5. Examplify: Enhancing Worked Examples for Better Learning

5.1 Introduction

In the contextual design study (Chapter 3) I observed that many students grappled with how to study most effectively. Both students and faculty affirmed that many students have poor study skills. Students wanted to study more efficiently for exams, by having a strong sense of that they know and what they need to spend more time on. I identified this as an another opportunity (after Nudge in Chapter 4) for which to design a new software system that tries to address this problem by operationalizing education theories and to provide data to inform such theories and their future applications (i.e., an operant probe system, defined in Chapter 2).

Through ideation of solutions, filtering by engagement with theory, and then by potential for uptake as determined by interviews with students and teachers, I settled on a rough description (and name) for the system: Examplify supports students in studying for exams by scaffolding the metacognitive skills needed to learn most effectively from example problems.

This chapter describes the iterative development of the Examplify system and a semester-long study in a large chemistry course to evaluate its efficacy as an operant probe. I evaluate the system through a pseudo-experimental comparison of course sections and a randomized controlled trial of two variants of the tool.

5.2 Background Theory

Humans generally overestimate their level of understanding, which hinders redress of their deficits ("Assessing our own competence: Heuristics and illusions.," 1999). For example, they are overconfident about their memories and are underestimating the amount they will learn by studying (Kornell, 2009). When students do study it is often by transcribing their notes until they don't feel confused, rather than testing themselves (Karpicke & Blunt, 2011). This overconfidence of understanding is more severe among less advanced learners (Falchikov & Boud, 1989)**,** who need most to be improve (Falchikov & Goldfinch, 2000). This work draws on three methods to improve learning through directing students' attention to their misconceptions: self-explanation, testing, and worked examples. The operation of Examplify differs from the procedures used in these studies but they do bear on its design and the hypotheses of its effects.

Testing that requires recall has both mediated effects (such as revealing a need for further study) and direct effects on learning (Pashler et al., 2007; Roediger & Karpicke, 2006a). Studies of the *testing effect* generally test paired associate learning and I found no studies testing complex cognitive skills. On paired associate learning tasks, the effect is greater the more difficult or intricate the test (e.g., Bjork, 1999; Karpicke & Roediger, 2007a). The testing effect has also been verified on a test of reading comprehension and retention but without demonstrating benefits from more demanding recall (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008). In a lab-style experiment, students studied prose passages and then restudied or took an open- or closed-book test. Taking either kind of test, with feedback, enhanced long-term retention relative to conditions in which subjects restudied material or took a test without feedback. On the initial test, open-book testing led to the best performance, but on a delayed assessment both types of testing produced equivalent retention. Bearing on the implementation of testing strategies, the students wrongly predicted they would recall more after repeated studying than through testing (Agarwal et al., 2008). This discrepancy between perceived and actual learning may result because students recall the feeling of knowing after they have restudied but feel less competent after testing. Students generally overestimate how quickly they have understood, for example, when people are allowed to decide when to stop studying, their memory performance can be worse than when the experimenter controls their timing (Kornell & Bjork, 2007; Metcalfe & Kornell, 2007) and they do not realize when extra study time will help (Koriat, 1997). Interestingly, a meta-analysis of testing effect studies noted that students who were tested frequently rated their classes more favorably in semester-end course ratings than students who were tested less frequently (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991). This is perhaps a selection effect due to selective reporting or collection of course ratings, but it does offer some hope for increasing the application of testing in classes.

The study of *worked examples* is another effective learning activity that breaks the illusion of understanding (Pashler et al., 2007; Renkl, 2002). "A worked example is a step-by-step demonstration of how to perform a task or how to solve a problem" (Clark, Nguyen, Sweller, 2006, p. 190) and studying worked examples is an effective instructional strategy to teach complex problem-solving skills (van Merriënboer, 1997). The theoretical rationale is based in Cognitive Load Theory (Sweller, 1988). Working memory has a limited capacity that can be filled by intrinsic, extraneous or germane cognitive load (Sweller, van Merriënboer, & Paas, 1998). When novices are first learning the schemas necessary to solve new types of problems, actually trying to solve the problem imposes an additional cognitive load, an *extraneous* cognitive load, and denies the limited working memory resources to cognition germane to learning. A large number of laboratory experiments and a smaller number of classroom studies have demonstrated that students learn more efficiently from problem solving activities when worked examples mixed in (Pashler et al., 2007). Others have compared learning only by problem solving to only by studying worked examples and found that pure worked example study was better for novices. As the learner develops in a domain, the benefit of worked examples recedes by the *expertise reversal effect* (Kalyuga, Ayres, Chandler, & Sweller, 2003).

Self-explanation has been demonstrated to improve student learning*:* students who explain examples to themselves learn better, make more accurate self-assessments of their understanding and use analogies more economically while solving problems (Pashler et al., 2007; VanLehn, Jones, & Chi, 1992). Seminal work on the *self-explanation effect* found that the students who learn best appeared to learn from examples by explaining to themselves (Chi, Bassok, Lewis, Reimann, & Glaser, 1989). Students can be taught to self-explain, and when they do, they learn more effectively

(Bielaczyc, Pirolli, & Brown, 1995). The theoretical basis of self-explanation is that it promotes generation and repair of a student's mental models (Chi, 2000). For learning that depends on paired association or probabilistic inference, self-explanation may not help (Wylie, Koedinger, & Mitamura, 2009).

Prompting students to self-explain generally causes higher learning gains from studying a material than without prompting. Many students do not self-explain naturally and the quality of self-explanations themselves can be highly variable (Lovett, 1992; Renkl, 1997). The positive effects of prompting on the frequency and quality of students' self-explanations has been demonstrated with verbal prompts from human experimenters (Chi, 1994), prompts automatically generated by computer tutors (Aleven & Koedinger, 2002), or embedded in the learning materials themselves (Hausmann & VanLehn, 2007). The latter study also asked whether the effects of self-explanation are due to the generation of the explanation or attention to an explanation and the authors found that generation of one's own explanation was more effective than paraphrasing an author-provided one. With the paraphrasing as a check on attention paid to the author-provided explanation, the authors contend that generating is more effective than mere attending. While earlier work (Lovett, 1992) found that learners who generate the key inferences have the same learning gains as learners who read the corresponding inferences, they point out that in the Lovett study, the student-produced and author-provided explanations were of different qualities. While explanation quality may be a confound in a study of human learning, it is an important experimental condition for education research given that this difference is to be expected in natural environments. An important factor in the utility of instructional explanations is whether they are for learning concepts or procedures; a recent meta-analytic review concluded that instructional explanations in example-based learning have greater benefit for conceptual than procedural knowledge, though not necessarily more than self-explanations (Wittwer & Renkl, 2010).

Much of the effectiveness of worked examples depends on the behaviors the students engage in, which vary significantly across both individuals and environments (Renkl, 1997). Various studies have experimented with different designs to elicit these beneficial behaviors, but from a cognitive psychological perspective. I contend that there is now a need to go beyond cognitive psychology methods and theory to include the concerns of interaction design. Interaction design can more rapidly explore the space of possible designs, driven by the needs and practicalities of use rather than only the needs of rigorous and incremental theory. For example, through the methods of psychological research, after over a decade of research in self-explanation only recently have researchers identified self-explanations in which students contrast their own with that of an expert (Hausmann, Van De Sande, & VanLehn, 2008). Existing theory can help constrain the space. For example, in the goal of designing optimal learning from worked examples, leading researchers have concluded that instructional explanations hinder a student's own self-explaining (Schworm & Renkl, 2002). I use these theoretical findings to guide the interaction design of Examplify.

