International Journal of Artificial Intelligence in Education, 10(4), 2010.

The Effect of Self-Explaining on Robust Learning

Robert G. M. Hausmann, *Carnegie Learning, Inc., Pittsburgh, PA, USA*
bhausmann@carnegielearning.com

Kurt VanLehn, *School of Computing and Informatics, Arizona State University, Tempe, AZ, USA*
kurt.vanlehn@asu.edu

Abstract. Self-explaining is a domain-independent learning strategy that generally leads to a robust understanding of the domain material. However, there are two potential explanations for its effectiveness. First, self-explanation generates additional *content* that does not exist in the instructional materials. Second, when compared to comprehension, *generation* of content increases understanding and recall. An *in vivo* experiment was designed to distinguish between these potentially orthogonal hypotheses. Students were instructed to use one of two learning strategies, self-explaining and paraphrasing, to study either a completely justified example or an incomplete example. Learning was assessed at multiple time points and levels of granularity. The results were consistent, favoring the generation account of self-explanation. This suggests that examples should be designed to encourage the active generation of missing content information.

Keywords. Self-explanation, physics, worked-out examples, problem solving

INTRODUCTION

An important domain-independent learning strategy is self-explaining, which is defined as the sense-making process that an individual uses to gain a greater understanding of some instructional material, including texts, worked-out examples, diagrams, and other multimedia materials by explaining it to themselves (as opposed to another person) (Roy & Chi, 2005). Self-explaining has consistently been shown to produce learning gains in several domains, including physics (Chi & Bassok, 1989; Chi, Bassok, Lewis, Reimann, & Glaser, 1989), the human circulatory system (Butcher, 2006; Chi, DeLeeuw, Chiu, & LaVancher, 1994; Hausmann & Chi, 2002), geometry (V. A. W. M. M. Aleven & Koedinger, 2002) and many others (Atkinson, Renkl, & Merrill, 2003; McNamara, 2004; McNamara, Levinstein, & Boonthum, 2004; Moreno, 2006; Renkl, 1997). Moreover, self-explaining has been used in several different learning contexts, including in the laboratory (Butcher, 2006; Chi, et al., 1989), in the classroom (McNamara, 2004; McNamara, et al., 2004), with prompting from humans (Chi, et al., 1994) and from computers (V. A. W. M. M. Aleven & Koedinger, 2002; Conati & VanLehn, 2000; Hausmann & Chi, 2002).

Although the effect has been widely replicated, it is still not clear why self-explanation works. Two potential explanations will be addressed in the paper that follows. The first explanation is that differences in the *content* are responsible for the increased learning gains. That is, self-explaining generates additional content that is not present in the instructional materials. The second explanation is that it is the *process*, rather than the product, that matters. That is, the process of generating the extra

content is more important for learning than the content itself. Let us provide names for the hypotheses: the Coverage hypothesis and the Generation hypothesis.

As an illustration, and a brief glimpse into the electrodynamics task domain used in the present experiment, suppose an example is being presented as a video. The video starts with a problem that shows an electric field (represented by parallel arrows; see Appendix B) and a positively charged particle (represented by a dot). The example video draws an arrow with its tail on the dot and says, "We draw a vector representing the electric force on the particle due to the electric field." This is a step in the example, so let us illustrate the process of self-explaining by considering two possible student responses:

1. Suppose the student says, "OK, so there's a force on the particle parallel due to the field." This does *not* count as a self-explanation; it is merely a paraphrase of the step.

2. On the other hand, suppose the student says, "The charge is positive, so the force is in the same direction as the field; if it had been negative, it would be in the opposite direction." This *does* count as a self-explanation because it goes well beyond what the example states.

However, suppose the example video itself said, "The charge is positive, so the force is in the same direction as the field; if it had been negative, it would be in the opposite direction." and the student read it and even paraphrased it. Would simply being aware of this extra content have the same benefits as generating it? The Coverage hypothesis says that it would, but the Generation hypothesis says that it would not.

Because examples are often addressed in Cognitive Load Theory (Paas, Renkl, & Sweller, 2003), it is worth a moment to discuss the theory's predictions. The theory defines three types of cognitive load: *intrinsic* cognitive load is due to the content itself; *extraneous* cognitive load is due to the instruction and harms learning; *germane* cognitive load is due to the instruction and helps learning. Renkl and Atkinson (2003) note that self-explaining increases measurable cognitive load and also increases learning, so it must be a source of germane cognitive load. This is consistent with both of our hypotheses. The Coverage hypothesis suggests that the students are attending to more content, and this extra content increases both load and learning. The Generation hypothesis suggests that load and learning are higher when generating content than when comprehending it. In short, Cognitive Load Theory is consistent with both hypotheses and does not help us discriminate between them.

The two hypotheses are pragmatically important. If the Coverage hypothesis is correct, then instruction should include examples that are as completely explained as possible—this is what many instructors tend to do in lectures and texts. On the other hand, if the Generation hypothesis is correct, then instructors should provide sparse examples and somehow motivate the students to fill in the explanations. Thus, the Coverage and Generation hypotheses have important implications for the design of instruction.

Explaining self-explaining

This research focuses on self-explanation of *examples*. An example is a solved problem, where the solution is derived in a series of steps (VanLehn, 1996); however, examples are typically incomplete. Although each step is produced by applying one or more knowledge components, neither the knowledge components nor the details of their application are mentioned in an incomplete example

(Zhu & Simon, 1987). The student is left to infer them instead. For instance, in the illustration above, the example asserted that a force existed in a certain direction, and the student had to apply the definition of electric fields to figure out why. Self-explaining an example consists mostly of applying knowledge components in order to justify and connect steps in an incomplete solution (Chi & VanLehn, 1991; VanLehn & Jones, 1993; VanLehn, Jones, & Chi, 1992). The self-explanation effect is the by-now common finding that students who self-explain incomplete examples learn more than students who do not (V. Aleven & Koedinger, 2000; Pirolli & Bielaczyc, 1989; Reimann & Neubert, 2000; Renkl, 1997; Renkl, Stark, Gruber, & Mandl, 1998). Although most examples found in textbooks are incomplete, it is possible to construct complete examples. A complete example justifies every step in terms of domain principles, definitions, and other knowledge components. This assumes that the domain has well-defined and accepted principles, definitions, etc. That is arguably the case for introductory physics, which is the task domain used in this study.

Several discussions of the self-explanation effect also mention the generation effect (Jacoby, 1978; Slamecka & Graf, 1978), which is a well-known finding in the memory literature, wherein subjects who participate in the generation of paired-associates have a higher probability of recall than participants who are merely presented with pairs (i.e., the read-only condition). It is quite a leap to apply a hypothesis about low-level recall to a whole cognitive skill, and its truth is not a foregone conclusion (for preliminary evidence, see deWinstanley, 1995; deWinstanley & Bjork, 2004; Peynircioglu & Mungan, 1993). As mentioned earlier, we call this conjecture the Generation hypothesis. To put it precisely, students learn more when they *generate* the explanation for example steps than when they merely read and *comprehend* the explanations. Thus, complete examples should be less effective than incomplete examples provided that students self-explain all the steps in the incomplete examples.

Lovett (1992) tested the Coverage and Generation hypothesis with permutation and combination problems. Lovett's 2 x 2 design crossed the source of the solution (subject vs. experimenter) with the source of the explanation for the solution (subject vs. experimenter). Thus, the experimenter-subject condition had students study an incomplete example without any prompting for self-explanation. The experimenter-experimenter condition had students study a complete example. Lovett found that the subject-experimenter condition was so confusing to the students that they required six times as many hints—thus, the fact that it produced low learning is not surprising and irrelevant to testing the hypotheses. Lovett analyzed errors and found that outcome differences among the remaining three conditions was mostly due to a single knowledge component, called numerator-starting-value (NSV) in her cognitive task analysis. Students in the experimenter-experimenter condition always heard about NSV during the experimenter's explanation, and they mostly applied it successfully during the post-test. About half the students in the experimenter-subject condition mentioned NSV during their explanations, and only those students applied it during testing. Students in the subject-subject condition probably had to apply NSV in order to generate the first step in their solutions, and they almost all applied it successfully again during testing. Thus, it appears that learning occurs if and only if NSV is either mentioned in experimenter's explanation, mentioned in students' explanation, or applied by the student. This is consistent with the Coverage hypothesis and not the Generation hypothesis. The Generation hypothesis would predict that when the experimenter mentions NSV, students are not likely to learn it.

Brown and Kane (1988) found that explanations provided by children between the ages of four and seven, either spontaneously or in response to prompts, were much more effective at promoting transfer than those provided by the experimenter. On the face of it, this is consistent with the

Generation hypothesis and not the Coverage hypothesis. However, the students who were told the rule may not have paid much attention to it, according to Brown and Kane. If so, then the experiment is not a good test of the hypotheses because the Coverage hypothesis requires that students attend to the explanations they are given.

Stark (1999) found that students who studied incomplete examples demonstrated stronger performance on near- and medium-transfer problems than students who studied complete examples. However, the benefit for incomplete problems did not reach traditional levels of statistical significance for far-transfer problems (summarized in Renkl, 2002, p. 533). Thus, the evidence from this study is mixed. The null result for far-transfer problems is consistent with the Coverage hypothesis only. The positive result for near- and medium-transfer problems is consistent with the Generation hypothesis only.

Manipulating the completeness of the examples is analogous to studies that compare learning from differentially elaborated texts. Across several experiments, McNamara et al. (McNamara, 2001; 1996) manipulated the completeness of a text and found that the learning outcomes depended on the prior knowledge of the students. Low-knowledge students learned best from elaborated texts. This is consistent with both the Coverage hypothesis and the Generation hypothesis because low-knowledge students are unlikely to generate the key connecting inferences left unsaid by the unelaborated text. On the other hand, the high-knowledge students learned better from the unelaborated texts. This is consistent with the Generation hypothesis and not with the Coverage hypothesis, which would predict a null result because students read the knowledge components in the elaborated text and they generated the knowledge components (during self-explanation) for the unelaborated texts. However, another possible explanation for this result is that the high prior-knowledge students' reading strategies were disrupted by the elaborated texts. Because the text supplied the knowledge components, they simply became passive readers and often failed to attend to the knowledge components presented by the text. Being passive thereby hurt their comprehension of the text. It also salvages the Coverage hypothesis, which requires that students attend to the presented knowledge components, and passive readers often do not do so.

Support for the passive-reader interpretation can be found in a study by Gilabert, Martinez, and Vidal-Abarca (2005). They designed elaborations for a text that simultaneously increased coherence, provided causal inferences, and encouraged active processing. They found that both high- and low-prior knowledge students learned more from the elaborated text than the original text. These results are consistent with the Coverage hypothesis and not with the Generation hypothesis. Moreover, this finding supports the passive-reader interpretation of the McNamara et al. results, thus undercutting its support for the Generation hypothesis.

In summary, the experimental record is mixed. Some studies (e.g., Lovett; Gilabert et al.) support the Coverage hypothesis; some studies (e.g., Brown & Kane; McNamara et al.) support the Generation hypothesis; and others studies (e.g., Stark et al.) support both. Although the main manipulation was providing incomplete versus complete examples and texts, other variables surfaced. Students with low prior knowledge may be unable to self-explain an incomplete text or an example. Students may not attend to the explanations in complete examples or texts.

Thus, to test the Coverage and the Generation hypotheses, an experiment is needed that controls for both the students' prior knowledge and their engagement in self-explaining or comprehending the examples. The experiments reviewed earlier seem not to have had adequate control of these important variables.

