Explaining across contrasting cases for deep understanding in science: An example using interactive simulations

Catherine C. Chase[1], Jonathan T. Shemwell, and Daniel L. Schwartz
Stanford University School of Education, 485 Lasuen Mall, Stanford, CA, 94305
cchase@stanford.edu, jshemwell@stanford.edu, danls@stanford.edu

Undergraduate students used a simulation to learn about electromagnetic flux. They were provided with three simulated cases that illustrate how changes in flux induce current in a coil. In the POE condition, students predicted, observed, and explained the outcome of each case, treating each case separately. In the GE condition, students were asked to produce a general explanation that would work for all three cases. A second factor crossed whether students had access to a numerical measurement tool. Effects of the measurement tool were less conclusive, but there was a strong effect of instructional method. Compared to POE students, GE students were better able to induce an underlying principle of electromagnetic flux during instruction and were better able to apply this principle to novel problems at post-test. Moreover, prior achievement predicted learning in the POE group, while students of all academic levels benefited equally from the GE condition.

Science education has learning goals that range from basic lab skills to beliefs about the sources of scientific knowledge. One enduring goal is for students to develop a deep understanding of phenomena so they can engage in the structure of scientific explanation. One way to characterize deep understanding is the capability and disposition to perceive and explain natural phenomena in terms of general principles. In this study, we show that deep understanding can depend critically on the way in which multiple instances of phenomena are presented to students and how students are instructed to explain those instances. The research is done in the context of undergraduate physics students learning about magnetic flux with a computer simulation.

It is common in science instruction to ask students to solve or conceptually explain a series of problems. One version of this approach is the Predict-Observe-Explain (POE) cycle (White & Gunstone, 1992). Students receive the set-up of an experiment and predict what will happen. They then observe the outcome and develop an explanation for why their prediction did or did not match the expected outcome. For POE and other sequenced formats, a series of questions or examples is carefully selected to help students instantiate a given core principle in multiple contexts, so that they develop a deeper, more abstract sense of the principle and learn the kinds of situations to which it applies. Formats such as POE are considered to be effective in part because they foster deep and often extended engagement with each new question or problem that students consider.

A risk of presenting students with a series of instances of a given principle is that students may treat each instance as unique and not grasp the underlying structure that links them together. Novices often have difficulty finding the underlying structure across instances that differ on the surface. In a classic study contrasting physics experts and novices, the experts categorized problems by their underlying concepts, such as energy conservation, whereas novices categorized them by their devices, such as springs or inclined planes (Chi, Feltovich, & Glaser, 1981). This encoding of surface instead of deep features would seem a likely pitfall of any pedagogy that engages students intensively with many instances of phenomena presented in series. For example, students doing POE might focus on the manipulation of a particular experiment, not noticing that it shares properties with a seemingly different manipulation. As a simple thought experiment, if a person adds a red solution to a beaker in one POE cycle to see what happens, and then adds a purple solution in the next cycle, it would be natural to treat red and purple as distinct manipulations, even though they are both cases of adding a color that contains red.

An alternative to instructional methods that have students work intensively with separate instances of phenomena is to have students explicitly consider multiple instances jointly. Contrasts among multiple, juxtaposed cases are known to support the induction of underlying structure if they differ on a few key dimensions (Beiderman & Shiffrar, 1987; Gibson & Gibson, 1955; Marton & Booth, 1997). Much like wine tasting, the process of comparing across cases helps people discern critical differentiating features that they might otherwise overlook (Bransford, Franks, Vye, & Sherwood, 1989). When students come to recognize invariant structure among cases with different surface features, they can schematize this invariant and more readily transfer this more general knowledge to new situations (Gick & Holyoak, 1980). Approaches to instruction that optimize contrasts have been successful in teaching statistics (Schwartz & Martin, 2004) and psychology (Schwartz & Bransford, 1998). O'kuma, Maloney & Hieggelke (2000) provide an example of this type instruction in science, wherein students are asked students to discover, apply, and explicitly state an underlying principle induced from a series of related cases.