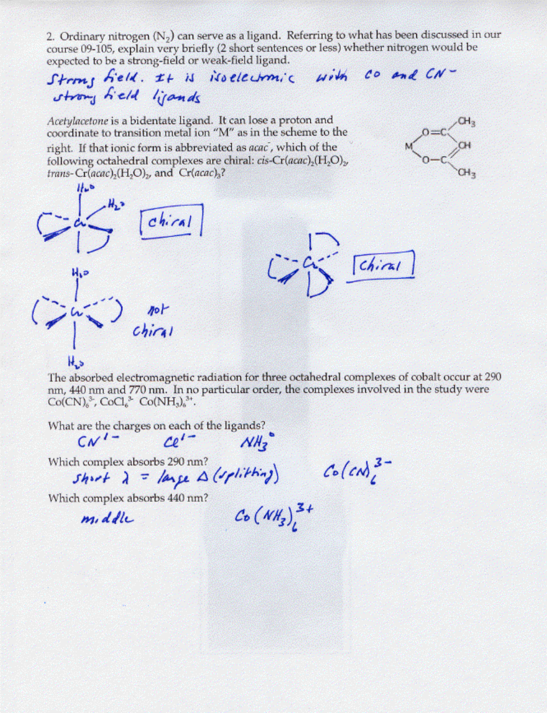
5.3 Core Features

Examplify began as an intention to develop a scalable software application to address the need perceived by both students and instructors to support students' in using study materials effectively. Following the fieldwork, I had established several design requirements for the application,

1. Scaffold effective study techniques for students that work even for students who don't know them
2. Be interactive enough that students are engaged
3. Help students to accurately assess what they know and don't know
4. Be self-paced so that students can go quickly over what they are already confident in
5. Map well to course assessments so that students know when they are prepared
6. Require no upfront action by the student in order to benefit
7. Require no changes to the instructor's curriculum or schedule
8. Require little or no time from the instructor to offer in her course

The key insight to the design of Examplify is that many instructors have a trove of exam preparation materials in their answer keys. I tried to conceive of a way to re-use these to help students prepare for exams. The field interviews (Chapter 3) made clear that instructors are reluctant to share good multiple-choice questions with students before the exam, but unless they re-use questions from semester to semester they would be willing to share questions from a previous exam. Some instructors do re-use questions from semester to semester because good multiple-choice questions can be so difficult to produce. However, worked solutions to problems can be easier to produce because they don't require tempting distractors or a simple unambiguous answer. Worked solutions do not have to demonstrate *the* right answer to a problem, just a valid answer. Further, some instructors offer these answer keys already to help their students prepare for exams. I noticed this during the pilot experience of Nudge and realized that building upon instructor answer keys would address design requirements 5 through 8.

Ideating on how to satisfy design requirements 2 and 4, I realized these answer keys could be made interactive and self-paced by letting students gradually reveal the expert work. Segmenting the work would not require the expertise necessary to author a new problem or worked example and could potentially be carried out quickly by students or outsourced remote workers. In the study below people with no chemistry knowledge took less than 1 minute per page. Figures 5-1 and 5-2 show an expert's example solution and the corresponding version covered up in steps. Further, by structuring the reveal interaction based in cognitive and metacognitive theory (described below), the activity could scaffold effective study technique (requirement 1) and help students accurately assess themselves (requirement 4).



2. Ordinary nitrogen (N_2) can serve as a ligand. Referring to what has been discussed in our course 09-105, explain very briefly (2 short sentences or less) whether nitrogen would be expected to be a strong-field or weak-field ligand.

Strong field. It is isoelectronic with CO and CN⁻ strong field ligands

Acetylacetone is a bidentate ligand. It can lose a proton and coordinate to transition metal ion "M" as in the scheme to the right. If that ionic form is abbreviated as *acac*⁻, which of the following octahedral complexes are chiral: *cis*-Cr(*acac*)$_2$(H_2O)$_2$, *trans*-Cr(*acac*)$_2$(H_2O)$_2$, and Cr(*acac*)$_3$?

The absorbed electromagnetic radiation for three octahedral complexes of cobalt occur at 290 nm, 440 nm and 770 nm. In no particular order, the complexes involved in the study were $Co(CN)_6^{3-}$, $CoCl_6^{3-}$ $Co(NH_3)_6^{3+}$.

What are the charges on each of the ligands?

CN^{1-} Cl^{1-} NH_3^{0}

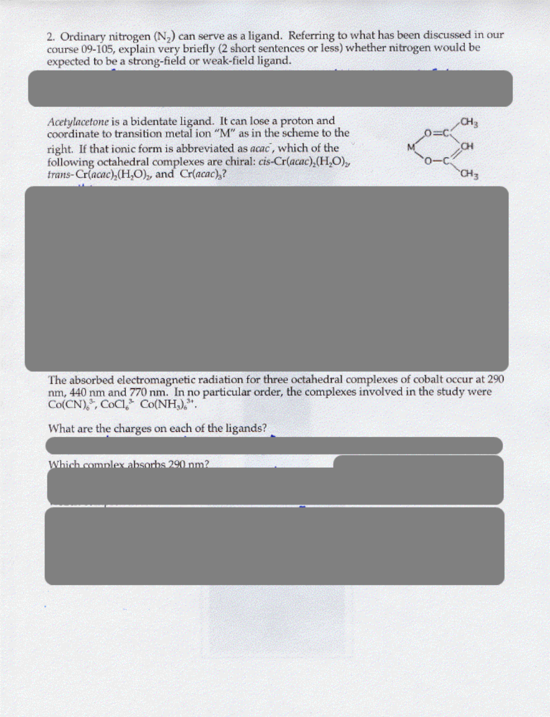
Which complex absorbs 290 nm?

short λ = large Δ (splitting) $Co(CN)_6^{3-}$

Which complex absorbs 440 nm?

middle $Co(NH_3)_6^{3+}$

Figure 5-2 An exam solution from an instructor



2. Ordinary nitrogen (N_2) can serve as a ligand. Referring to what has been discussed in our course 09-105, explain very briefly (2 short sentences or less) whether nitrogen would be expected to be a strong-field or weak-field ligand.

[redacted]

Acetylacetone is a bidentate ligand. It can lose a proton and coordinate to transition metal ion "M" as in the scheme to the right. If that ionic form is abbreviated as *acac*⁻, which of the following octahedral complexes are chiral: *cis*-Cr(*acac*)$_2$(H_2O)$_2$, *trans*-Cr(*acac*)$_2$(H_2O)$_2$, and Cr(*acac*)$_3$?

[redacted]

The absorbed electromagnetic radiation for three octahedral complexes of cobalt occur at 290 nm, 440 nm and 770 nm. In no particular order, the complexes involved in the study were $Co(CN)_6^{3-}$, $CoCl_6^{3-}$ $Co(NH_3)_6^{3+}$.

What are the charges on each of the ligands?

[redacted]

Which complex absorbs 290 nm?

[redacted]

[redacted]

Figure 5-1 Solution covered up in steps

With this rough design in mind, I reviewed relevant theoretic literature to settle upon an array of theoretically grounded features. Table 5-1 details each feature, the design claim behind it, and the warranting evidence. The overall design is a self-paced interactive study aid that helps students' to engage more actively with worked examples.

Examplify supports cognitive engagement by scaffolding a step-by-step walk-through of a problem posed and its solution. Based on evidence that students who explain to themselves learn more from examples (Chi et al., 1989), at each step the tool focuses the student on a part of the solution and prompts for an explanation. To motivate this activity, to help students check themselves, and to provide support to students who are still unsure, after submitting an explanation the system shows the student explanations that others have submitted. Students can click up or down to give feedback on how helpful the other explanation is and the more helpful explanations will be shown more frequently. After seeing as many as they want, they can revise their own explanation and resubmit. This helps to enhance the question resource with byproducts of the learning activity.