DESIGN AND PREDICTIONS

The ideal experiment would compare Generation versus comprehension of the same explanations for example steps. The first challenge is to get students in the generation condition to actually generate most of the explanations. To do this, we prompted for self-explanations of steps in incomplete examples because such prompting has been shown to vastly increase the frequency of self-explanation (Chi, et al., 1994; Renkl, 1997). The second challenge is to get students in the comprehension condition to attend to the presented justifications, while simultaneously not self-explaining them. To do this, we had students paraphrase complete examples because paraphrasing provides an overt, behavioral indication that the individual has at least attended to the instructional materials and tends to reduce the frequency of self-explanation (Hausmann & Chi, 2002). Thus, the two main conditions are self-explanation of incomplete examples and paraphrase of complete examples. Let us use *SE-incomplete* and *P-complete* as the names for these conditions.

If compliance of the participants with the instructions was the same in both conditions, then the Generation hypothesis predicts that the learning gains of SE-incomplete should be larger than P-complete, and the Coverage hypothesis predicts a tie in learning gains. That is, if students in SE-incomplete condition generated 90% of the content that they should, and the students in the P-complete condition comprehended 90% of the content that they should, then we would expect a tie (Coverage hypothesis) or a difference in favor of the SE-incomplete condition (Generation hypothesis). However, suppose it is less likely for students to self-explain than to comprehend, as seems plausible. That is, suppose students in the SE-incomplete condition generate 75% of the content that they should, and students in the P-complete condition comprehend 90% of the content that they should. Then the Coverage hypothesis predicts that the SE-incomplete condition would learn *less* than the P-complete condition. Written symbolically, the predictions are:

1. Generation hypothesis: P-complete < SE-incomplete
2. Coverage hypothesis: SE-incomplete ≤ P-complete

As a manipulation check, we included two more conditions: P-incomplete and SE-complete. In the P-incomplete condition, students are given an incomplete example and prevented from self-explaining it by being prompted to paraphrase it. Thus, they should neither generate nor comprehend the target content, so both hypotheses predict that students in the P-incomplete conditions will learn the least. In the SE-complete conditions, students were prompted to self-explain a complete example. The self-explanation prompting should act like the active engagement manipulation of Gilabert *et al.* (2005), so we'd expect the same amount of self-explanation in SE-complete as in SE-incomplete. Thus, the Generation hypothesis predicts SE-complete = SE-incomplete. We'd also expect students who are prompted to self-explain a complete example to comprehend at least as much as students prompted to paraphrase it. Thus, the Coverage hypothesis predicts P-complete ≤ SE-complete. Here is a summary of all the predictions:

1. Generation hypothesis:	P-incomplete	<	*P-complete*	<	*SE-incomplete*	=	SE-complete
2. Coverage hypothesis:	P-incomplete	<	*SE-incomplete*	≤	*P-complete*	≤	SE-complete

In the results section, we will report measures for all four conditions, with a special focus on comparing P-complete to SE-incomplete (italicized above), as they are the conditions where the hypotheses' predictions differ.

METHOD

LearnLab: Knowledge components

The data were collected in the Pittsburgh Science of Learning Center's physics LearnLab.[1] A LearnLab is a course that is designed for conducting rigorous, *in vivo* experiments on issues relating to robust learning. Robust learning is defined in three parts. First, learning is considered robust when the knowledge is retained over significant periods of time (i.e., long-term retention). Second, robust learning is the opposite of inert knowledge in the sense that students are able to broadly apply their knowledge to other problems within the same class of problems, as well as across different classes of problems (i.e., far transfer). Finally, robust learning expedites the acquisition of new information (i.e., acceleration of future learning). The benefit of collecting data in this environment, as opposed to the laboratory, is that the realism of the classroom increases the generalizability of the results, without sacrificing randomization or control over other extraneous variables. Therefore, the results obtained in a LearnLab classroom should be easily assimilated into non-LearnLab classrooms.

One assumption made by the Pittsburgh Science of Learning Center is that knowledge can be decomposed into smaller units, which will be referred to as *knowledge components* (VanLehn, 2006). A knowledge component is defined as any piece of knowledge that can be learned and applied independent of other knowledge components. An example of a knowledge component for the domain of electrodynamics is, "If a charged particle is in an electric field, then there will be an electric force on the particle due to the field." Another knowledge component is, "If a particle is positively charged, then its electric force is parallel to the electric field; otherwise, the electric force is anti-parallel to the field." Because the first piece of knowledge can be known without knowing the second, they are both treated as knowledge components.

Participants

To test the learning of these knowledge components, students were recruited from five, second-semester, calculus-based physics courses (Physics II: Electricity and Magnetism) taught at the U.S. Naval Academy. Of the 113 available students, 106 students volunteered and provided informed consent. Two students, who provided informed consent, were absent the day of the experiment; therefore, the total sample size was $N = 104$ students. The student volunteers were given course credit for their participation.

---

[1] http://www.learnlab.org/learnlabs/physics

Materials

The materials used for the experiment were developed in association with one of the LearnLab instructors and two other physicists. The domain covered during the experiment was electrodynamics, with an emphasis on the forces acting on a charged particle located in a region with an electric field. A complete specification of the problems can be found in Appendix A.

The problems were solved by the participants using the Andes Physics Tutor (Gertner & VanLehn, 2000; VanLehn, et al., 2005). Andes is an intelligent tutoring system designed to replace the paper-and-pencil problems located at the end of each chapter (see Appendix B for a screenshot of Andes). Students are expected to enter each solution step, which is analogous to instructors expecting that students "show all of their work." The benefit of using Andes for homework is that students receive feedback on each step of the solution, as well as on-demand help. The feedback is in the form of red and green flags that alert the student to incorrect and correct entries, respectively. The assistance Andes provides is either in response to the student asking what's wrong with an incorrect entry (i.e., what's wrong help) or the student asking what step should be taken next (i.e., next-step help).

In addition to solving problems with Andes, the students also studied examples. The examples were videos of a screen-logging program that captured the actions of an expert solving the problems in Andes. In addition, the examples' completeness was manipulated by using an audio track that described each action either with or without a justification for the action. Excluding justifications as a manipulation of example completeness has been used with success in other studies [33, 40-42]. For instance, in a complete example, the voice-over included a statement of why the definition of an electric field equation ($\mathbf{F} = q\mathbf{E}$) was used in the solution (see Appendix D).

The second manipulation was the type of study strategy that the students were instructed to use at the beginning of the experiment and prompted throughout the study. For the self-explaining conditions, students were given a description of what self-explaining entails, as well as an example of a hypothetical student self-explaining a concept from first-semester physics (see Appendix C). For the paraphrasing conditions, they received almost identical instructions, with the major change being a description of what paraphrasing is and an example of a student paraphrasing the same physical concept.

In addition to the given descriptions and examples, students were prompted throughout the experiment to engage in their respective study strategies. The annotated, worked-out examples were broken into seven to ten segments, depending on the complexity of the problem. At the end of each segment, the voice-over prompted the student to either, "Please begin your self-explanation" or, "Please begin your paraphrase." Below the window where the video was replayed, there were four prompts that were taken from the instructions (see the bulleted lists in Appendix C). These written prompts served as an additional cue for the student to engage in a particular study strategy.

The materials were designed with two constraints in mind. First, every example was isomorphic to the previously solved problem. The only difference between the two was the surface features of the problem, such as the values of the givens and, in some cases, the directions along which the motion or force was directed (i.e., the vertical or the horizontal axis). Second, the problems were designed such that they grew in conceptual complexity (see the right-most column in Appendix A). That is, the principles from the first problem were also addressed by the second problem, and the second problem's principles were included by the third. More importantly, each subsequent problem added a new principle to the previous problem. This nested structure of problems allowed the experimenters to track performance on specific knowledge components over time. The implication is that the current

experiment does not conform to the conventional pretest-intervention-posttest design used in most educational research. Instead, it takes individual knowledge components as the unit of analysis and tracks the student's performance over several opportunities to apply them.

Design

The experiment was a 2 x 2 x 3 mixed-factorial design. Two independent variables were crossed. Study strategy (paraphrase vs. self-explain) and example type (complete vs. incomplete) were between-subjects variables; whereas problem (PROB1, PROB2, vs. PROB3) was a within-subjects variable.

Students were block-randomized into condition. That is, care was taken to ensure that students were randomly assigned to condition under the constraints that grade point average (GPA), prior Andes usage, major, and the letter grade from the previous physics course (i.e., General Physics I) were equally represented in each experimental condition. The sample size for each condition was: P-complete: paraphrasing complete examples (n = 26), P-incomplete: paraphrasing incomplete examples (n = 23), SE-complete: self-explaining complete examples (n = 27), and SE-incomplete: self-explaining incomplete examples (n = 28).

Procedure

As stated previously, participants were recruited from five sections of second-semester physics at the U.S. Naval Academy (i.e., General Physics II). The experiment took place during one of the class periods, which was approximately 110 minutes in duration. As students logged into the system at the start of class, they were randomly assigned to an experimental condition. The system then introduced students to the experiment and displayed the learning strategy instructions that corresponded with their experimental condition (see Appendix C). Upon reading the instructions, the students were then prompted to solve the first problem. The first problem was relatively easy and only required one principle application (see the practice problem in Appendix A). When the first problem was completed, the students were then instructed to play the first example video (i.e., EX1). This process, alternating between solving problems and studying video examples, repeated for three cycles so that, by the end of the experiment, four problems were solved and three examples were studied. The procedure of iterating between studying examples and solving problems was analogous to the Alternating Example condition in Trafton and Reiser (1993).

It should be noted, however, that the student solved a problem *first*, and then they studied an isomorphic example *afterwards*. This sequencing of problems and examples departs from traditional methods used in prior research (Catrambone & Yuasa, 2006; Sweller & Cooper, 1985; Trafton & Reiser, 1993). We inverted the isomorphic pairs so that we could use learning curves to plot increasing competence. That is, for every knowledge component introduced to the students during the experiment, it first occurred in a problem, then in an isomorphic example. The problem played the role of a pre-test and furnished the first point on the learning curve. The problem that followed the example played the role of a post-test or mid-test, and furnished the next point on the learning curve.

While the students were studying the examples, they were prompted to either paraphrase or self-explain at the end of each segment. To capture their verbalizations, each student wore a pair of headphones equipped with a close-talk, noise-cancelling microphone (Andrea ANC-750 CTI Stereo Headset). The headphones and noise cancellation were necessary because each classroom had

approximately 24 students, each studying and solving problems simultaneously. The audio was digitally recorded and stored on the local machines. In addition to audio, all of the on-screen activity was recorded using a built-in, screen-logging facility. The following data-streams were created for each student: a.) an audio track of their verbalizations; b.) a movie of their on-screen activities; and c.) a text-only log file of each action in the Andes interface; d.) homework log files from the entire semester; and e.) exam performance on electrodynamics.

The training period ended when the student logged off of the Andes system. Students were free to leave the training at any time; however, only nine students left early due to scheduling conflicts. The remaining students left either after the last problem was solved, when the class period was over (i.e., after 110 minutes had elapsed), or whichever occurred first.

RESULTS

Equivalent groups check

Before turning to the problem-solving and learning results, it was necessary to ensure that the randomization procedure was effective in producing equivalent groups. Participants were block-randomized such that there were initially equal sample sizes for each of the four conditions (i.e., P-incomplete, P-complete, SE-incomplete, SE-complete). Five variables were checked to ensure equal groups. The variables included GPA (based on a four-point system), major (broken down into three categories: Engineering, Science, and Other), previous Andes usage (i.e., a subset of the students used Andes in Physics I: SP211), the letter grade for Physics I, and performance on the practice problem, as measured by the normalized assistance score.