However, merely engaging with contrasting cases does not automatically produce deep understanding. Our hypothesis was that *how* students were instructed to process multiple, related cases would be critical for determining whether they would notice and encode the underlying structure. In particular, without explicit prompting, students would be likely to treat each problem separately and miss the common underlying structure. This hypothesis is supported by research in the domain of analogical reasoning. For example, Loewenstein, Thompson, & Gentner, (1999) showed that simply asking students to process two cases presented together was not nearly as effective for schema abstraction as explicitly prompting them to compare the cases and describe their similarities. Likewise, Catrambone & Holyoak (1989) found that transfer of underlying principles was improved when students were explicitly asked to identify the deep features that were common to two analogs. In the current study, we furnished all students with a set of contrasting cases embodying a single underlying principle. We asked one group of students to provide a general explanation (GE condition) for all the cases, whereas the other group followed the more typical approach of predicting, observing, and explaining each case in turn (POE condition). Our hypothesis was that the GE approach would lead students to induce the underlying principle during the activities, which in turn, would lead to better understanding at post-test.

Scientific principles that explain natural phenomena often involve complex relationships that are difficult to conceptualize using everyday language. Mathematics can provide crucial vocabulary and syntax to support students' conceptual reasoning in the face of complexity. For example, researchers (Schwartz, Martin & Pfaffman, 2005) had younger students use POE with the balance scale (i.e., will the scale tip or stay balanced given weights on pegs at various distances from the fulcrum). They found that encouraging students to "invent math" to predict and explain the results led to much greater learning than encouraging students to explain in "words." Representing distances and weights as numbers enabled students to test possible relationships (i.e. the multiplicative relationship of weight and distance that balanced the scale) and make precise comparisons that were difficult to make using words.

The simulation used in the current study features a measurement tool that mathematizes the concept of magnetic field by expressing field intensity as numerical values separated into their vector components. We gave half of the students in the study access to this measurement tool and encouraged its use on the presumption that it would help them identify and reason more precisely about the contrasts and similarities across the three configurations.

In the current study, undergraduates in an introductory physics course learned about magnetic flux in the context of an interactive computer simulation. Simulations offer exciting new possibilities for science learning (de Jong, 2006;), but pedagogies for their use are new and evolving. Instructional design has focused on providing embedded scaffolds to support student inquiry in relatively open-ended tasks, so students produce optimal experimental runs of a simulation (e.g., de Jong, 2006). Rather than focus on inquiry, we took advantage of simulations' affordances for engaging students with a set of contrasting cases. To do this, we asked students to generate conceptual explanations from a series of three scenarios within a simulation. We expected that using the common POE model of instruction, which encourages intensive processing of individual cases, would lead students to see different scenarios in the simulation as unique, unrelated instances, like the red and purple solutions in our thought experiment. Therefore, we wanted to determine if the simple switch of asking students to find a general explanation for all the cases could overcome this likely problem and produce superior learning outcomes.

Thus, the design of the study was a 2 x 2, crossing the factors of General Explanation (GE) v. Predict, Observe, Explain (POE) by Measurement Tool (MT) v. No Measurement Tool (No-MT). We expected the GE group to gain a deeper understanding of magnetic flux because in comparing across cases, they would be more likely to induce the general principle. We also predicted that the GE-With Measurement Tool (GE-MT) condition would perform the best of all on our learning assessments, because the precision of mathematical representation would help them identify and reason about relevant contrasts.

Methods

Participants

Participants were 103 undergraduates in an introductory physics course on electricity and magnetism at a highly selective university. The study took place during one of the 50-min recitation sections associated with the course. Because many students needed to leave before the end of the section (often to get to another class), 23 students did not complete at least one of the four questions on the post-test, leaving us with complete data for only 80 students.

Design

Sections were assigned intact but at random to the four treatments: GE-MT (n=20; 3 sections), GE-No-MT (n=25; 3 sections), POE-MT (n=20; 3 sections), and POE-No-MT (n=15; 2 sections). The unequal numbers of students in each condition were due to variations in section size (6-15 students) and the odd number of sections. The eleven different sections were taught by six different teaching assistants, and all but one of the teaching assistants taught two sections. To compensate for teacher effects, each teaching assistant taught one GE and one POE section. Both sections for a given teaching assistant were then randomly assigned to either the MT or No-MT condition.

Procedure and Materials

During the lesson, students completed worksheets, which directed them to use the simulation to learn about magnetic flux in the context of electromagnetic induction (Faraday's Law). Students worked in groups of two to three with one laptop computer. All groups spent 25-30 minutes on the worksheets. Throughout the class, the teaching assistants moved from group to group, answering student questions. Teaching assistants were unaware of the study's hypotheses. At the end of class, students completed a brief post-test to assess learning outcomes.