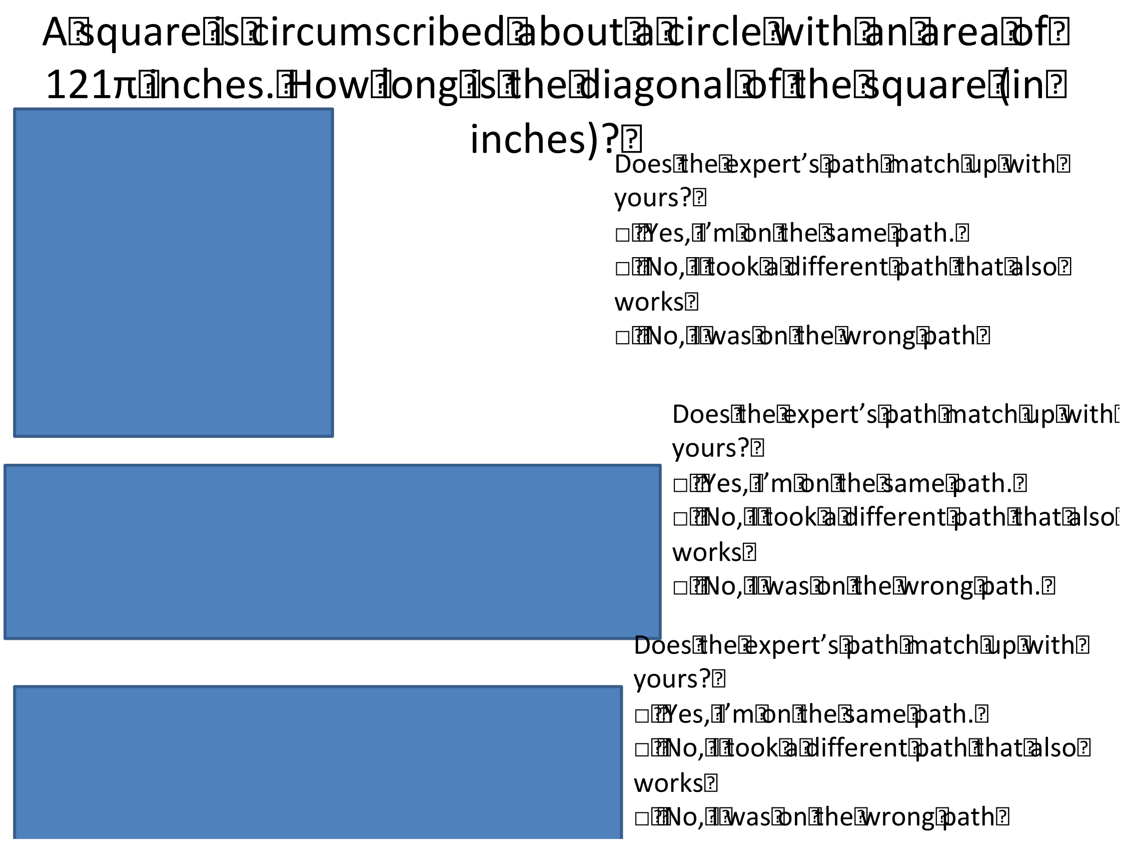
Examplify has been designed to support accurate self-assessment while learning. Students are often deceived by their illusions of understanding. For example, they often read through a practice or past exam problem without making a real effort to answer them or even think about the content (Renkl, 2002). Students may convince themselves of good performance by assuming or feeling like they could produce the answer shown. Many students study by transcribing their notes until they are not confused, but this may be less effective than taking a test on that material. (Karpicke & Blunt, 2011). Students often go into passive learning while reading or in lecture. They overestimate how quickly they have understood (e.g. when people are allowed to decide when to stop studying, their memory performance can be worse than when the experimenter controls their timing (Kornell & Bjork, 2007; Metcalfe & Kornell, 2007). They do not realize when extra study time will help (Koriat, 1997).

Like Nudge, Examplify was developed through a series of iterations. The prototypes have many visible states fluctuated widely from iteration to iteration, due to the cognitive complexity of the task. Figure 5-3 shows a screenshot of the PowerPoint prototype, which used animation to reveal the parts of the model solution. Figure 5-4 shows the sequence of screens in the implemented version of Examplify. Figure 5-6 steps through the interactions with the screens. After trying to solve the problem on paper, the student clicks *Check my work*. The expert's work appears and they click to indicate how well their work matches (e.g. "Partly right"). They are then prompted to explain, "Why did the expert do this?" They can then click that *Yes* they understand the work or *Not yet*. Either button proceeds to prompt them to work out the next step. This version was user tested with students in a summer version of the course. Like Nudge, the feature set evolved by observing their use and drawing on evidence-based learning science principles. The final set of features is presented in Table 5-1.

Table 5-1 Examplify feature matrix

Feature	Claim	Warrant	Status
Present model solutions step-wise	Breaking a problem into steps focuses attention productively	Modular steps reduce task-related “intrinsic” cognitive load and shift it to the germane (Gerjets & Scheiter, 2006)	Implemented; Piloted
Reuses instructor’s extant materials	Instructors are more likely to adopt a technology that a) doesn’t require more work and b) teaches the way they do.	Results of contextual inquiry	Implemented; Piloted
Prompt students for explanations of explanations of the expert’s work	Explaining correct examples improves learning but students need scaffolds to do so.	Students studying worked examples do not spontaneously explain (Chi et al., 1989; Renkl, 1997)	Implemented; Piloted
Require valid explanation in order to advance	Students won’t explain unless required to	Learner control causes students to not use the prompts (Scheiter, Gerjets, & Vollmann, 2006) Instructional explanations hinder learners in generating explanatory justifications of solution steps (Schworm & Renkl, 2002)	Piloted; Rejected by user testing
Shows explanations for each step	Step-based explanations from students will improve learning	Coupling worked examples with instructional explanations of steps improves learning (Catrambone & Yuasa, 2006; van Gog, Paas, & van Merriënboer, 2006)	Implemented
Source the explanations from	Student explanations may be more effective	Non-experts often make better explanations of work than experts do	Implemented

other students		(Aleahmad, Aleven, & Kraut, 2009) Expert knowledge creates blind spots in instruction (Nathan, Koedinger, & Alibali, 2001)	
Prompt students to attempt solving a step of the problem before seeing expert's work	Prompting work leads to better learning from the example	Taking memory tests improves long-term retention (Roediger & Karpicke, 2006a); both in the lab and classroom (McDaniel, Roediger, & McDermott, 2007)	Implemented; Controlled in experiment
At end of step prompt for cognitive load	Proper cognitive load is an important factor in the effectiveness of an example	Excessive information can produce too much cognitive load and interfere with schema development (Sweller et al., 1998) Simple measures of cognitive load can be reliable (Gerjets & Scheiter, 2006)	Design driven during study



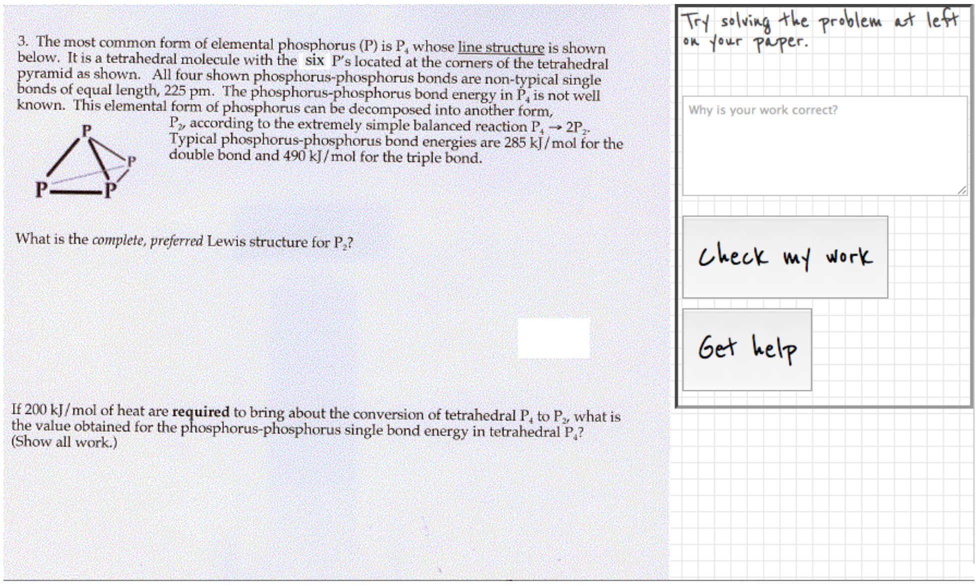
A square is circumscribed about a circle with an area of 121π inches. How long is the diagonal of the square (in inches)?

Does the expert's path match up with yours?
□ Yes, I'm on the same path.
□ No, I took a different path that also works
□ No, I was on the wrong path

Does the expert's path match up with
yours?
□ Yes, I'm on the same path.
□ No, I took a different path that also
works
□ No, I was on the wrong path.