According to an Analysis of Variance (ANOVA), there was not a statistically reliable difference between experimental conditions for GPA, $F(3, 100) < 1$. Furthermore, a chi-squared analysis revealed that there were no differences between conditions for major ($\chi^2(6, N = 104) = 3.70$, $p = .72$), previous Andes usage ($\chi^2(3, N = 104) = 1.50$, $p = .68$), nor the first semester physics grade ($\chi^2(12, N = 104) = 11.49$, $p = .49$). An ANOVA on practice problem assistance scores revealed no reliable differences, $F(3, 99) = .37$, $p = .78$. Based on these five metrics, it is reasonable to assume that the experiment started with equivalent groups. Therefore, the statistical analyses that follow were conducted without controlling for any of these variables.

When the log files from the homework assignments were analyzed, it was discovered that some of the students had done homework problems on electric fields before the experiment started; however, these students turned out to be equally distributed among the conditions, $\chi^2(3, N = 104) = 4.96$, $p = .18$.

Manipulation check

In addition to prior knowledge and ability, we also evaluated whether or not the students followed the directions outlined at the beginning of the experiment. The students were instructed to alternate between two activities. They were asked to first solve a practice problem using Andes. Then they were asked to watch a short video on how to solve an isomorphic problem. The problems were represented in a list, starting with the first activity at the top and the next activity below it. The results of this study are based on three assumptions.

First, it is assumed that the students would read the instructions at the beginning of the experiment. To check this assumption, an analysis of the students' reading duration of the instructions was undertaken. It was found that this assumption was essentially correct. Ninety-two percent (96 / 104 = 92%) of the students read the instructions.

Second, it was expected that the students worked on their problems and watch the videos in order. About one-third of the students (35 / 104 = 34%) completely solved all of the problems and watched all of the videos in the exact order prescribed by the instructions.

Finally, it was assumed that the students would watch the whole video. Approximately one-third (101 / 312 = 32%) of the videos were viewed in their entirety. Based on these analyses, the participants may not have acted in accordance to our assumptions. Therefore, our conclusions drawn from the data analyses will require a caveat that not all of the videos were completely viewed by all participants at each step of the experiment.

Example study time and time-to-solution

The amount of time dedicated to studying the examples did not differ between conditions, $F(3, 100) = .31$, $p = .82$. Likewise, the amount of time taken to solve the problems did not differ between conditions, $F(3, 100) = .81$, $p = .49$. The lack of difference in both the time-to-study and the time-to-solution argues against a time-on-task interpretation of the results.

Normalized assistance score

A *normalized assistance score* was used as our dependent measure to gauge the impact of the experimental conditions on learning. The normalized assistance score was defined as the sum of the help requests and errors per problem, divided by the number of entries made in solving that problem. That is, we count the amount of assistance students receive (summing help requests and immediate feedback on errors) normalized by the number of opportunities for such assistance. Thus, lower assistance scores indicate that the student derived a solution while making fewer mistakes and getting less help, and thus demonstrating better performance and understanding.

Normal learning

Problems solved during the experimental session

As a measure of normal learning (as opposed to robust learning, which was defined earlier), normalized assistance scores were averaged over individuals for all three problems in the training set (see *Fig. 1*). Notice that the y-axis in this figure is inverted from the customary display, so that higher bars represent *less* learning. The bars are ordered left-to-right in accord to the predictions of the Generation hypothesis (see top portion of *Table 1*). If the Generation hypothesis holds, then the bars should get shorter as they move rightward. The predictions for the Coverage hypothesis are listed in the bottom portion of *Table 1*. If the Coverage hypothesis holds, then the heights of the two middle bars should be reversed. To test these hypotheses statistically, contrast-coded predictors (found in parentheses) were created and tested as single degree of freedom, planned comparisons within the overall repeated-measures ANOVA.

Table 1
The predictions and contrast codes for the two rival hypotheses.

Generation Hypothesis	P-incomplete	<	P-complete	<	SE-incomplete	=	SE-complete
Contrast Codes	(-4)		(-2)		(3)		(3)
Coverage Hypothesis	P-incomplete	<	SE-incomplete	≤	P-complete	≤	SE-complete
Contrast Codes	(-4)		(-2)		(3)		(3)

The pattern of means for the normalized assistance scores is consistent with the predictions of the Generation hypothesis (see *Fig. 1*). A repeated-measures ANOVA confirmed that the visually apparent difference between the P-incomplete and P-complete conditions were higher than the SE-incomplete and SE-complete conditions. This relationship was statistically reliable with a medium effect size, $F(1, 73) = 6.17$, $p = .02$, $\eta_p^2 = .078$.[2]

[2] Partial eta squared (η_p^2) is an effect-size measure, which is equivalent to R^2 or the variance accounted for by the predictor variable. Traditional interpretations of partial eta squared are as follows: >.2 is a large effect size; >.1 is a medium effect size; >.05 is a small effect size (Cohen, 1988).

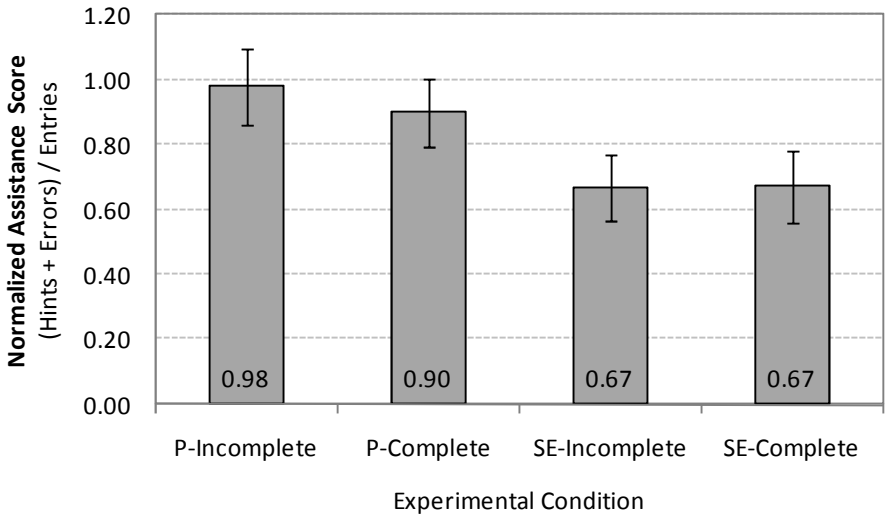


Fig. 1. Mean normalized assistance scores (± standard error), collapsing across the three training problems.

Andes help usage: Next-step help and bottom-out hints

As stated in a previous section (see Materials, p. 7), the Andes tutoring system offers several different types and levels of help. For instance, when a user is uncertain as to which step to take next, she is then able to ask Andes for a hint (i.e., next-step help). Thus, next-step help requests are a measure of the students' confusion or uncertainty during problem solving. To assess the extent to which students relied on the next-step help, we counted the total number of next-step hint requests and divided by the number of entries made per problem (i.e., the hint rate).

The pattern of results was consistent with the Generation hypothesis (see *Fig. 2*). To test the reliability of this pattern, an repeated-measure ANOVA confirmed that the SE-complete and SE-incomplete were equal to each other, but requested reliably fewer next-step hints per entry than both of the paraphrase conditions, $F(1, 73) = 8.70$, $p = .004$, $\eta_p^2 = .11$.

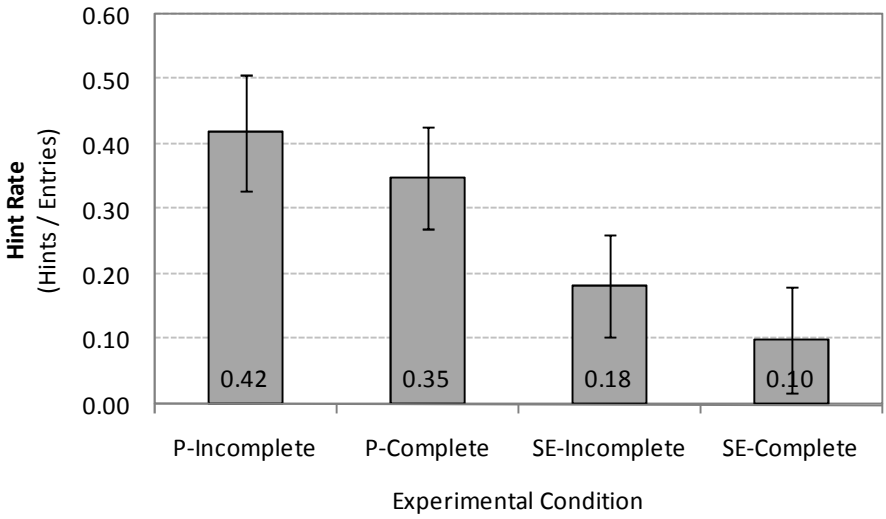


Fig. 2. Mean next-step hint rate (± standard error) across the three training problems.

Once a student has asked for help on a step, the student can then proceed to ask for hints until Andes tells the user directly what to enter. This last hint will be referred to as a *bottom-out hint*. Note that both next-step help and what's wrong help include bottom-out hints. There was a reliable main effect of study strategy on bottom-out help usage, which collapsed across both types of hints (i.e., what's wrong and next-step help). The students in the P-complete and P-incomplete conditions requested marginally more bottom-out hints than the students in the self-explain conditions, $F(1, 73) = 5.47$, $p = .02$, $\eta_p^2 = .07$ (see *Fig. 3*).

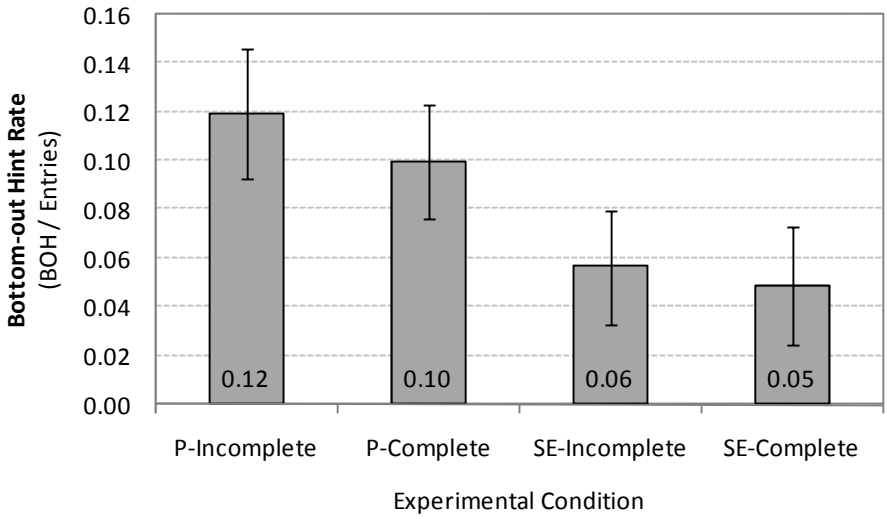


Fig. 3. The mean (±SE) frequency of the bottom-out hint rate per experimental condition.

Again, this pattern of results was more consistent with the Generation hypothesis than the Coverage hypothesis. Furthermore, the results from the next-step help and bottom-out help usage replicates a finding reported in Catrambone and Yuasa (2006). They found evidence that participants who were in the "active learning" condition (i.e., those asked to self-explain examples) asked for fewer hints than the students in the "passive learning" condition (i.e., those asked to study examples).