The simulation (Figure 1) was a PhET interactive simulation (Wieman, Adams, & Perkins, 2008; available at http://phet.colorado.edu/simulations). This simulation allows students to move a magnet around the screen to light a bulb attached to a conducting coil. According to Faraday's Law, a changing magnetic flux in the coil will induce a voltage and light the bulb. The simulation represents the magnetic field as a sea of tiny test compass needles. Changes in field direction and strength are depicted by rotation of the needles and changes in their brightness.

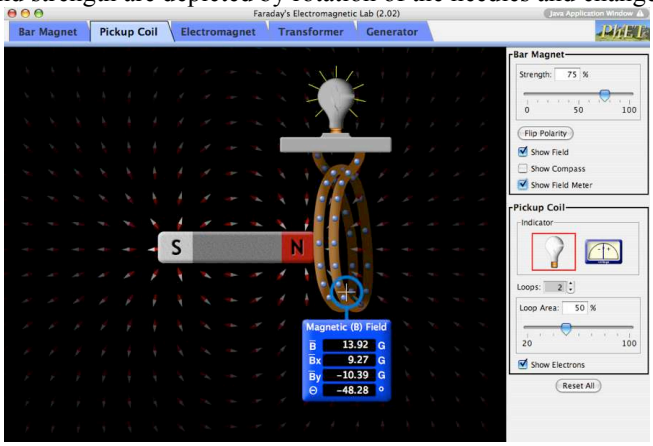

Figure 1. Phet Simulation. The magnet can be moved around the screen at varying speeds and positions to demonstrate how voltage is induced from magnetic field changes. The "field meter" measures field strength.

The worksheets presented all students with the same set of three cases (Figure 2). Two of the cases vary the magnet's position and one flips the magnet's polarity. A comparison of these cases reveals the invariant cause of voltage induction – a change in the component of the field within and perpendicular to the face of the coil. This translates to a change in the magnetic flux. The cases were designed to demonstrate three different manifestations of this underlying principle. Case A shows that a change in overall field strength can produce a voltage. Case C shows that a change in the field's direction can produce voltage. Case B illustrates that a change in field in the vertical direction does not produce voltage. Taken individually, it can appear that different kinds of changes are causing the voltage in each case. In Case A, the strength is changing; in Case C the direction is changing; in Case B there is almost no change in direction, and the change in strength is not particularly salient. Thus, a change in the field's strength or direction would seem sufficient to induce a voltage. But taken as a group, it is possible to induce the deep principle – that only a change in strength of the horizontal component of the field qualifies as a change in magnetic flux, which produces voltage.

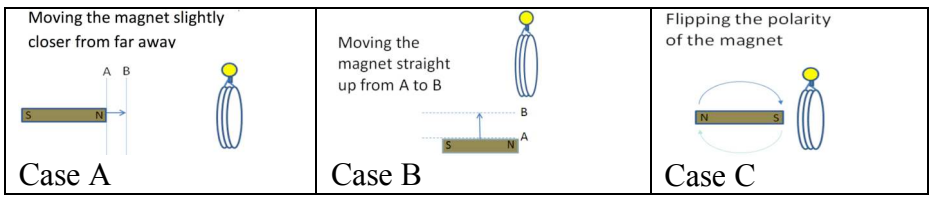

Figure 2. Cases. Students recreated these cases in the simulation and observed their effects.

Description of Conditions

In the POE condition, for each of the three cases, students made predictions, observed what happened, and explained why. In the predict phase, the worksheets instructed students to make predictions for each of the three cases in Figure 2. Specifically, they had to predict what the light bulb would do in each situation and draw expected changes to the magnetic field. In the observe and explain phase, students used the simulation to test their predictions by recreating each of the three cases. For each case, there was a space on the worksheet for students to record "what the light did," describe the light's brightness, draw the changes that happened to the magnetic field, and "explain the change in the magnetic field that caused the bulb to light." Thus, the POE condition was unlike many POE cycles in that students worked with all three cases at the same time. This was done so that the POE condition would be comparable to the GE condition, which also worked with the three cases simultaneously.