Does the expert's path match up with yours?
□ Yes, I'm on the same path.
□ No, I took a different path that also works
□ No, I was on the wrong path

Figure 5-3 PowerPoint prototype of Examplify

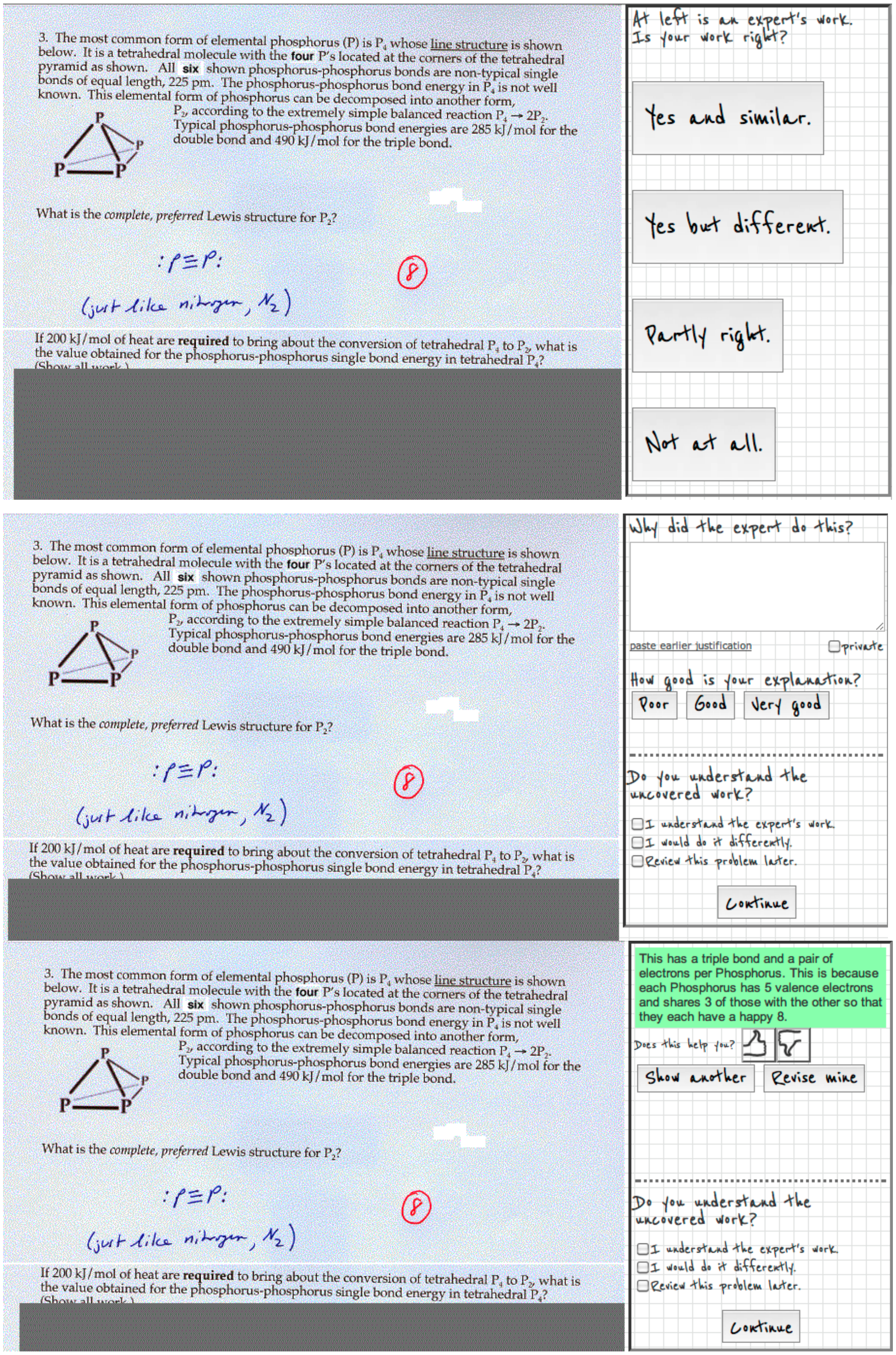


3. The most common form of elemental phosphorus (P) is P_4 whose line structure is shown below. It is a tetrahedral molecule with the six P's located at the corners of the tetrahedral pyramid as shown. All four shown phosphorus-phosphorus bonds are non-typical single bonds of equal length, 225 pm. The phosphorus-phosphorus bond energy in P_4 is not well known. This elemental form of phosphorus can be decomposed into another form, P_2, according to the extremely simple balanced reaction $P_4 \rightarrow 2P_2$. Typical phosphorus-phosphorus bond energies are 285 kJ/mol for the double bond and 490 kJ/mol for the triple bond.

What is the *complete, preferred* Lewis structure for P_2?

If 200 kJ/mol of heat are **required** to bring about the conversion of tetrahedral P_4 to P_2, what is the value obtained for the phosphorus-phosphorus single bond energy in tetrahedral P_4? (Show all work.)

Try solving the problem at left on your paper.



3. The most common form of elemental phosphorus (P) is P_4 whose line structure is shown below. It is a tetrahedral molecule with the **four** P's located at the corners of the tetrahedral pyramid as shown. All **six** shown phosphorus-phosphorus bonds are non-typical single bonds of equal length, 225 pm. The phosphorus-phosphorus bond energy in P_4 is not well known. This elemental form of phosphorus can be decomposed into another form, P_2, according to the extremely simple balanced reaction $P_4 \rightarrow 2P_2$. Typical phosphorus-phosphorus bond energies are 285 kJ/mol for the double bond and 490 kJ/mol for the triple bond.

What is the *complete, preferred* Lewis structure for P_2?

$:P\equiv P:$

(just like nitrogen, N_2)

8

If 200 kJ/mol of heat are **required** to bring about the conversion of tetrahedral P_4 to P_2, what is the value obtained for the phosphorus-phosphorus single bond energy in tetrahedral P_4? (Show all work.)

At left is an expert's work. Is your work right?

3. The most common form of elemental phosphorus (P) is P_4 whose line structure is shown below. It is a tetrahedral molecule with the **four** P's located at the corners of the tetrahedral pyramid as shown. All **six** shown phosphorus-phosphorus bonds are non-typical single bonds of equal length, 225 pm. The phosphorus-phosphorus bond energy in P_4 is not well known. This elemental form of phosphorus can be decomposed into another form, P_2, according to the extremely simple balanced reaction $P_4 \rightarrow 2P_2$. Typical phosphorus-phosphorus bond energies are 285 kJ/mol for the double bond and 490 kJ/mol for the triple bond.

What is the *complete, preferred* Lewis structure for P_2?

$:P\equiv P:$

(just like nitrogen, N_2)

8

If 200 kJ/mol of heat are **required** to bring about the conversion of tetrahedral P_4 to P_2, what is the value obtained for the phosphorus-phosphorus single bond energy in tetrahedral P_4? (Show all work.)

Why did the expert do this?

How good is your explanation?

Do you understand the uncovered work?

□ I understand the expert's work.
□ I would do it differently.
□ Review this problem later.

3. The most common form of elemental phosphorus (P) is P_4 whose line structure is shown below. It is a tetrahedral molecule with the **four** P's located at the corners of the tetrahedral pyramid as shown. All **six** shown phosphorus-phosphorus bonds are non-typical single bonds of equal length, 225 pm. The phosphorus-phosphorus bond energy in P_4 is not well known. This elemental form of phosphorus can be decomposed into another form, P_2, according to the extremely simple balanced reaction $P_4 \rightarrow 2P_2$. Typical phosphorus-phosphorus bond energies are 285 kJ/mol for the double bond and 490 kJ/mol for the triple bond.

What is the *complete, preferred* Lewis structure for P_2?

$:P\equiv P:$

(just like nitrogen, N_2)

8

If 200 kJ/mol of heat are **required** to bring about the conversion of tetrahedral P_4 to P_2, what is the value obtained for the phosphorus-phosphorus single bond energy in tetrahedral P_4? (Show all work.)

This has a triple bond and a pair of electrons per Phosphorus. This is because each Phosphorus has 5 valence electrons and shares 3 of those with the other so that they each have a happy 8.