Robust learning

Far transfer and retention

On an average of 29 days after the completion of the experiment, the students were administered an exam that covered a subset of the material from the experiment. One exam question was similar to one of the problems used in the experiment (see "PROB2" and "Chapter Exam" from Appendix A). The exam question is considered a far-transfer assessment item because the problems from the training set were in one dimension, while the exam problem included motion in two dimensions. Thus, the chapter exam question served as both a far-transfer assessment and a retention test.

To analyze the effect of the experimental conditions on exam performance, an ANOVA was conducted on the midterm exam question, and the score was expressed as a percentage of the total possible points (20 point). Unfortunately, not every section in the experiment used Andes throughout the duration of the semester. Because Andes is known to affect learning (VanLehn et al, 2005), analyses of exam scores were restricted to the three out of the five sections (N = 63 participants) that used Andes.

Activity and problem type were entered as between-subjects variables. The results were consistent with neither the Generation (η_p^2 = .036) nor the Coverage (η_p^2 = .001) hypotheses.

However, a Fischer's least significant difference (LSD) post-hoc analyses revealed a trend with the SE-complete group (M = 90.83, SD = 9.96) demonstrating a marginally higher score on the chapter exam question than the P-complete group (M = 73.00, SD = 22.51), (LSD: p = 0.06, η_p^2 = .08).

A second measure of retention was the students' performance on an isomorphic homework problem that was completed well after the training session (see "Homework: Exam isomorph" in Appendix A). Although some students complete their homework by the date it is due, others do their homework later, typically just before an exam. This provides an opportunity for LearnLab researchers to measure the students' performance on the homework items at different points in the semester, albeit knowing that students self-selected when they would do their homework.

We analyzed the students' performance on a homework problem that was isomorphic to the chapter exam, in the sense that they shared an identical deep structure (i.e., both analyzed the motion of a charged particle moving in two dimensions – see Appendix A). The homework problem was solved after the training session. There was a statistically reliable effect and a medium to large effect size favoring the Generation hypothesis (see *Fig. 4*). The SE-incomplete and SE-complete conditions demonstrated lower normalized assistance scores than both of the paraphrase conditions, $F(1, 27)$ = 4.81, p = .04, η_p^2 = .15. This pattern of results replicates and strengthens the finding from the chapter exam.

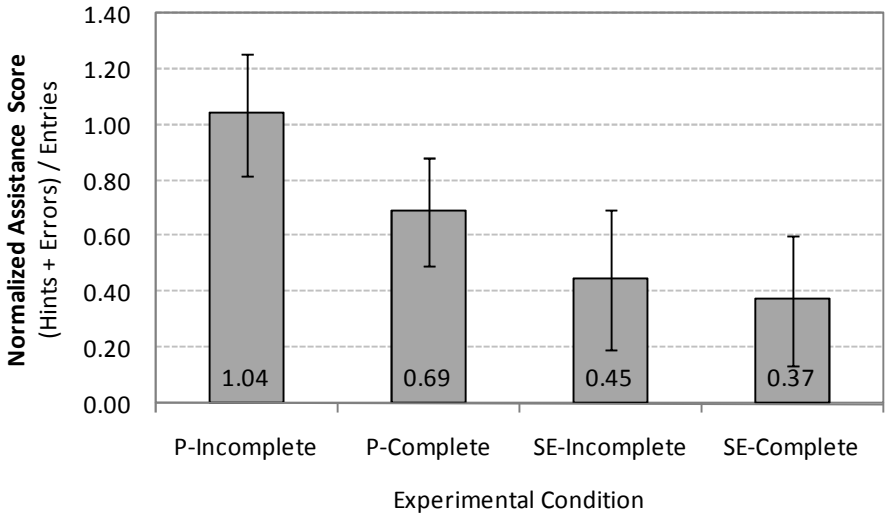


Fig. 4. Mean normalized assistance scores (± standard error) for an isomorphic homework problem.

Accelerated future learning

To assess the acceleration of future learning, the chapter on magnetism was selected because some of its concepts overlap those of the electrodynamics chapter. For example, both include a charged particle, a force, and a field (either magnetic or electric). A deep understanding of electrodynamics might accelerate students' learning of magnetism.

There were significant differences between experimental conditions on the magnetism homework problem that was most similar to the electrodynamics problems (see *Fig. 5*). The pattern of data supported the Generation hypothesis. The students in the P-complete and P-incomplete conditions demonstrated higher normalized assistance scores than the SE-complete and SE-incomplete conditions, $F(1, 46) = 3.70$, $p = .06$, $\eta_p^2 = .075$. Once again, this pattern of results was more consistent with the Generation hypothesis than the Coverage hypothesis.

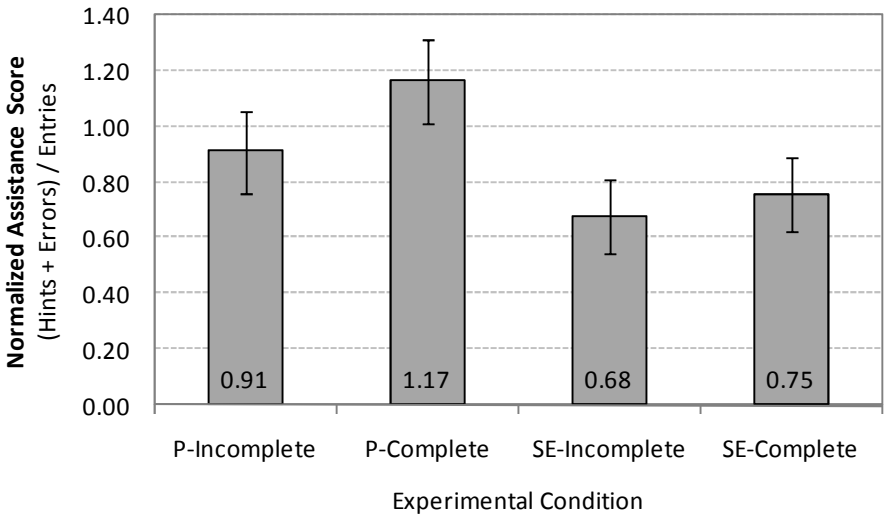


Fig. 5. Mean normalized assistance scores (± standard error) for a homework problem on magnetism.

Knowledge component analysis

To obtain a finer-grained analysis of learning over time, we focused on knowledge components that occurred in all three problems.[3] Four knowledge components met this criterion, which included: KC1: applying the definition of the electric field (**F** = *q***E**); KC2: drawing an electric-field (E-field) vector; KC3: drawing an electric-force vector; and KC4: defining the charge on a particle. The most important knowledge component was applying the definition of the electric field (KC1) because it was the main principle taught in the chapter on electric fields. However, not every student was able to complete all three problems. Therefore, the data were restricted to those who applied the knowledge components across all three problems.

To measure performance on each of these knowledge components, we used a repeated-measures ANOVA on the assistance scores for each of the four knowledge components across the three problems. Opportunity was entered as the repeated, within-subjects factor, and the contrast between

[3] The data from this experiment can be accessed from the Pittsburgh Science of Learning Center DataShop, which can be found here: https://learnlab.web.cmu.edu/datashop/

the P-complete and SE-incomplete was evaluated at each level of Opportunity. In order to simplify the analysis, we compared only these two conditions because they are the ones where the Generation hypothesis and the Content hypothesis make different predictions.

KC1. Applying the definition of the electric field

For the most important knowledge component, KC1, the assistance score decreased for all of the experimental conditions (Hotelling's trace; $F(2, 29) = 5.26$, $p = .01$). Moreover, the assistance score, as a function of opportunity, differed between conditions (see *Fig. 6*). Collapsing across opportunities, there was a reliable difference between the P-complete and SE-incomplete conditions, $F(1, 30) = 4.40$, $p = .05$, $\eta_p^2 = .13$. The difference between conditions was largely driven by a reliable difference for the second opportunity, $F(1, 30) = 14.32$, $p = .001$, $\eta_p^2 = .32$. This pattern of results replicated the normalized assistance scores findings reported earlier, which were measured at the level of the experimental session (i.e., collapsing across knowledge components *and* opportunities). The findings at this fine-grain size were also consistent with the Generation hypothesis.

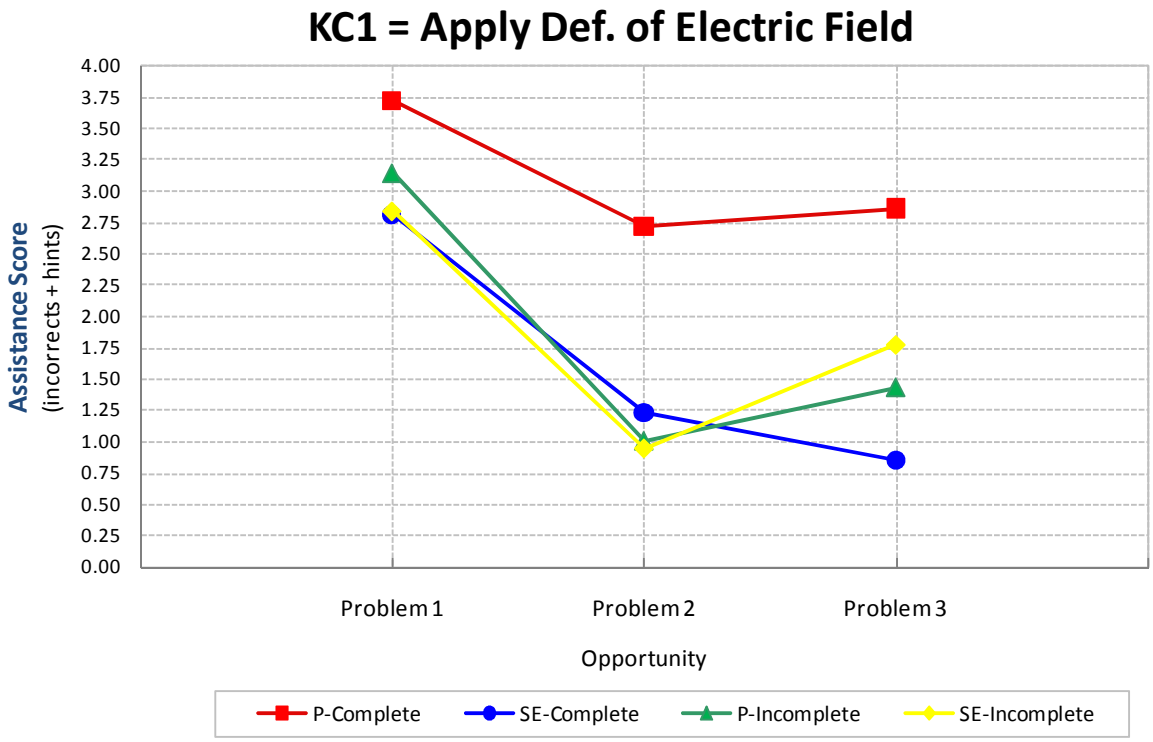


Fig. 6. Assistance score per opportunity to apply KC1, the definition of an electric field.

KC2. Drawing the electric force vector

The pattern of results for the second knowledge component (i.e., KC2: Drawing an electric-force vector) was slightly different (see *Fig. 7*). Unfortunately, there was no evidence that the assistance scores decreased over time (Hotelling's trace; $F(2, 54) = .47$, $p = .63$). However, similar to KC1, the assistance score differed between conditions when performance was collapsed across opportunities (see *Fig. 6*). There was a marginal difference between the P-complete and SE-incomplete conditions ($F(1, 55) = 3.31$, $p = .07$, $\eta_p^2 = .06$), and the difference was largely driven by a reliable difference for the first opportunity, $F(1, 55) = 4.40$, $p = .04$, $\eta_p^2 = .07$. This was also consistent with the Generation hypothesis.