The GE worksheets did not contain a prediction phase. Instead, GE students were told whether the bulb would light brightly or dimly, after which they observed the cases and worked to generate a single, unifying explanation that would work across them all. The worksheet contained an example general explanation for how three cases of objects of varying masses and volumes would sink to varying depths of a liquid (the example general explanation described density as a ratio of mass to volume, which determines sink "depth"). After looking over this example, students were instructed to open the simulation, produce each of the cases, draw and record observations of the magnetic field, and then write "a single general explanation that will address what the magnetic field must do for the bulb to light or not light in any given case."

The field meter, an optional feature of the simulation (depicted in Figure 1) allowed users to take numerical measurements of the magnetic field. The field meter measured horizontal and vertical components of the field, the angle between the field vector and the vertical, and the overall magnetic field strength. Groups in the MT condition were given access to the field meter and told to use it to record horizontal and vertical components of field strength inside the coil. Studentsin the No-MT condition were told not to use the meter.

Dependent Measures and Coding

During the last 10 minutes of the lesson, students individually completed a six-item test assessing their understanding of the vector (perpendicular component) contribution to changes in magnetic flux in the context of Faraday's Law. Two of the items were dropped from our analyses because they proved to be unreliable measures of student understanding of magnetic flux. Figure 3 shows an example post-test item.

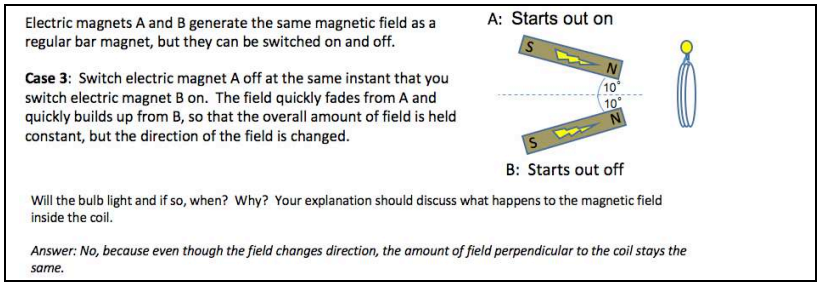

Figure 3. Sample post-test item.

Post-test responses were coded for whether or not the deep structure (the vector component nature of flux) was discussed, using a 1-0 coding scheme. An answer with a score of 1 applied the principle that changes in magnetic flux depend on changes in the component of magnetic field perpendicular to the coil. We further subdivided the non-deep answers into two categories: shallow and vague. Shallow answers depended on surface features by referring to a change in the strength or direction of the magnetic field as the causal agent. Vague answers referred to a general change in magnetic field as the causal agent, without further specifying the type of change. In this shallow-vague coding scheme, shallow answers earned a score of 1, while vague answers earned a score of 0. Worksheet explanations were also coded along these two dimensions: deep structure and shallow-vague. All questions were coded by two primary coders. For each question, a random sample (20%) of the data was double-coded to achieve inter-rater reliability. Percent coder agreement ranged from 80-100% across questions.

Results

Equivalence of Groups

To check the equivalence of students across experimental conditions, we compared groups on prior achievement as measured by students' course midterm scores and found no significant differences. A factorial ANOVA on midterm scores crossed instructional method (GE or POE) with measurement tool (MT or No-MT). There were no differences in scores by instructional method, $M_{GE} = 27.3$, $SE_{GE} = 1.3$, $M_{POE} = 26.5$, $SE_{POE} = 1.5$, $F(1,73) = 0.17$, $p = .68$, nor was there an interaction of instructional method with measurement tool, $F(1,73) = 0.02$, $p = .90$. There was a near main effect of measurement tool, as the MT group had lower scores than the No-MT group, $M_{MT} = 25.3$, $SE_{MT} = 1.4$, $M_{No\text{-}MT} = 28.7$, $SE_{No\text{-}MT} = 1.4$, $F(1,73) = 3.30$, $p = .07$. However, this difference was in the opposite direction of experimental effects (described below).

Post-Test Performance

Post-test measures revealed that the GE students developed a deeper understanding of the vector component nature of magnetic flux than POE students. There was a near-significant trend for MT students to outperform No-MT students, which suggests that using the field meter might also have helped students arrive at a deep understanding. Figure 4 depicts these patterns.