Does this help you?

Do you understand the uncovered work?

□ I understand the expert's work.
□ I would do it differently.
□ Review this problem later.

Figure 5-4 Screenshots of implemented Examplify

5.3.1 Competing Predictions

In the course of iterating, a feature was questioned that could be resolved by neither user testing nor the literature: whether students should be prompted to solve the problem before seeing the solution. User testing cannot answer this because the evaluation function is not user preference or facility, but long-term learning. The learning science literature is contradictory (Koedinger, Corbett, & Perfetti, in press): the testing effect literature has shown that being tested improves retention (Roediger & Karpicke, 2006b) but the worked example literature has shown that adding worked examples (more study trials and fewer test trials) improves novice learning (Pashler et al., 2007) and are sometimes most effective without the addition of problem solving (Paas, 1994). Later work helps clarify when worked examples are best and when to interleave them with problem solving (McLaren, Lim, & Koedinger, 2008; Salden, Aleven, Renkl, & Schwonke, 2009; Salden, Koedinger, Renkl, Aleven, & McLaren, 2010).

There is no clear consensus on the optimal design for a system with the goals of Examplify. This system and evaluation differ from prior related work on testing, worked examples and self-explanation in several ways:

1) These model solutions are “found” from materials designed as assessments, not authored as example-based instruction like in most worked example studies.

2) The test here requires active problem solving, not mere recall as in the testing effect studies.

3) Because the problem-solving test has no single correct response with which to compare one’s answer, the benefits of the model solution rely on the learner’s ability to compare it against their own solution.

4) The prompted self-explanation is a form of testing, albeit neither with correctness feedback or paired associations in most testing effect studies.

5) The examples are used voluntarily in a real course students are taking.

To resolve whether students should be prompted to solve, I experimented with two alternative versions: one emphasizing self-testing as shown above and the other emphasizing worked examples and self-explanation by omitting the first two screens of the interaction.

5.3.2 Benefits of Worked Examples

Cognitive load theory (à la “worked example effect”) suggests that novice students learn new procedures more efficiently by replacing many problems with worked examples. For novices, the cognitive load of attempting to solve problems takes away mental resources that could be more effectively used to learn from the example (Sweller, 1988). When learners are novice in a domain, studying worked examples requires less cognitive load than solving matched problems, leaving cognitive resources needed to learn.

When students are proficient in a domain, the worked out part of the examples can hinder rather than help by adding extraneous cognitive load that distracts students from productive problem solving. This “expertise reversal effect” has been observed for expertise in multiple domains, including chemistry (Sweller, Ayres, & Kalyuga, 2011). However in most studies, the number of examples is controlled by the experiment. How would the use of worked examples play out when students can use examples completely at their own discretion? In most studies the worked examples are carefully designed. How well would "found" worked examples from instructors' archives do?

"Modular" worked examples break down complex problem solutions into smaller meaningful solution elements to "convey knowledge on problem categories together with category-specific solution recipes" (Gerjets, Scheiter, & Catrambone, 2004). This lowers intrinsic cognitive load and thus improves learning. In this case the worked examples are not authored, but found. They lack the instructional explanations and explicit category labeling of the solution recipes. Can this crude modularization method, produced by covering up parts or steps of a written solution, offer similar benefits to learning?

5.3.3 Benefits of Problem Solving

Worked examples help by removing problem solving to reduce cognitive load when learning. However problem solving can also help by promoting the active construction of knowledge (Anderson, Corbett, Koedinger, & Pelletier, 1995).

Studying through testing requires more retrieval of knowledge, which facilitates future performance (Karpicke, 2010). The "testing effect" literature suggests that students learn more robustly by executing mental effort, as they would have to on a future assessment. For example for some problems, such as in chemistry, the hard part is to know how to frame the problem rather than the mechanics of solving within that frame. If students are not confronted with the task of generating the frame, they may accurately self assess their ability to execute the mechanics yet not realize that they are not prepared for an exam.

The apparent tension between worked examples and problem solving can be reconciled by adaptively presenting the more appropriate activity based on the performance of the learner. Intelligent tutoring systems adapt the learning activity in sophisticated ways but are computationally complex and require 100-1000 hours of time from skilled experts to produce each hour of student instruction (Murray, 1999). Another effective technique is simply to fade from worked examples when students are naïve to problem solving when they are more knowledgeable (Atkinson, Renkl, & Merrill, 2003; Renkl, Atkinson, & Maier, 2000). Worked examples can be made more cheaply by less skilled authors than required for intelligent tutoring systems (Aleahmad et al., 2009) and fading can be directed by the learner instead of a complicated artificial intelligence.

5.3.4 Two kinds of worked example interaction

Examplify creates adaptive learning activities using existing answer keys as content. Figure 5-5 presents the learner interaction flow and Figure 5-6 shows screen shots of each state. Each worked example starts in the Try state as a problem solving activity, with all the expert's work occluded. The learner tries to solve the problem on their paper as they would on an exam. If they aren't able to produce any work or feel it is too difficult, they can click *Get Help* to reveal part of the expert's work. If they are able to make any progress on the problem, they click *Check my work* which also reveals part of the expert's solution. To proceed they reflect and indicate how similar their work is. After a new portion of the expert work is revealed, learners are prompted to reflect on why that is the appropriate work for the problem. To advance, they reflect and indicate whether they understand the work shown. That click takes them to the Try state again but for the next portion of work. The nonsolving variant is used as a control condition in the evaluation study (described below).