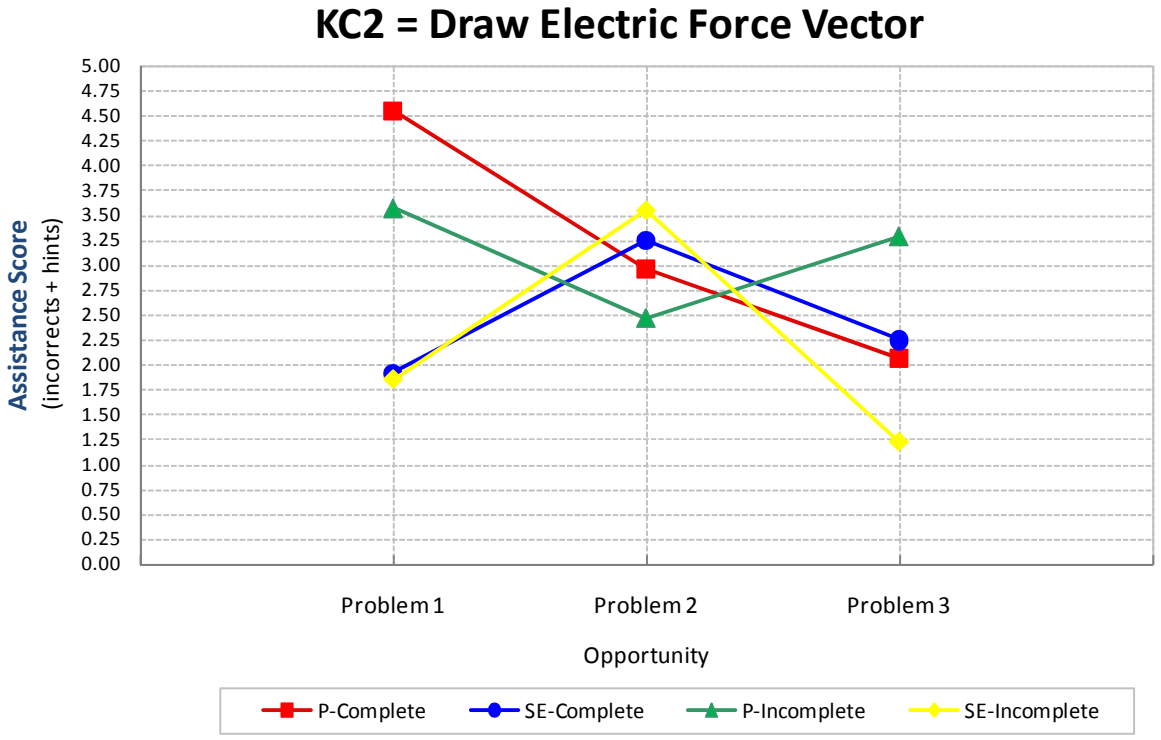


Fig. 7. Assistance score per opportunity to apply KC2: drawing the electric-force vector.

KC3. Drawing an electric-field vector

A similar main effect was found for drawing the electric-field vector. There was a strong within-subjects effect of time, with students in all conditions becoming increasingly competent over time

(Hotelling's trace; $F(2, 72) = 26.57$, $p < .001$, $\eta_p^2 = .43$; see *Fig. 8*). This is a marked contrast with KC2, drawing an electric force vector, where the learning curves were more flat. Both knowledge components involve drawing a vector, which is a complex act that requires filling out a multi-entry dialogue box. However, the different shapes of their learning curves may differ due to the fact that drawing a mechanical force vector (e.g., gravity, tension, and friction) is familiar to these students, whereas drawing a field vector is a more novel concept.

The difference between conditions was consistent with the previous two knowledge components. The assistance score differed between the P-complete and SE-incomplete conditions when performance was collapsed across opportunities ($F(1, 73) = 4.82$, $p = .03$, $\eta_p^2 = .06$; see *Fig. 6*). The difference was largely driven by a reliable difference for the first opportunity, $F(1, 73) = 7.80$, $p = .007$, $\eta_p^2 = .10$. Again, the difference between the P-complete and SE-incomplete conditions was consistent with the Generation hypothesis.

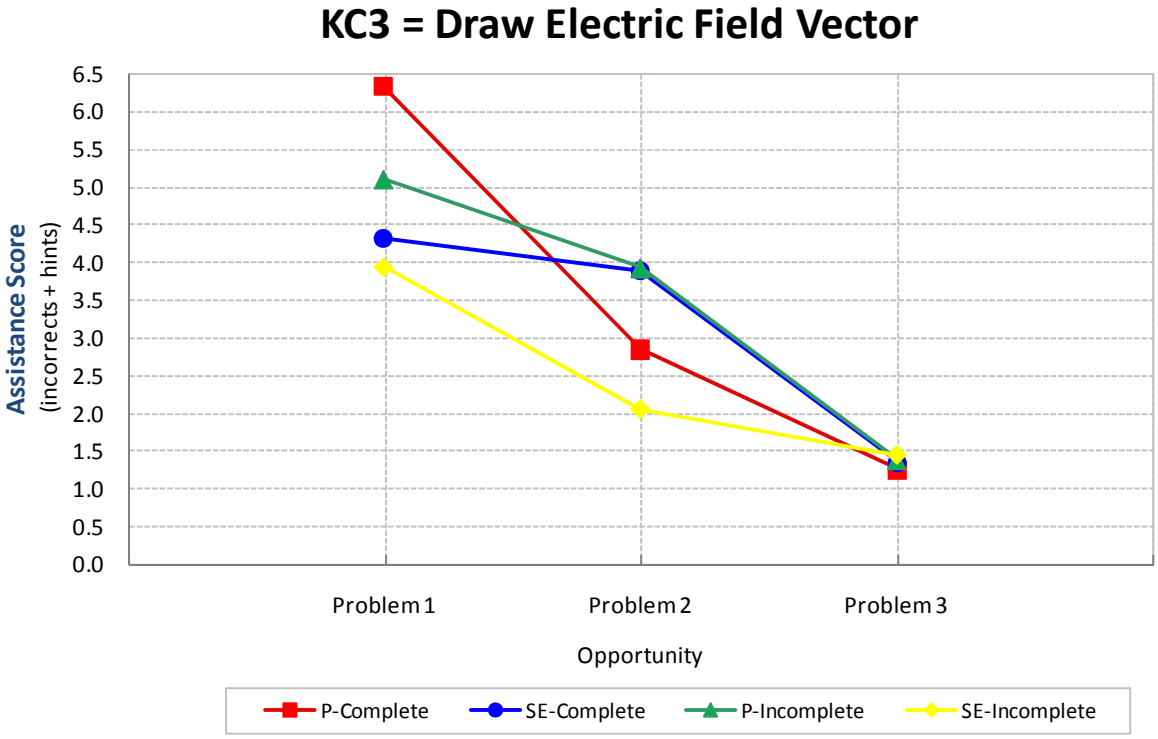


Fig. 8. Assistance score per opportunity to apply KC3, drawing an electric-field vector.

KC4. Defining the charge on a particle

Finally, for the last knowledge component, defining the charge on a particle, there were no reliable differences between conditions (all Fs < 1). The assistance scores were, however, lower for KC4 than the average assistance score for all the other KCs, $F(1, 32) = 76.59$, $p < 0.001$, $\eta_p^2 = .70$.

The reason why this knowledge component was unaffected by the experimental manipulations was because the students were committing very few errors, even on the first opportunity (M = .59, SD = .17). Defining the charge on a particle is extremely easy because all of the information is given in the problem statement. Therefore, there is little surprise why there was a large effect between knowledge components, yet no differences between experimental conditions.

Analysis of classroom verbal data

To better understand the results, an analysis of the verbal protocols was undertaken. While the students studied the worked-out examples, they were asked to either self-explain or paraphrase the content of the instructional materials. As stated previously, every student was outfitted with a close-talk, noise-canceling microphone so that their verbalizations could be recorded, transcribed, and coded. The data were coded according to the procedure outlined in Chi (1997).

Data selection

Because of the volume of data that verbal protocols generate, we reduced the data to only those students who watched all the video-based examples and focused on only the verbalizations that occurred during the example-studying phase of the experiment. That is, we did not analyze the verbalizations made while solving the problems. Restricting our analyses to these data, 994 episodes were hand-coded, which constituted approximately eight percent of the overall sample (994 / 12,749 = 7.80%).

Segmentation and coding scheme

The data were segmented according to idea units, which roughly correspond to individual sentences. Coding was conducted in three passes through the verbal data (see Fig. 9). Within studying examples, the talk was coded as either a *self-explanation* or *paraphrase* (Pass 1). Self-explanations and paraphrases were decomposed by their topic (Pass 2). One topic was the *user interface*. The Andes user interface is nontrivial, and both the videos and the students sometimes discuss how to use it. In contrast, students' talk was coded as reflecting *content* when it was directly tied to the instructional content either as a part of the student's prior knowledge, or embedded in the video-based example. Finally, self-explanations were further categorized as being either *meta-cognitive* or *justification-based* (Pass 3). The purpose for the final coding category was to explore differences between conditions that might be due to the meta-cognitive prompts used in the self-explanation condition (see Appendix C).

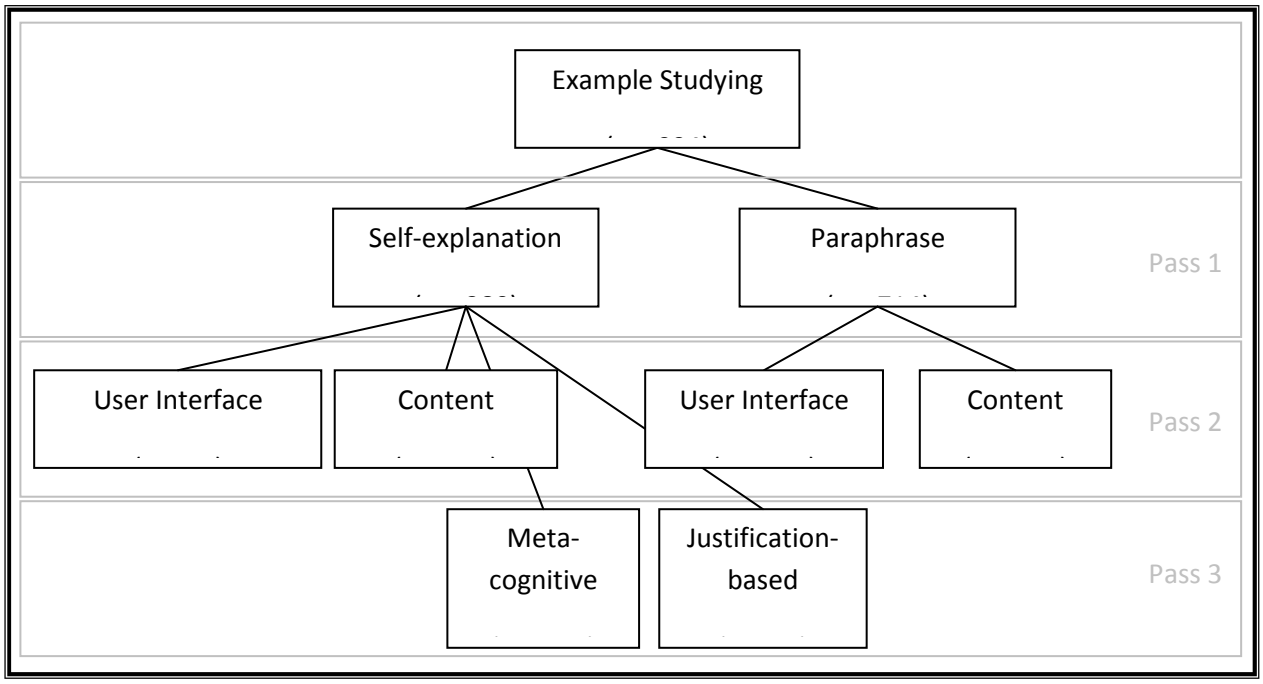


Fig. 9. The coding scheme used to categorize student utterances while studying examples.