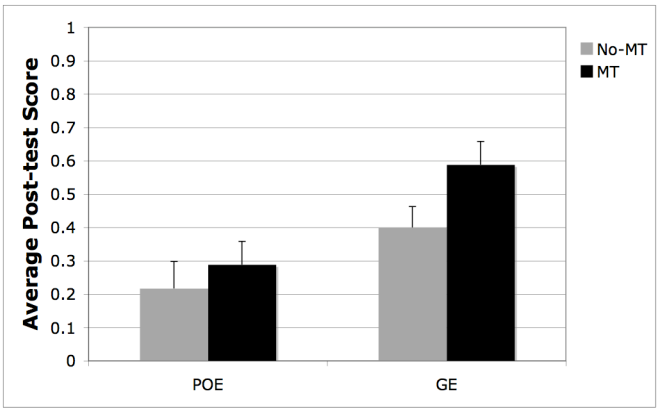

Figure 4. Average deep structure score across all post-test items, broken out by condition.

To test the effects of treatment on learning outcomes, a factorial ANOVA crossed method of instruction with measurement tool, using students' average deep structure score across all post-test items as the dependent variable. The ANOVA yielded a main effect for GE instruction, $F(1, 76) = 11.57$, $p = .001$, $d = 0.39$. There was also a near main effect of measurement tool, $F(1, 76) = 3.30$, $p = .07$. The interaction effect was not significant, $F(1, 76) = 0.67$, $p = .41$, though descriptively, the difference between MT and No-MT conditions was larger in the GE group.

Effects of Prior Achievement

Pre-existing achievement levels predicted learning outcomes, but only for the POE condition. Correlations between post-test and course midterm scores were non-significant for the GE group, $r = 0.03$, $p = .83$, but moderate and significant for the POE group, $r = 0.39$, $p = .03$. Both MT, $r = 0.28$, $p = .09$, and No-MT, $r = 0.12$, $p = .46$, students' post-test scores were uncorrelated with achievement. The low correlations between post-test and midterm for the GE groups suggest that the positive effect of GE instruction acted independently of students' prior achievement levels. The opposite occurred in POE instruction, where high achievers learn more from the instruction.

Worksheet Explanations

While working with the simulation, students in the GE condition wrote deep explanations on worksheets at a much higher rate than POE students (Table 1). This effect was pronounced. For the 80 students completing the experiment, only 1 out of 35 (2.9%) in the POE condition wrote a deep explanation compared with 14 out of 45 (31.1%) in the GE condition, χ^2 (1, N = 80) = 1.03, $p = .001$. Measurement tool, in contrast, did not significantly affect worksheet performance, χ^2 (1, N = 80) = 0.08, $p = .78$. So GE students were far more likely to induce the deep structure during the worksheet activity than POE students.

Table 1. Percentages of students who wrote deep explanations on worksheets (n deep/n total), by condition.

	No-MT	MT	Total
POE	0/15 (0.0%)	1/20 (5.0%)	1/35 (2.9%)
GE	7/25 (28.0%)	7/20 (35.0%)	14/45 (31.1%)
Total	7/40 (17.5%)	8/40 (20.0%)	15/80 (18.8%)

Relating Worksheet Explanations and Post-Test Performance

How students performed on worksheets predicted how they performed at post-test. A one-way ANOVA used worksheet explanation (deep or non-deep) as the independent variable and post-test score as the dependent measure. There was a substantial main effect of worksheet explanation, F (1, 78) = 30.08, $p < .001$, $d = .62$. Students who noticed the deep structure during instruction (and wrote about it in their explanations) were far more likely to apply the deep structure to novel problem situations on the post-test. Only one POE student discussed the deep structure of the worksheet, so it is impossible to determine whether worksheet performance predicts performance equally for both instructional conditions. Nonetheless, the findings are clear. Worksheet performance strongly predicted post-test performance, and students in the GE group were far more likely to perform well on the worksheet.

Descriptive Trends in Non-Deep Worksheet and Post-Test Responses

What were students' worksheet explanations and test responses, if not deep? Figure 5 shows the percentages of student worksheet explanations and post-test responses that were deep, shallow, and vague, broken out by condition. Overall, students across all four conditions gave similar proportions of vague responses, meaning that POE students did not simply neglect the task, at least not more than the GE students. However, the POE group had a higher percentage of shallow responses on both the worksheet and post-test, indicating that POE students tended to focus on locally salient aspects of the three cases (i.e. the field changed strength, or it changed direction). In comparison to MT students, No-MT students wrote a higher proportion of shallow explanations at test, though this pattern was far less prominent on worksheet performance. As is evident from the data already presented, the GE group generated a higher percentage of deep responses on both worksheets and tests.