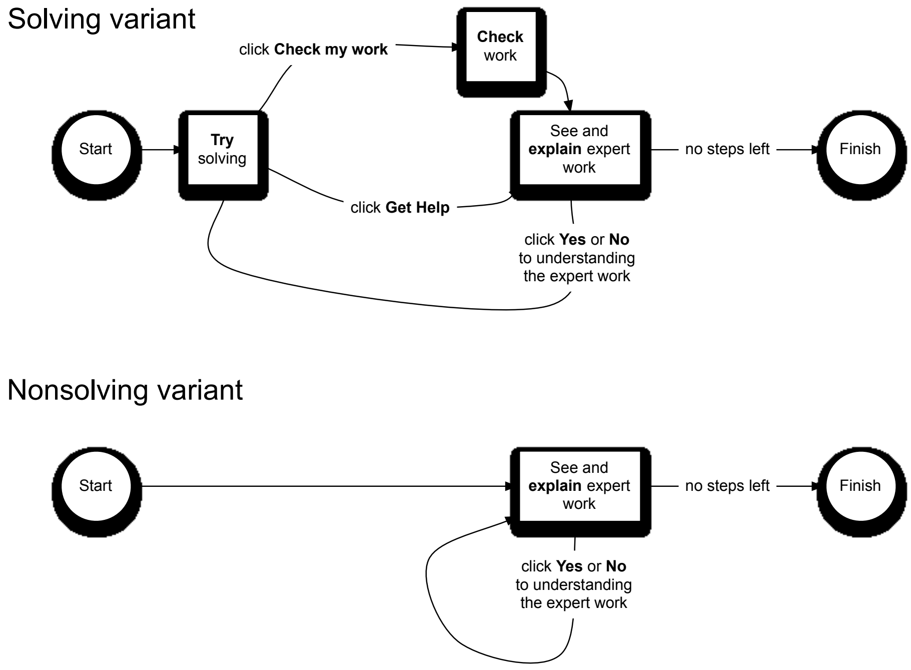
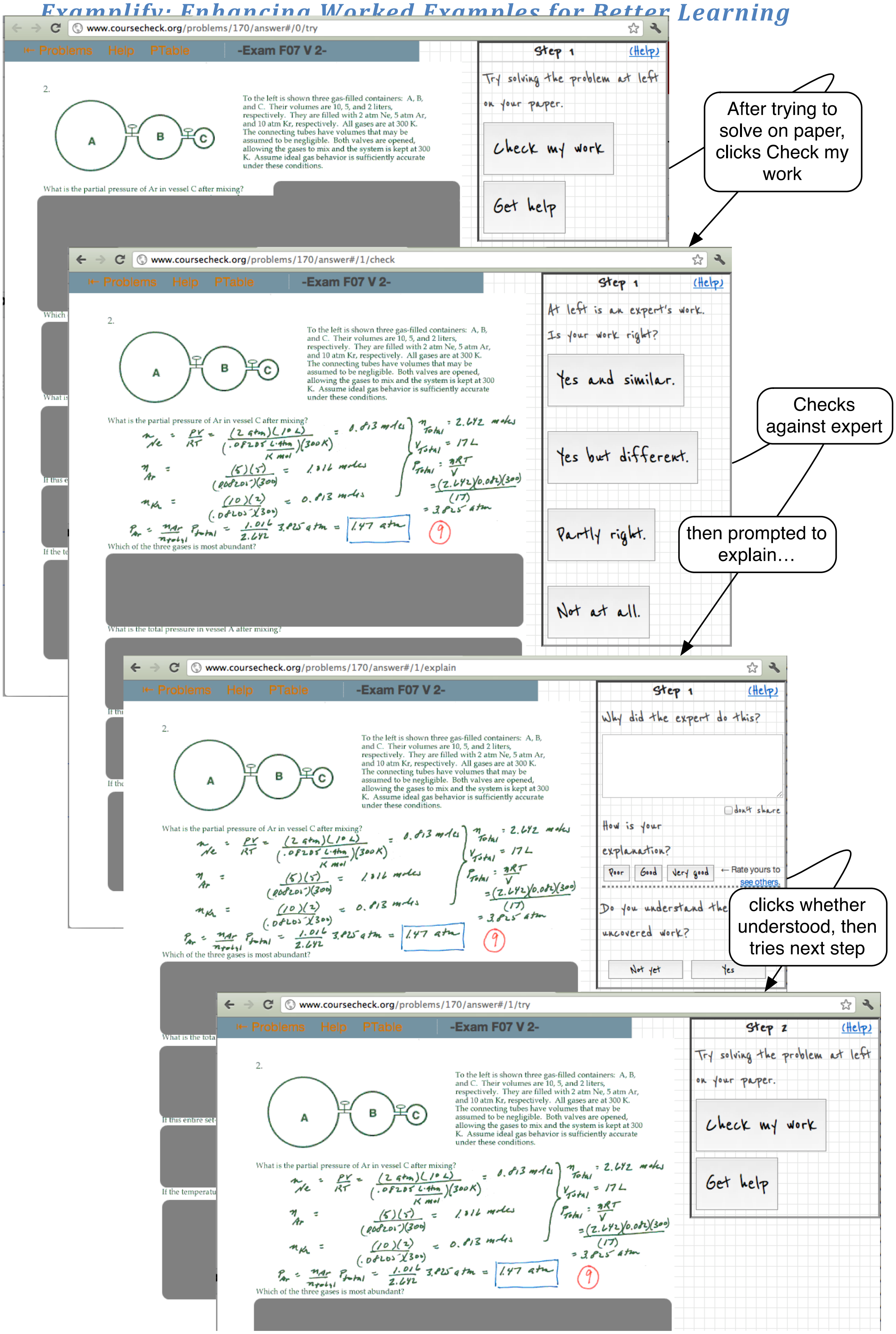

Solving variant
click **Check my work**
Check work
Start
Try solving
See and **explain** expert work
no steps left
Finish
click **Get Help**
click **Yes** or **No** to understanding the expert work
Nonsolving variant
Start
See and **explain** expert work
no steps left
Finish
click **Yes** or **No** to understanding the expert work

Figure 5-5 States and transitions of Examplify worked example interaction



www.coursecheck.org/problems/170/answer#/0/try

2.

To the left is shown three gas-filled containers: A, B, and C. Their volumes are 10, 5, and 2 liters, respectively. They are filled with 2 atm Ne, 5 atm Ar, and 10 atm Kr, respectively. All gases are at 300 K. The connecting tubes have volumes that may be assumed to be negligible. Both valves are opened, allowing the gases to mix and the system is kept at 300 K. Assume ideal gas behavior is sufficiently accurate under these conditions.

What is the partial pressure of Ar in vessel C after mixing?

Step 1 (Help)

Try solving the problem at left on your paper.

After trying to solve on paper, clicks Check my work

www.coursecheck.org/problems/170/answer#/1/check

What is the partial pressure of Ar in vessel C after mixing?

$$n_{Ne} = \frac{PV}{RT} = \frac{(2\ \text{atm})(10\ L)}{(.08205\ \frac{L\cdot\text{atm}}{K\ mol})(300K)} = 0.813\ \text{moles}$$

$$n_{Ar} = \frac{(5)(5)}{(0.08205)(300)} = 1.016\ \text{moles}$$

$$n_{Kr} = \frac{(10)(2)}{(.08205)(300)} = 0.813\ \text{moles}$$

$$n_{Total} = 2.642\ \text{moles}$$

$$V_{Total} = 17\ L$$

$$P_{Total} = \frac{nRT}{V} = \frac{(2.642)(0.082)(300)}{(17)} = 3.825\ \text{atm}$$

$$P_{Ar} = \frac{n_{Ar}}{n_{total}} P_{total} = \frac{1.016}{2.642} 3.825\ \text{atm} = \boxed{1.47\ \text{atm}}$$

Which of the three gases is most abundant?

What is the total pressure in vessel A after mixing?

Step 1 (Help)

At left is an expert's work. Is your work right?

Checks against expert

then prompted to explain…

www.coursecheck.org/problems/170/answer#/1/explain

Step 1 (Help)

Why did the expert do this?

□ don't share

How is your explanation?

← Rate yours to see others.

Do you understand the uncovered work?

clicks whether understood, then tries next step

www.coursecheck.org/problems/170/answer#/1/try

Step 2 (Help)

Try solving the problem at left on your paper.

Figure 5-6 Screen shots from an example usage of the Solving variant

5.3.5 Implementation

Examplify is implemented as a web application that runs in any modern web browser. The backend was developed in Ruby on Rails 3.1 with a PostgreSQL database and hosted on Heroku (PaaS) servers. The frontend was developed in HTML5, jQuery and Backbone.js.

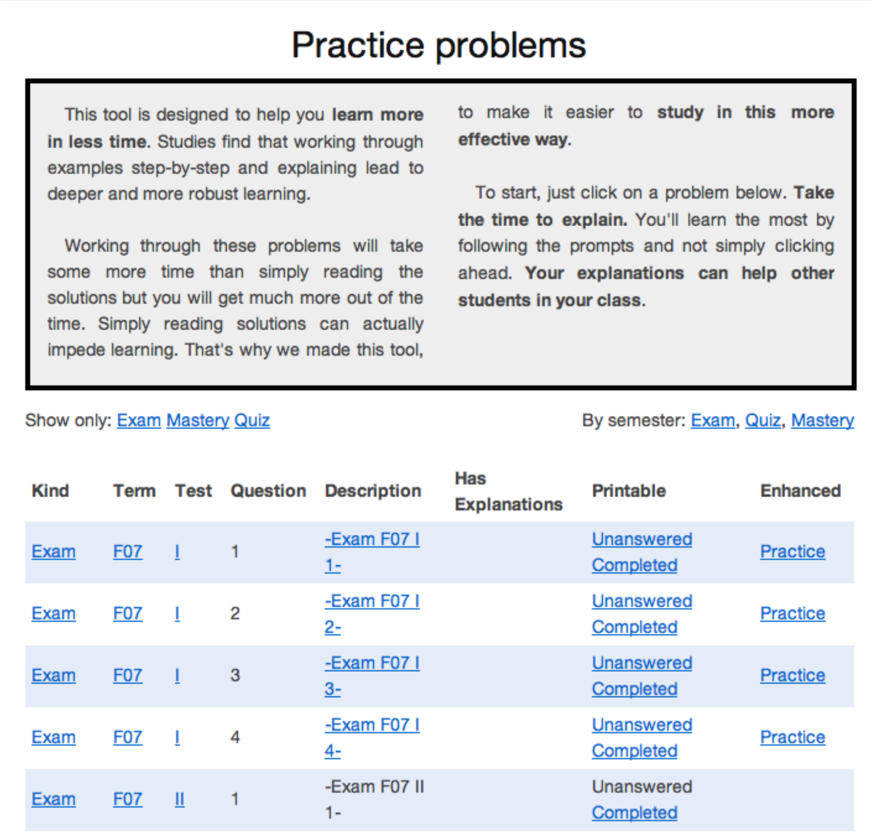
5.3.6 Problem browser

Students find examples to open through the problem browser. The browser started as in Figure 5-7 but was later improved as in Figure 5-8. For each problem, the student can choose an empty version of the problem like on the test, a completed version like a printed answer key, or the interactive version specific to Examplify. All references to "examples" in the evaluation study refer to these interactive examples.