For the first pass (Pass 1), the coding scheme used the following four criteria to distinguish self-explanations from paraphrases (see Table 2). First, if the segment's content went beyond the information presented in the current step of the worked-out problem, it was coded as self-explanation. Second, segments with meta-cognitive comments were also counted as a self-explanation. Third, if the student raised a question during the segment in response to the content of the step, the segment was coded as self-explanation. Finally, the segment was coded as a self-explanation if the student made some integrative statement with either their prior domain-relevant knowledge or previously presented material. Finally, if none of the above criteria were met, and the statement merely repeated what was in the example, then the segment was coded as a *paraphrase*.

Table 2
The definitions and examples for the third level of coding.

Code	Definition	Example
Self-explanation (user-interface)	A meta-cognitive statement or justification that goes beyond the information provided in the example that deals with the interface.	So he just put 'E' out in the middle of nowhere. That's interesting. I guess it really kind of doesn't really matter. I don't know. Huh.
Self-explanation (content)	A meta-cognitive statement or justification of the physics content.	I have no earthly idea what goes... what's going on with Andes right at this point. So, I'm still a little confused.
Paraphrase (user-interface)	A restatement of the example step that deals primarily with how to accomplish an action in the user interface.	So, I read the problem and assign a variable, which is 'P.' Click on the little dot in the left-hand corner. Drag it down, left click, a box comes up. You select particle, click 'OK.'
Paraphrase (content)	A restatement of the physics content.	'F' equals 'Q E' and solve for 'F.' Force is defined on the body... uh... the same direction that 'E' is.

Coding-scheme results

The results for the coding scheme are presented in the next three sections, corresponding to each of the three passes through the data. For each pass, the results are reported in a table that summarizes the descriptive statistics using the proportions of each code.[4] On the other hand, to test for differences

[4] For example, the 994 segments that comprise the data of Level 1 (see *Table 3*, bottom row) were first partitioned according to condition (e.g., there were 213 segments in the P-complete condition; see *Table 3*, top row) and then the counts for each code in that condition were divided by the total number of segments in that condition.

between experimental conditions, the amount of total talk was controlled statistically using an ANCOVA. The reason for reporting the results as proportions in tabular formats is because they are easier to interpret than the estimated marginal means that are generated by the ANCOVA.

Pass 1: Paraphrasing vs. Self-explaining. In an effort to detect if the experimental manipulation had its intended effects, the average number of *paraphrase* and *self-explanation* segments were contrasted for each of the four conditions. A post-hoc analysis revealed that the SE-incomplete condition produced more self-explanation segments than both the P-incomplete (LSD: $p = .05$) and P-complete (LSD: $p < .03$) conditions (see Table 3). This suggests that the SE vs. P manipulation was successful.

Moreover, the SE-complete condition produced marginally more paraphrases than the SE-incomplete condition, $F(1, 73) = 3.38$, $p = 0.07$. This suggests that students in both SE conditions articulated justifications of steps as intended, but because those justifications were mentioned in the complete examples, such articulations were counted as paraphrases for the SE-complete students and as self-explanations for the SE-incomplete students. This is consistent with the intended effect of the complete vs. incomplete manipulation. In short, the Pass 1 coding suggests that both experimental manipulations were operating as intended.

Table 3
The mean number of paraphrases and self-explanation episodes for each experimental condition.

	n	**Paraphrase**	**Self-explanation**
P-Incomplete	23	165 / 213 = 0.77_a	48 / 213 = 0.23_a
P-Complete	26	244 / 321 = 0.76_a	77 / 321 = 0.24_a
SE-Incomplete	28	125 / 219 = 0.57_b	94 / 219 = 0.43_b
SE-Complete	27	180 / 241 = 0.75	61 / 241 = 0.25
Total	104	714 / 994 = 0.72	280 / 994 = 0.28

Note. Means in columns with different subscripts differ reliably at $p < .05$ by Fisher's least significant difference test.

Pass 2: User-interface vs. Content. For the second pass through the verbal data, paraphrases and self-explanations were partitioned into *user-interface* vs. *content* statements. There were two reliable effects. First, a post-hoc analysis revealed that the SE-incomplete condition produced more user-interface self-explanation segments than both the P-incomplete (LSD: $p < .04$) and P-complete (LSD: $p = .003$) conditions (see Table 4). If details of the user interface are easily overlooked, which seems likely, then asking students to self-explain may bring the user-interface details to the students' attention, whereas asking students to paraphrase may not increase the salience of such details.
Second, across all conditions, there were more *content* self-explanations than *user interface* self-explanations (250 vs. 30). This suggests that, although the Andes user interface is nontrivial, the physics content was even more challenging. This supports our assumption that students are mostly learning physics in this experiment, and that user-interface learning was a minor source of variance.

Table 4
The mean number of user interface and content paraphrases and self-explanation episodes for each experimental condition.

		SE	**SE**	**Paraphrase**	**Paraphrase**
	n	**User Interface**	**Content**	**User Interface**	**Content**
P-Incomplete	23	$5 / 48 = 0.10_a$	43 / 48 = 0.90	66 / 165 = 0.40	99 / 165 = 0.60
P-Complete	26	$4 / 77 = 0.05_a$	73 / 77 = 0.95	69 / 244 = 0.28	175 / 244 = 0.72
SE-Incomplete	28	$15 / 94 = 0.16_b$	79 / 94 = 0.84	50 / 125 = 0.40	75 / 125 = 0.60
SE-Complete	27	6 / 61 = 0.10	55 / 61 = 0.90	75 / 180 = 0.42	105 / 180 = 0.58
Total	104	30 / 280 = 0.11	250 / 280 = 0.89	260 / 714 = 0.36	454 / 714 = 0.64

Note. Means with different subscripts differ reliably at $p < .05$ by Fisher's least significant difference test.

Pass 3: Meta-cognitive vs. Justification-based. For the final pass through the data, only the self-explanations segments were decomposed and coded as being either meta-cognitive or justification-based (see Table 5). There were a total of 86 justification episodes and 194 meta-cognitive episodes across the four conditions. A post-hoc analysis revealed that the SE-incomplete condition produced more justification-based self-explanation episodes than the P-complete (LSD: $p = .02$) and marginally more than the P-incomplete (LSD: $p = .11$) conditions. This suggests that the students in the SE-incomplete were in fact actively trying to fill in the missing information from the examples, as we intended that they should. The SE-incomplete group also produced more justification-based self-explanations segments than the SE-complete group (LSD: $p = .02$). In short, these results suggest again that the students' verbalizations are consistent with the activities that they were requested to perform.

Table 5
The mean meta-cognitive and justification-based self-explanation episodes that focused on user-interface elements.

	n	**SE-Justification**	**SE-Meta-cognitive**
P-Incomplete	23	17 / 86 = .198	31 / 194 = .160
P-Complete	26	23 / 86 = .267	54 / 194 = .278
SE-Incomplete	28	34 / 86 = .395	60 / 194 = .309
SE-Complete	27	12 / 86 = .140	49 / 194 = .253
Total	104	86 / 86 = 1.00	194 / 194 = 1.00

None of the pair-wise contrasts were statistically reliable for the meta-cognitive self-explanations. However, the overall results from the third pass (the "Total" line of Table 5) suggest that the focus of most self-explanations was targeted more toward meta-cognitive reflections of one's current state of understanding and comprehension (194 episodes) than filling in the missing justifications (86 episodes). This makes sense because mentioning justifications was suppressed in one condition (P-incomplete) and counted as paraphrasing in two conditions (SE-complete and P-complete).

Coding scheme summary

The pattern of results for the verbal protocols suggests a few conclusions. First, it may have been more challenging to self-explain than paraphrase the material. If we collapse across experimental conditions, we find that the average number of paraphrase segments *(M* = 6.87, *SD* = 8.50) is higher than the average number of self-explanation segments (*M* = 2.69, *SD* = 5.19). It may be that the instructions to self-explain were too difficult or that paraphrasing is the first step toward producing high quality self-explanations.

Second, when students successfully self-explained, the focus of their statements was targeted toward physics content. This was encouraging because some students were new to the Andes interface. Given that most self-explanation segments were about the content may be one of the reasons why they performed well on the problem-solving tasks (e.g., the training set, homework problems, and the exam question).

DISCUSSION

Self-explanation can be viewed both as a process (i.e., the activity of generation) and as an outcome (i.e., new content). Although many believe that active learning strategies, such as self-explanation, necessarily increase learning from worked-out examples, a confound is introduced for self-explanation in that it makes students aware of content that is not contained in the original example. In order to tease apart these two explanations, we defined two hypotheses. The Generation hypothesis claims that generating content makes is better understood and easier to recall than comprehending it. The Coverage hypothesis claims that generation vs. comprehension makes no difference; the benefits of self-explanation are due to increasing the content experienced by self-explainers compared to non-self-explainers.

To test these hypotheses, we had students to engage in two different learning strategies. The first was self-explaining, which is a classic active learning strategy. The second was paraphrasing, which might also be considered an active learning strategy because the student is at least attempting to put into his or her own words the content of the example. Yet paraphrasing, by definition, does not produce any new content.

Orthogonal to the learning strategy, we manipulated the content of what was studied. The examples that the students studied were either completely or incompletely justified. The completely justified examples articulated the reasons for taking each problem-solving step. The incompletely justified examples, on the other hand, omitted the necessary reasons for the step and merely explained the mechanics of taking the step in the learning environment (i.e., Andes).

Ideally, students who paraphrased complete examples should experience the same set of justifications as the students who self-explained incomplete examples. Thus, these two groups would become aware of the same content, but one would have generated it while the other would have comprehended it. If the Coverage hypothesis holds, then these two groups should have the same learning gains. If the Generation hypothesis holds, then the students who self-explain incomplete examples should learn more than the students who paraphrase complete examples.

The results were remarkably consistent across two grain sizes of analysis (i.e., problem-set vs. knowledge-component), across situations (i.e., experimental session, homework, and exams), across assessments (i.e., near transfer, far transfer, and accelerated future learning topics), and across dependent measures (i.e., normalized assistance scores, errors, and help usage). The results usually supported the Generation hypothesis. Students in the experimental condition who were instructed and prompted to self-explain while studying the examples demonstrated lower errors and requested fewer hints than the students who merely paraphrased the information. This suggests that generation is an important feature of self-explaining. In particular, students who paraphrased completely justified examples did *not* learn more than students who self-explained incompletely justified examples, which would occur if the Coverage hypothesis were true and so were certain plausible assumptions about the relative frequency of paraphrasing and self-explaining (see the Introduction).

Limitations of the study

There are a few limitations to the current study. First, students were not required to watch every frame of every video, and only about a third of them did so. Perhaps this occurred because the new content tended to occur at the beginning of the set of videos, so some students stopped watching when the content became familiar. Another reason might be due to the students' preferred method of learning.

For instance, some students may prefer to try a problem themselves first. That is, they prefer to learn on their own and from their own mistakes. Alternatively, other students prefer to attempt to solve a problem only after they have observed a worked example. The materials were structured such that the examples came after the problems were solved. This may have violated their traditional study methods.