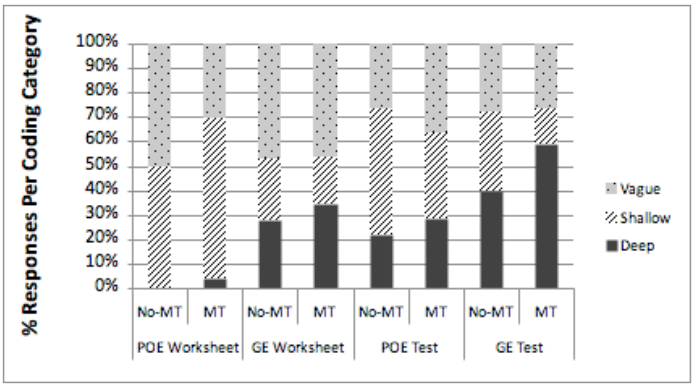

Figure 5. Percentages of deep, shallow, and vague responses on worksheets and post-tests.

Mortality Threat

As described in the methods section, 23 students did not complete one or more of the assessment items because they left the experiment during the assessment phase. These students' worksheet and partial test performances are omitted from the analyses described above. To check for mortality effects, we calculated the percentage of students who answered in terms of deep structure for each post-test question and on worksheets, including all student data (Table 2). Importantly, the pattern of worksheet performance including all 103 students is identical to the pattern for the 80 students who completed all post-test items. At post-test, GE students gave deep responses at a higher rate than POE students on all four items. Also, with the exception of Item 3, MT students gave deep responses at a higher rate than No-MT students. In a final check, we examined differences in prior achievement by comparing leaving and remaining students' midterm scores. Scores in the two groups did not differ, $M_{leave} = 26.8$, $SE_{leave} = 1.5$, $M_{remain} = 27.0$, $SE_{remain} = 1.0$, $t(1,97) = 0.09$, $p = .93$, nor did they differ by condition. Given the equivalent prior achievement and the results shown in Table 2, and especially considering the strong relationship between worksheet and test performance, it is unlikely that results described above were caused by attrition.

Table 2. Percentages of students with deep worksheet explanations and post-test responses, broken out by condition and item. Includes all student data.

		Worksheet (N = 103)	**Item 1** (N = 102)	**Item 2** (N = 101)	**Item 3** (N = 90)	**Item 4** (N = 82)
POE	No MT	0.0%	0.0%	5.6%	55.6%	22.2%
	MT	4.4%	13.0%	17.4%	47.8%	26.1%
GE	No MT	27.8%	16.7%	33.3%	58.3%	27.8%
	MT	34.6%	46.2%	57.7%	69.2%	38.5%

Discussion

Three strands of evidence show the superiority of working to create single, unified explanations of several simulated cases of a phenomenon (GE condition) over predicting and explaining those cases separately (POE condition). First, GE students had higher post test scores, showing their deeper understanding of the vector component nature of magnetic flux. Second, better worksheet performance in the GE group strongly linked this deeper understanding to the process of constructing a single explanation that satisfied all cases. Finally, in contrast to post-test scores for POE, post-test scores for GE were uncorrelated with prior achievement, indicating that this pathway to deeper understanding was open to students of all levels.

It is worth emphasizing that the students in all the conditions were engineering students at a selective enrollment university who had received instruction on magnetic flux in their main lecture class, many of whom had extensive prior experience with procedural basics including vector decomposition. Therefore, the lesson we taught was far from an introduction to the topic but rather an opportunity for students to rediscover "old" ideas in a new context. These special conditions suggest caution when generalizing our findings to contexts where students know less about the topic beforehand. Nevertheless, the non-correlation of test scores with midterm scores indicates that GE-type instruction could be beneficial for students with a broad range of prior knowledge. Here we note that many students in the study "forgot" about the component nature flux when working with the simulation despite their exposure to this topic in their course. Apparently, recitation activities that drive students to the task of "rediscovering" what they have been told in lecture is a valuable – perhaps indispensable – addition to student learning.