At the top of the problem browser a blurb reads:

> This tool is designed to help you **learn more in less time**. Studies find that working through examples step-by-step and explaining lead to deeper and more robust learning.
>
> Working through these problems will take some more time than simply reading the solutions but you will get much more out of the time. Simply reading solutions can actually impede learning. That's why we made this tool, to make it easier to **study in this more effective way**.
>
> To start, just click on a problem below. Try solving the problem shown. You can click to check your work or get help. **Take the time to explain.** You'll learn the most by following the prompts and not simply clicking ahead. **Your explanations can help other students in your class**.

For students with the nonsolving control variant (described below) the underlined text is omitted.



Practice problems

This tool is designed to help you **learn more in less time.** Studies find that working through examples step-by-step and explaining lead to deeper and more robust learning.

Working through these problems will take some more time than simply reading the solutions but you will get much more out of the time. Simply reading solutions can actually impede learning. That's why we made this tool, to make it easier to **study in this more effective way.**

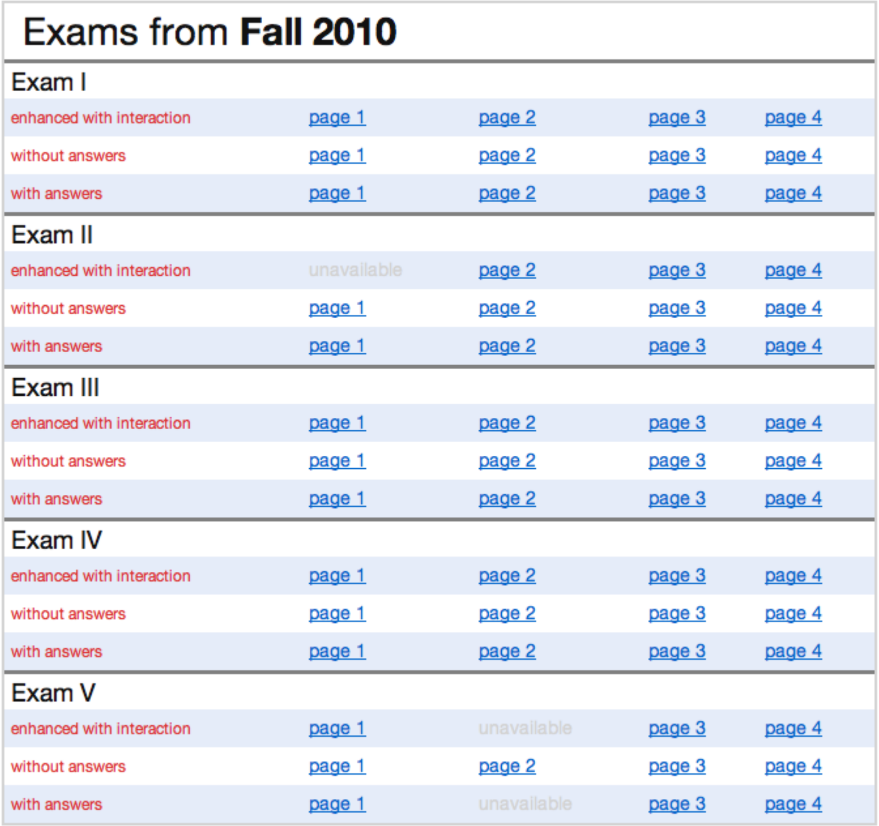
To start, just click on a problem below. **Take the time to explain.** You'll learn the most by following the prompts and not simply clicking ahead. **Your explanations can help other students in your class.**

Show only: Exam Mastery Quiz

By semester: Exam, Quiz, Mastery

Kind	Term	Test	Question	Description	Has Explanations	Printable	Enhanced
Exam	F07	I	1	-Exam F07 I 1-		Unanswered Completed	Practice
Exam	F07	I	2	-Exam F07 I 2-		Unanswered Completed	Practice
Exam	F07	I	3	-Exam F07 I 3-		Unanswered Completed	Practice
Exam	F07	I	4	-Exam F07 I 4-		Unanswered Completed	Practice
Exam	F07	II	1	-Exam F07 II 1-		Unanswered Completed	

Figure 5-8 Problem browser from start of semester until Exam 3



Exams from **Fall 2010**

Exam I				
enhanced with interaction	page 1	page 2	page 3	page 4
without answers	page 1	page 2	page 3	page 4
with answers	page 1	page 2	page 3	page 4
Exam II				
enhanced with interaction	unavailable	page 2	page 3	page 4
without answers	page 1	page 2	page 3	page 4
with answers	page 1	page 2	page 3	page 4
Exam III				
enhanced with interaction	page 1	page 2	page 3	page 4
without answers	page 1	page 2	page 3	page 4
with answers	page 1	page 2	page 3	page 4
Exam IV				
enhanced with interaction	page 1	page 2	page 3	page 4
without answers	page 1	page 2	page 3	page 4
with answers	page 1	page 2	page 3	page 4
Exam V				
enhanced with interaction	page 1	unavailable	page 3	page 4
without answers	page 1	page 2	page 3	page 4
with answers	page 1	unavailable	page 3	page 4

Figure 5-7 Problem browser from Exam 3 until end of term

5.4 Experimental Design

5.4.1 Context

The study took place in a large introductory chemistry class at a competitive private university. The course curriculum is stable and the instructor has a large bank of old exams. For the past 12 years, after every exam the instructor has solved the test, scanned the solutions and put them online. Each exam question is a separate page and there are four pages per exam for 25 points each.

Each page is one interactive example within Examplify. To add them into Examplify required covering each step of expert work with a gray box (see Figure 5-1). This task was distributed among several paid assistants with no chemistry expertise. They each took less than 1 minute per page.

The course was taught in two lecture sections between which students chose (10:30am, n=136 and 11:30am, n=86). Students may have chosen based on earliness in the day or constraints of their schedules.

5.4.2 Conditions

I compare Examplify (with solving) to a nonsolving control variant of Examplify and to a business-as-usual (BAU) control section. Students self-selected into the Examplify or BAU sections, presumably to meet the constraints of their course schedules, and had no knowledge there would be any differences between them.

In the Examplify section, students had access to the default version of Examplify with the solving prompt. They accessed the Examplify site through Blackboard or course announcement emails. Accessing from Blackboard required generating and remembering a password.

Students who opted into the study were randomly assigned to receive either the solving variant of Examplify or an alternate, nonsolving variant of Examplify. In the nonsolving control, the prompt to solve the step of the example was removed (Figure 5-5). Instead students immediately saw the first step of the solution and were prompted to explain before clicking through to the next step (Figure 5-6). References to solving were also removed from the explanatory text on the problem browser page (Figure 5-8).

The BAU control section operated no differently from previous years of the course, except students enrolled in the study filled out questionnaires and polls about what they had done in the class.

5.4.3 Hypotheses

The hypothesis that Examplify with solving will improve learning on both immediate and delayed measures follows from past theory in that this condition combines the benefits of worked examples and testing. That is, it prompts for self-testing but students can quickly get a worked example step if needed. This hypothesis is novel and, in fact, application of cognitive load theory might suggest the opposite, namely, that the prompt for self-testing (problem solving) may be extraneous load and thus the non-solving variant would be predicted to be better.

5.4.3.1 H-immediate

Students with Examplify with solving interaction score higher on immediate assessments.

This H-immediate hypothesis is operationalized as higher scores across the four non-cumulative exams, both versus the nonsolving control variant and BAU control section.

Examplify is designed to reduce the cognitive load of problem solving by decomposing the steps of the problem and allowing students to see a solution immediately if they choose. While this interaction uses more cognitive load than a simple worked example, this may be germane cognitive load that helps them assess their understanding.