The prompts given to the students were meta-cognitive in nature, and did not explicitly remind students to produce justifications for steps. The prompts were selected from Chi et al. (1994) because they successfully induced self-explanation of a biology text. The prompts mainly focused on the reflection of knowledge, instead of the production of justifications. In a follow-up study, we plan to contrast meta-cognitive prompts with justification prompts in order to investigate if they produce different behaviors.

Finally, one qualification should be made with respect to the completeness manipulation. The students could flesh out the incompletely justified examples by using the help system during problem solving. That is, Andes presents the hints in a graded fashion. The first hint is supposed to serve as a reminder for which step to apply. If the student asks for a second hint on the same step, it provides a reason for taking the step. The last hint in the sequence directly tells the student which action to take. Therefore, the second-level of help may have washed out our completeness manipulation. This possibility will be the focus of future analyses.

Conclusions

The results of the current experiment were remarkably consistent across problems, time, and granularity. The evidence favors the Generation hypothesis, which states that the robust learning that results from self-explanation is largely due to the active generation of the missing information in the form of justifications for the application of each problem-solving step. These results suggest that new learning technologies should be designed to encourage students to actively generate missing information.

ACKNOWLEDGEMENTS

Funding for this research is provided by the National Science Foundation, Grant Number SBE-0354420 to the Pittsburgh Science of Learning Center (PSLC, http://www.learnlab.org). Portions of the results were presented at the 13th International Conference on Artificial Intelligence in Education Conference. The authors would like to thank Anders Weinstein and Dale Walters for their help integrating the screen-capturing software into Andes; Michael A. Ringenberg and Min Chi for their expertise in setting up the databases; Anders Weinstein, Adaeze Nwaigwe, Alida Skogsholm, and Benjamin Billings for their help in migrating Andes log files into the DataShop, Robert Shelby and Brett van de Sande for their assistance in developing the materials; and Donald J. Treacy, John J. Fontanella, and Mary C. Wintersgill for graciously allowing us to collect data in their classroom.

REFERENCES

Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? In G. Gauthier, C. Frasson & K. VanLehn (Eds.), *Proceedings of the 5th International Conference on Intelligent Tutoring Systems, ITS 2000* (pp. 292-303). Berlin: Springer Verlag.

Aleven, V. A. W. M. M., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explain with a computer-based Cognitive Tutor. *Cognitive Science, 26*, 147-179.

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology, 95*(4), 774-783.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology, 20*(4), 493-523.

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology 98*(1), 182-197.

Catrambone, R., & Yuasa, M. (2006). Acquisition of procedures: The effects of example elaborations and active learning exercises. *Learning and Instruction, 16*(2), 139-153.

Chi, M. T. H. (1997). Quantifying qualitative analysis of verbal data: A practical guide. *The Journal of the Learning Sciences, 6*(3), 271-315.

Chi, M. T. H., & Bassok, M. (1989). Learning from examples via self-explanations. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 251-282). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Chi, M. T. H., Bassok, M., Lewis, M. W., 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. H., DeLeeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439-477.

Chi, M. T. H., & VanLehn, K. A. (1991). The content of physics self-explanations. *The Journal of the Learning Sciences, 1*(1), 69-105.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2 ed.). Hillsdale, NJ: Lawrence Earlbaum Associates, Inc.

Conati, C., & VanLehn, K. (2000). Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education, 11*, 398-415.

deWinstanley, P. A. (1995). A generation effect can be found during naturalistic learning. *Psychological Bulletin & Review, 2*(4), 538-541.

deWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition, 32*(6), 945-955.

Gertner, A., & VanLehn, K. (2000). Andes: A coached problem solving environment for physics. In Gauthier, Frasson & K. VanLehn (Eds.), *Intelligent Tutoring Systems: 5th International Conference* (pp. 133-142). Berlin: Springer.

Gilabert, R., Martinez, G., & Vidal-Abarca, E. (2005). Some good texts are always better: Text revision to foster inferences of readers with high and low prior background knowledge. *Learning and Instruction, 15*(1), 45-68.

Hausmann, R. G. M., & Chi, M. T. H. (2002). Can a computer interface support self-explaining? *Cognitive Technology, 7*(1), 4-14.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution *Journal of Verbal Learning and Verbal Behavior, 17*(6), 649-667.

Lovett, M. C. (1992). Learning by problem solving versus by examples: The benefits of generating and receiving information *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 956-961). Hillsdale, NJ: Erlbaum.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55*(1), 61-62.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38*(1), 1-30.

McNamara, D. S., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1-43.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavioral Research Methods, Instruments, and Computers, 36*, 222-233.

Moreno, R. (2006). When worked examples don't work: Is cognitive load theory at an Impasse? *Learning and Instruction, 16*(2), 170-181.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1-4.

Peynircioglu, Z. F., & Mungan, E. (1993). Familiarity, relative distinctiveness, and the generation effect. *Memory & Cognition, 21*(3), 367-374.

Pirolli, P., & Bielaczyc, K. (1989). Empirical analyses of self-explanation and transfer in learning to program. In C. M. Olson & E. E. Smith (Eds.), *Proceedings of the 11th Annual Conference of the Cognitive Science Society* (pp. 450-457). Hillsdale, NJ: Erlbaum Associates, Inc.

Reimann, P., & Neubert, C. (2000). The role of self-explanation in learning to use a spreadsheet through examples. *Journal of Computer Assisted Learning, 16*, 316-325.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21*(1), 1-29.

Renkl, A. (2002). Worked-out examples: Instructional explanations support learning by self-explanations. *Learning and Instruction, 12*(5), 529–556.

Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational Psychologist, 38*(1), 15-22.

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary Educational Psychology, 23*(1), 90-108.

Roy, M., & Chi, M. T. H. (2005). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 271-286). Cambridge: Cambridge University Press.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592-604.

Stark, R. (1999). *Lernen mit Lösungsbeispielen. Einfluß unvollständiger Lösungsbeispiele auf Beispielelaboration, Motivation und Lernerfolg* Bern, Switzerland: Huber.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*(1), 59-89.

Trafton, J. G., & Reiser, B. J. (1993). The contributions of studying examples and solving problems to skill acquisition *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 1017-1022). Hillsdale, NJ: Erlbaum.

VanLehn, K. (1996). Cognitive skill acquisition. In J. T. Spence, J. M. Darly & D. J. Foss (Eds.), *Annual Review of Psychology* (pp. 513-539). Palo Alto, Ca: Annual Reviews, Inc.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*, 227-265.

VanLehn, K., & Jones, R. M. (1993). Learning by explaining examples to oneself: A computational model. In S. Chipman & A. L. Meyrowitz (Eds.), *Foundations of knowledge acquisition: Cognitive models of complex learning* (pp. 25- 82). Boston: Kluwer.

VanLehn, K., Jones, R. M., & Chi, M. T. H. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences, 2*(1), 1-59.

VanLehn, K., Lynch, C., Schultz, K., Shapiro, J. A., Shelby, R., Taylor, L., et al. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education, 15*(3), 147-204.

Zhu, X., & Simon, H. A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction, 4*(3), 137-166.

APPENDIX A

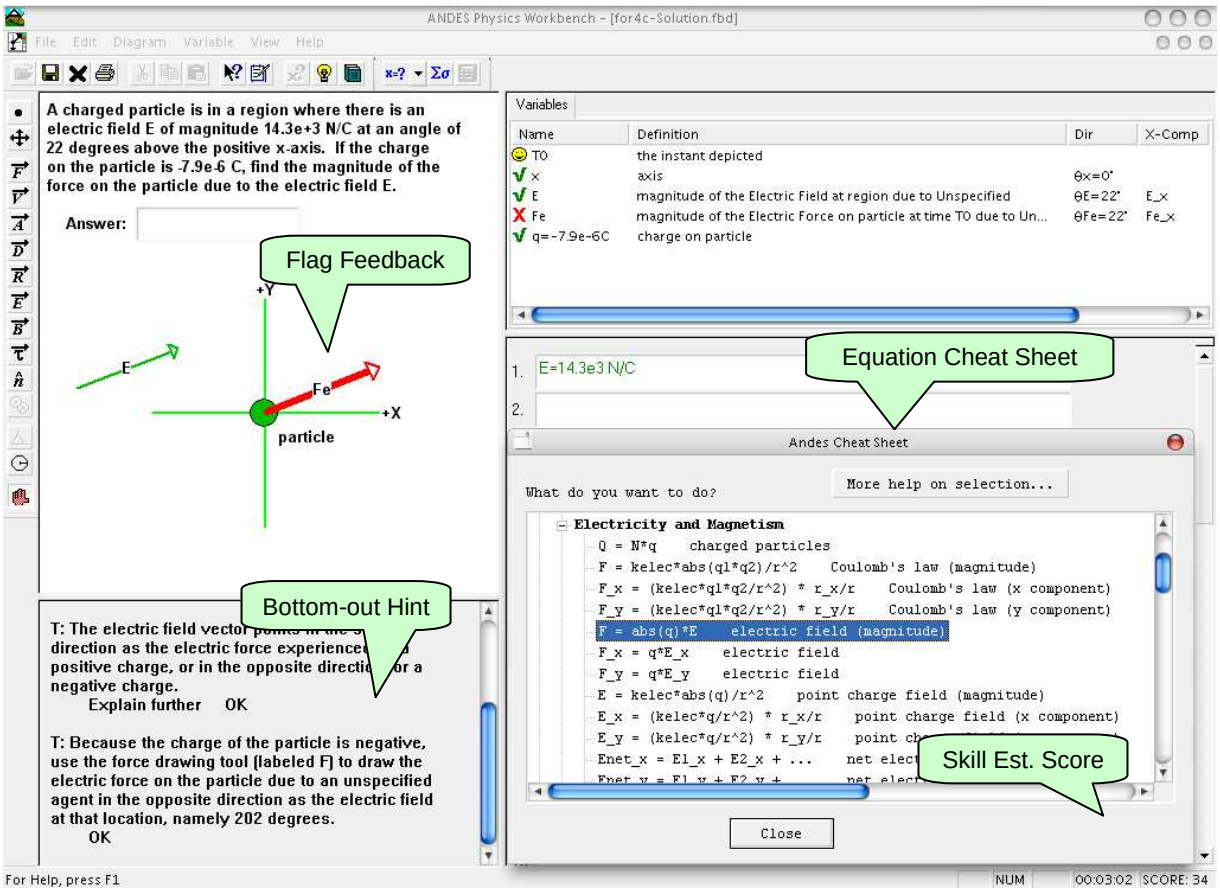
Electrodynamics problem statements used for problem solving, example studying, homework and an in-class chapter exam.