How did producing a general explanation help students develop a deeper understanding? Our favored interpretation is that when students search for a single explanation that applies to all cases, they are seeking and inducing the invariant under transformation. They are asking, "What is it about the three simulated experiments that are the same?" When successful, they induce an invariant relation that gives a principled account of different surface transformations. By contrast, students who work on cases in isolation, as the POE students did, are more apt to notice and think in terms of surface features that differ in each case. Of course, simply seeking a general explanation is insufficient if the data one collects are poor. The three simulation scenarios were designed *a priori* to include optimal contrasts that would help highlight the invariant structure. The GE students took greater advantage of these optimal contrasts than the POE students.

In addition to treating the cases either separately or together, there were other differences between the GE and POE conditions. Students in the POE condition predicted and observed the outcome of each case while those in the GE condition observed each case but were told the outcome. We cannot rule out these other differences as potential causes of the learning effects. However, given the nature of students' worksheet explanations, the "cross-case vs. within-case" explanation seems most plausible.

These results do not warrant a negative judgment on POE or similar pedagogies (such as sequential problem solving) because our implementation was not meant to represent optimal use of POE. Rather, the results show that asking students to induce a general explanation across a set of cases has the important consequence of facilitating a deeper understanding. Moreover, the results show that it can be risky to expect that students will generalize across multiple cases when they are asked to give a separate answer or explanation for each one.

In addition to the GE effect, the data included a trend in which the measurement tool helped both GE and POE students learn the underlying principle, according to post-test data. However, MT students did not have better worksheet explanations than No-MT students. We conclude from the worksheet data that, in opposition to our prediction, the reasoning afforded by this additional mathematical representation did not help students induce the underlying structure of magnetic flux from the three cases. Assuming that mathematical representations can facilitate students' conceptual reasoning, it seems that those we made available to students via the measurement tool were not appropriate, or were not appropriately structured, for the task of inducing a general explanation. Additionally, it is puzzling that MT students came near to outperforming No-MT students at post-test even though

their worksheet explanations were not different. One explanation for this disparity is that the post-test contained stronger cues for students' prior knowledge of the vector component nature of flux than the worksheet activity. These cues would likely have been of greater help to MT students as a result of their having worked explicitly with vectors via the measurement tool.

Pending further research, the current study provides two suggestions. One is for the design of simulations and another is for the design of instruction. Currently, the design of simulations suggests running multiple experiments, each with a single condition. This design pulls for something like a series of cases using a POE pedagogy, where students set parameters, make a prediction, observe, and explain. This makes it difficult to compare across multiple conditions, which is what real world experimenters often do. Perhaps simulations should allow the presentation of multiple cases simultaneously on the screen. This would permit the production of optimal contrasts, and with proper orientation, students could be guided to generate general explanations. A second learning from this study, confirming prior research and extending it to the context of conceptual learning in science, is that when students are presented with several instances of a phenomenon, they will not automatically search for the common structure that exists across them. Rather, students need to be pushed to do this. Science instruction that compels students to generate explanations that work for several different experiments or situations can help students construct deep understandings of general principles and the conditions to which they apply.

Endnotes

(1) The first two authors contributed equally to this work and are listed alphabetically.

References

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, and Cognition, 13*, 640-645.

Bransford, J. D., Franks, J. 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). NY: Cambridge University Press.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(6), 1147-1156.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

De Jong, T. (2006). Technological advances in inquiry learning. *Science, 312* (5773), 532-533.

Dean, D.J., & Kuhn, D. (2006). Direct instruction vs. discovery: The long view. *Science Education, 91*, 384-397.

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment. *Psychological Review, 62*, 32-51.

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

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review, 6*(4), 586-597.

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

Marton, F., & Booth, S. (1997). *Learning and awareness*. Mahwah, NJ: Erlbaum.

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

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.

Schwartz, D. L., Martin, T., & Pfaffman, J. (2005). How mathematics propels the development of physical knowledge. *Journal of Cognition and Development*, *6*, 65-88.

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

White, R. & Gunstone, R. (1992). *Probing Understanding*. London: Falmer Press.

Acknowledgements

We would like to thank Chaya Nanavati for her guidance and support with implementation, and the Physics Education Technology (PhET) Project for use of their simulation.