Mogese, T. F. (2013). Students' conceptual change in electricity and magnetism using simulations: A comparison of cognitive perturbation and cognitive conflict. Journal of Research in Science Teaching, 50(6), 677–698.

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. Review of research in education, 32(1), 268–291.

Felder, R. (1993). Reaching the second tier: Learning and teaching styles in college science education. Journal of College Science Teaching, 23(5), 286–290.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive science, 7(2), 155–170.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. Journal of Educational Psychology, 95, 393–408.

Gentner, D., & Markman, A. (1997). Structure mapping in analogy and similarity. American Psychologist, 52(1), 45–56.

Gibson, J. J. (1979). The ecological approach to visual perception. Boston: Houghton Mifflin.

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment?. Psychological review, 62(1), 32–49.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15, 1–38.

Goodwin, C. (1994). Professional vision. American anthropologist, 96(3), 606–633.

Grandy, R., & Duschl, R. A. (2007). Reconsidering the character and role of inquiry in school science: Analysis of a conference. Science & Education, 16(2), 141–166.

Joshua, S., & Dupin, J. J. (1987). Taking into account student conceptions in instructional strategy: An example in physics. Cognition and Instruction, 4(2), 117–135.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math. Science, 320(5875), 454–455.

Kapur, M. (2008). Productive failure. Cognition and Instruction, 26, 379–424.

Karplus, R., Karplus, E., Formisano, M., & Paulsen, A. (1977). A survey of proportional reasoning and control of variables in seven countries. Journal of Research in Science Teaching, 14(5), 411–417.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). 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.

Khishfe, R., & Abd-El-Khalick, F. (2002). Influence of explicit and reflective versus implicit inquiry-oriented instruction on sixth graders' views of nature of science. Journal of research in Science Teaching, 39(7), 551–578.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. Psychological Science, 15, 661–667.

Lawson, A. (2010). How "scientific" is science education research. Journal of Research in Science Teaching, 47(3), 257–275.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning. American Psychologist, 59(1), 14–19.

Medawar, P. B. (1979). Advice to a young scientist. New York: Harper & Row.

Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem-solving, and induction. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 5, pp. 189–223). Hillsdale, NJ: Erlbaum.

McDermott, L. C. (1991). Millikan lecture 1990: What we teach and what is learned—closing the gap. American Journal of Physics, 59, 301–315.

Minda, J. P., & Ross, B. H. (2004). Learning categories by making predictions: An investigation of indirect category learning. Memory & cognition, 32(8), 1355–1368.

National Research Council. (2000). Inquiry and the National Science Education Standards: A Guide for Teaching and Learning. Washington, DC: The National Academies Press.

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

Nussbaum, J., & Novick, S. (1982). Alternative frameworks, conceptual conflict and accommodation: Toward a principled teaching strategy. Instructional science, 11(3), 183–200.

O'Kuma, T. L., Maloney, D. P., & Hieggelke, C. J. (2000). Ranking task exercises in physics. Upper Saddle River, NJ: Prentice Hall.

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. Instructional science, 32(1), 1–8.

Rittle-Johnson, B., & Star, J. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge: An experimental study on learning to solve equations. Journal of Educational Psychology, 99, 561–574.

Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. Review of Educational Research, 77(4), 534–574.

Roscoe, R. D., & Chi, M. T. H. (2008). Tutor learning: The role of explaining and responding to questions. Instructional Science, 36(4), 321–350.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. Journal of Experimental Psychology: Learning, 13, 629–639.

Ryoo, K., & Linn, M. C. (2014). Designing guidance for interpreting dynamic visualizations: Generating versus reading explanations. Journal of Research in Science Teaching, 51(2), 147–174.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. Cognition & Instruction, 16, 475–522.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. Journal of Educational Psychology, 103(4), 759–775.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for learning: The hidden efficiency of original student production in statistics instruction. Cognition & Instruction, 22, 129–184.

Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. Cognitive Psychology, 28, 225–273.

Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Granott & J. Parziale (Eds.), Microdevelopment: Transition processes in development and learning (pp. 31–58). New York: Cambridge University Press.

Strike, K. A., & Posner, G. J. (1992). A revisionist theory of conceptual change. In R. Duschl & R. Hamilton (Eds.), Philosophy of science, cognitive psychology, and educational theory and practice (pp. 147–176). New York: State University of New York Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive science, 12(2), 257–285.

White, R. T., & Gunstone, R. F. (1992). Probing understanding. London: Falmer.

Wieman, C. E., Adams, W. K., & Perkins, K. K. (2008). PhET: Simulations that enhance learning. Science, 322(5902), 682–683.

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. Science education, 92(5), 941–967.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. Cognitive Science, 34, 776–806.