Problem	Activity	Problem Statement	Principle
Practice	SOLVE	A charged particle is in a region where there is an electric field E of magnitude 17.8 V/m at an angle of 37 degrees above the positive x-axis. If the charge on the particle is 3.4 C, find the magnitude of the force on the particle P due to the electric field E.	$\mathbf{F_e} = q\mathbf{E}$
EX1	STUDY	A charged particle is in a region where there is an electric field E of magnitude 14.3 V/m at an angle of 22 degrees above the positive x-axis. If the charge on the particle is -7.9 C, find the magnitude of the force on the particle P due to the electric field E.	$\mathbf{F_e} = q\mathbf{E}$
PROB1	SOLVE	An electron (qe = -1.60e-19 C; me = 9.11e-31 kg) is in a region where there is a uniform electric field E. The force on the electron due to the electric field exactly cancels its weight near the Earth's surface. If the y-component of the net force on the particle near Earth due to the electric field and gravity is zero, what is the magnitude (and show the direction) of the electric field E?	$\mathbf{F_e} = q\mathbf{E}$ $\mathbf{F_w} = m\mathbf{g}$ $\mathbf{F_e} + \mathbf{F_w} = 0$
EX2	STUDY	A charged particle (q = 52.0 mC) is in a region where there is a uniform electric field E of magnitude 120 N/C at an angle of 90 degrees above the positive x-axis. If the y-component of the net force on the particle near Earth due to the electric field and gravity is zero, what is the mass of the particle?	$\mathbf{F_e} = q\mathbf{E}$ $\mathbf{F_w} = m\mathbf{g}$ $\mathbf{F_e} + \mathbf{F_w} = 0$
PROB2	SOLVE	An electron (qe = -1.60e-19 C; me = 9.11e-31 kg) is in a region, between two parallel charged plates, that produce a uniform electric field E of magnitude 2.0e+4 N/C. The separation between the plates is 1.5 cm. The electron undergoes a constant acceleration from rest near the negative plate and passes through a tiny hole in the positive plate (see Figure below). Find the velocity of the electron as it leaves the hole. In this problem, gravity can be ignored.	$\mathbf{F_e} = q\mathbf{E}$ $\mathbf{F_e} = m\mathbf{a}$ $\mathbf{v}_1^2 = \mathbf{v}_0^2 + 2\mathbf{ad}$

EX3	STUDY	A proton (qp = 1.6e-19 C; mp = 1.7e-27 kg) is in a region where there is a uniform electric field E of magnitude 320 N/C, directed along the positive x-axis. The proton accelerates from rest and reaches a speed of 1.20e+5 m/s. How long does it take the proton to reach this speed? In this problem, gravity can be ignored.	$\mathbf{F_e} = q\mathbf{E}$ $\mathbf{F_e} = m\mathbf{a}$ $\mathbf{v}_1^2 = \mathbf{v}_0^2 + 2\mathbf{ad}$
PROB3	SOLVE	An electron (qe = -1.60e-19 C; me = 9.11e-31 kg) is in a region where there is a uniform electric field E of magnitude 4.0e-12 N/C, directed along the negative y-axis. The electron is moving in the positive y-direction at an initial velocity of 4.3 m/s. How far will the electron travel before it comes to rest? In this problem, please include gravity.	$\mathbf{F_e} = q\mathbf{E}$ $\mathbf{F_w} = m\mathbf{g}$ $\mathbf{F_e} + \mathbf{F_w} = m\mathbf{g}$ $\mathbf{v}_1^2 = \mathbf{v}_0^2 + 2\mathbf{ad}$
Chapter Exam Question	SOLVE	An electron enters a region of space where the vector electric field is E = (2i + 3j)x 10^4 N/C. If the initial velocity of the electron when it entered the field was V0 = 4x10^6i m/s what is its velocity 6E-6s after entering the field?	$\mathbf{F_e} = q\mathbf{E}$ $\mathbf{F_e} = m\mathbf{a}$ $\mathbf{v}_1 = \mathbf{v}_0 + \mathbf{a}t$
Homework: Exam isomorph	SOLVE	A fully ionized alpha particle with a charge of 3.2 E-19 C and a mass of 6.64 E-27 kg is initially moving in the x direction with a velocity of 1.5 E3 m/s. The particle enters a region where there is a uniform electric field in the positive y direction with a magnitude of 2.5 E2 N/C. What will be the velocity of the particle after it has moved 0.5 m in the x direction?	$\mathbf{F_e} = q\mathbf{E}$ $\mathbf{F_e} = m\mathbf{a}$ $\mathbf{v}_1 = \mathbf{v}_0 + \mathbf{a}t$
Homework: Magnetism	SOLVE	A particle with a charge of -3.2e-6 C enters a region with a velocity given by v = (56 m/s) i. The uniform magnetic field in the region is given by B = -(0.15 T) k, and there is a uniform electric field pointing downward in the vertical direction. If the particle is to keep moving in a straight line, what must be the magnitude of the electric field?	$\mathbf{F_B} = \mathbf{vqB}$ $\mathbf{F_e} + \mathbf{F_B} = 0$ $\mathbf{F_e} = q\mathbf{E}$

APPENDIX B

A screenshot of the Andes Physics Tutor



APPENDIX C

Self-explanation and paraphrase instructions and examples

Self-explanation Instructions

In this experiment, we would like you to study a few worked-out examples. The examples are presented one step at a time so that you will have time to really think about what information each step provides and how this relates to what you've already seen.

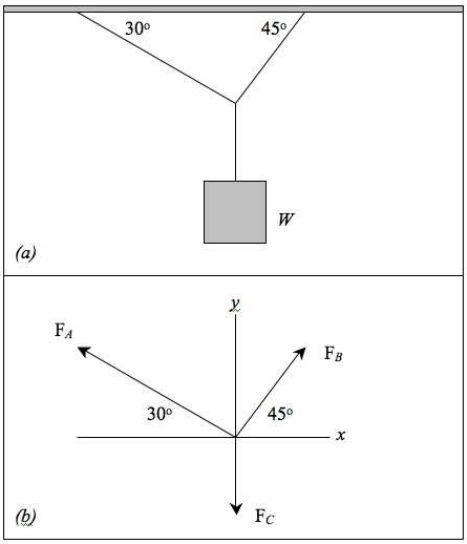
We would like you to watch and listen to each step and then explain what it means to you. That is:

- What new information does each step provide for you?
- How does it relate to what you've already seen?
- Does it give you a new insight into your understanding of how to solve the problems?
- Does it raise a question in your mind?

Tell us whatever is going through your mind - even if it seems unimportant.

Self-explanation Example

In this case, a hypothetical student is studying how to solve a statics problem. The example reads: "Consider the knot at the junction of the three strings to be 'the body'" and is accompanied by the following diagram:



At first, this confuses the student because she initially thought that the block should be the body. She says:

"I thought the block would be the body. The weight force is acting down and the strings are pulling up. Oh I see, *the sum of the forces is zero at the knot*, that's why it should be the body."

The student in this example provides a reason or **explanation** for choosing the knot. This is the essence of self-explaining.

Paraphrase Instructions

In this experiment, we would like you to study a few worked-out examples. The examples are presented one step at a time so that you will have time to really think about what information each step provides and how this relates to what you've already seen.

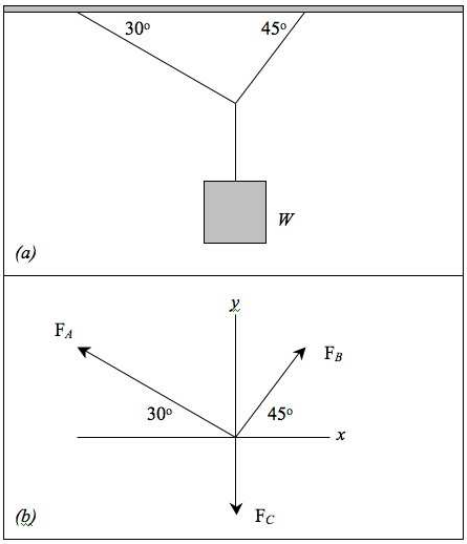
We would like you to watch and listen to each step and then paraphrase what it means in your own words. That is:

- How can I put this in my own words?
- How can I say this in a different way?
- In other words:_____.
- What's another way that I could put this?

Tell us whatever is going through your mind - even if it seems unimportant.

Paraphrase Example

In this case, a hypothetical student is studying how to solve a statics problem. The example reads: "Consider the knot at the junction of the three strings to be 'the body'" and is accompanied by the following diagram:



The student tries to restate that the knot should be the body. She says:

"The knot would be the body. It is in the middle of the three strings."

The student in this example restates the reasoning for choosing the knot. This is the essence of paraphrasing.

APPENDIX D

Completely-justified and Incompletely-justified Examples

The completely justifications are in bold, which were omitted from the incompletely justified examples. This appendix only includes the first (of three) examples that were presented to the students.

This is example number one.

We are going to start by reading the problem statement:

A charged particle is in a region where there is an electric field E of magnitude 14.3 V/m at an angle of 22 degrees above the positive x-axis. If the charge on the particle is -7.9 C, find the magnitude of the force on the particle P due to the electric field E.

The first step in solving this problem is to choose a body. In this case the body of interest is the charged particle. To draw the body, we select the Body Tool from the Diagram Toolbar, place the cursor in a convenient position in the Diagram Window and left-click. This brings up a dialog box in which we select our body. **We choose the charged particle because both of the forces mentioned in the problem affect the particle.** Next, we assign it a name. In this case, we will use "p" to refer to the particle.

[PROMPT]

The second step is to draw a coordinate system. To draw the axis, we select the Axis Tool from the Diagram Toolbar and we click and drag. This brings up a dialog box in which we select the Orientation of our axis. In this case, we will use a zero degree orientation. **We choose an un-rotated axis because all of the motion is going to be straight up and down.**

[PROMPT]

For the next few steps, we are going to enter the scalar and then the vector givens mentioned in the problem statement. The only scalar given is the charge on the particle. To assign a charge on the particle, we right-click in the Variable Window and select "Add new variable". This brings up a submenu with several options. **Because we are defining the charge on the particle**, we select "charge." This brings up a dialog box. Here we select the body on which the charge is located. In this case, the charge is on the particle, and we use the default name, "q" as the name for the charge. We also know the value because it is mentioned in the problem statement, so we enter that in the next field [paste: -7.9C].

[PROMPT]

The next given is the magnitude and direction of the electric field. An electric field is a vector, so we need to indicate the direction in the free-body diagram. To draw the electric field, select the Electric Field tool from the vector tools in Diagram toolbar, and we place the cursor in the diagram window, click and drag. This brings up a dialog box. The electric field is in a region and it is due to some unspecified source. **It is unspecified because the source of the electric field is not mentioned in the problem statement.** The field is acting at only one time point, and the problem statement says that the orientation is 22 degrees in the plane. We will use the default name "E" to refer to the electric field. We also know the magnitude, so we can assigned it by typing it in the Equation Window [paste: E=14.3N/C]

[PROMPT]

Now that all the given information has been entered, we need to apply our knowledge of physics to solve the problem.

One way to start is to ask ourselves, "What quantity is the problem seeking?" In this case, the answer is the magnitude of the force on the particle due to the electric field.

We know that there is an electric field. **If there is an electric field, and there is a charged particle located in that region, then we can infer that there is an electric force on the particle. The direction of the electric force is in the opposite direction as the electric field because the charge on the particle is negative**.

We use the Force tool from the vector tool bar to draw the electric force. This brings up a dialog box. The force is on the particle and it is due to some unspecified source. We do know, however, that the type of force is electric, so we choose “electric” from the pull-down menu. For the orientation, we need to add 180 degrees to 22 degrees to get a force that is in a direction that is opposite of the direction of the electric field. Therefore we put 202 degrees. Finally, we use “Fe” to designate this as an electric force.

[PROMPT]

Now that the direction of the electric force has been indicated, we can work on finding the magnitude. **We must choose a principle that relates the magnitude of the electric force to the strength of the electric field, and the charge on the particle. The definition of an electric field is only equation that relates these three variables.** We write this equation, in the equation window.

[PROMPT]

Now that all of the equations have been entered, we can solve for the unknown. To do so, right click anywhere in the equation window and select the sought for quantity. In this case, we’re solving for the electric force and Andes gives us a value so we can cut and paste the answer in the answer field. When the answer turns green, we know that we have the answer correct.

That is the conclusion of Example 1