Both course sections have worked examples, but in Examplify they are broken up into steps. This modular form has been found to be more efficient and to reduce cognitive load. When students are ready to solve problems, Examplify may be more motivating than the BAU static questions and also scaffold better study strategies.

5.4.3.2 H-delayed

The benefits of Examplify with solving will be greater on delayed assessments than immediate assessments.

Examplify with solving should increase the frequency of students recalling information (testing effect) and proceduralization (learning by doing). Both these activities improve robustness of learning, which I measure by comparing delayed and immediate tests on the same topics, and again versus both controls.

5.4.4 Knowledge measures

All knowledge measures came from the normal course assessments. Accordingly, there are no formal pretest measures.

There were 4 non-cumulative exams (E1-4) distributed evenly over the term such that each exam covered the immediately preceding material. During the final exam period, a fifth exam was given of which half was on topics from the latest exam (E4) and half was on earlier topics (E2-3). A student's score on this could replace their lowest exam grade.

I use the half of the fifth exam that is on early topics as a delayed measure of learning, referred to below as "Delayed exam scores on early topics". The paired immediate measure is the average score of the two exams on those earlier topics (E2-3), referred to below as "Immediate exam scores on early topics".

5.4.5 Explanatory measures

Each student's personal attributes affect how she uses Examplify, which in turn affect how the tool affects her and her learning. To understand how the tool works differently for different students, I logged user activities and collected several large questionnaires over the term. (These measures are the same as in the Nudge study in Chapter 4.)

Behavioral measures include their interactions with Examplify and questionnaires about their time and study behaviors.

Cognitive measures include their math aptitude, operationalized as the SAT or ACT Math score reported on the questionnaires. (ACT scores were normalized to SAT.)

Metacognitive and motivation measures were numerous on the questionnaires. One factor that comes up in the results is mastery-avoidance from the 2 X 2 Achievement Goal Framework (Elliot & McGregor, 2001). In a mastery-avoidance goal orientation, students strive to avoid misunderstanding or failing to learn course material. The 7-pt scale is of agreement with statements such as, "I am often concerned that I may not learn all that there is to learn in this class."

5.4.6 Attrition and Missing Observations

17 students signed up for the study, but never did any coursework and were omitted from all analysis.

11 of these non-starters were in the Examplify section (11%) and 6 in the BAU control section (9%). Within the Examplify section 7 (13%) were in the solving condition and 4 (8%) were in the nonsolving condition.

Of students who started the course, four (2.6%) dropped before the end. They are included in analyses for which their data are available.

5.4.7 Timeline

To help interpret the following results, Figure 5-9 Timeline of Examplify study shows a timeline of the course, assessments, questionnaires and when changes were made to Examplify. The questionnaires were given before instruction, after the 3rd exam, and after the course final exam. After the 2nd exam, links were added to the Examplify tool allowing students to use the traditional static example problems. After the 3rd exam, based on the results of that questionnaire, the ease of accessing and navigated the tool was improved.

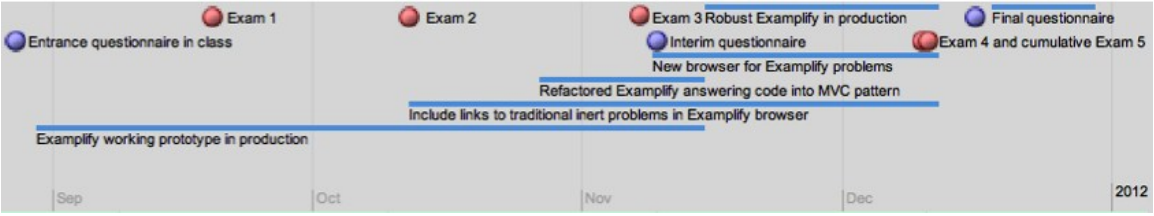


Figure 5-9 Timeline of Examplify study

5.5 Results

5.5.1 Descriptive statistics

5.5.1.1 Pre-existing differences

Because the lecture sections are not randomly assigned, I tested for any natural differences between them. Table 5-2 details the incoming attributes of students in each condition. Within the Examplify section, I separate people who never used the tool (Never opened) from people who opened the solving or nonsolving variants of the tool, because the conditions make no difference for students who never opened it (confirmed statistically).

Table 5-2 Incoming attributes and usage

Group	Freshman proportion	Math aptitude (200-800)	Mastery-avoidance (1-7)
Control section	38% (23/60)	711 (n=51)	5.2 (sd=1.1, n=50)
Examplify section	65% (60/93)	723 (n=78)	4.5 (sd=1.6, n=82)
- *Never opened*	33% (1/3)	730 (n=3)	5.4 (sd=1.9, n=3)
- *Nonsolving*	61% (28/46)	717 (n=37)	4.3 (sd=1.5, n=41)
- *Solving*	70% (31/44)	726 (n=38)	4.6 (sd=1.8, n=38)

The Examplify section had significantly higher proportion of freshman (X^2=10.1, p=.0014) and students reported significantly lower self-ratings on mastery-avoidance goal orientation (α=.84, F(1,130)=8.0, p=.006).

As a check against differential attrition across the randomly assigned variant conditions, there no were no significant differences on any of these measures between the solving and nonsolving conditions.

5.5.1.2 Subjective rating

A questionnaire was given after the 3rd exam asking the usefulness of several features of the course. 40% of respondents (n=54) rated "Interactive tool to study past problems" as "Good" or "Great" (15%). There were no differences by condition. 29% didn't perceive it as useful and 26% didn't yet know about it. Ease of accessing the tool was improved for the 4th quarter of the term by making the web link more prominent in Blackboard and updating the problem browser from as in Figure 5-7 to as in Figure 5-8.

5.5.1.3 Examplify usage

Ninety-seven percent of students in the study opened the Examplify tool (no difference by Examplify condition) and every one of those opened at least one exam example. Students in the Solving variant went on to open more overall (p=.0015) and across example types (F(1,88)=11.3, p=.001) than students in the Non-solving variant.

Freshman status, math aptitude and mastery-avoidance motivation did not predict open rates nor did they interact with solving condition to predict open rates.

At the beginning of the study there were some usability kinks in browsing examples that were slowly worked out over the term. All changes were in common between conditions and the last changes were deployed immediately after the 3rd exam. To see the change in use over time, I count the number of examples opened in each quarter of the term (before each exam). The days between exams were similar, though the period from the 2nd to 3rd exam was shorter than the others. The 4th exam period follows the improved navigation described above as a response to the questionnaire after the 3rd exam.

Table 5-3 shows the average number of interactive examples opened during the periods between each exam. High users are those who opened more than 3 exam examples over the term (the median usage among who had access to the tool). The number opened goes up over the term (F(3,276)=31.0, p<.0001) but among solving students increases more (F(3,306)=3.8, p=.011).

Table 5-3 Average number of interactive examples opened during each exam preparation period

	Overall usage	**Exam 1**	**Exam 2**	**Exam 3**	**Exam 4**
Nonsolving	All (n=48)	0.1	0.3	0.4	2.1
	Low only (n=33)	0.1	0.1	0.4	0.6
	High only (n=15)	0.2	0.7	0.6	5.5
Solving	All (n=45)	0.1	1.0	0.8	4.7
	Low only (n=21)	0.0	0.2	0.5	0.9
	High only (n=24)	0.3	1.7	1.2	8.0

Table 5-4 Activity over term

Group	Immediate exam scores (Exams 1-4)	Example opens on early topics *before* exams	Example opens on early topics *after* exams	Early (immediate) exam scores on early topics	Delayed exam scores on early topics	Delayed minus early
Control section	69.3 (n=55, sd=12.6)	n/a	n/a	72.9 (n=59, sd=13.7)	65.1 (n=55, sd=18.9)	−8.0 (n=55, sd=17.8)
Examplify section	70.1 (n=86, sd=12.3)	…	…	74.3 (n=91, sd=14.1)	68.9 (n=88, sd=19.9)	−5.4 (n=87, sd=18.0)
- *Never opened*	57.9 (n=2, sd=22.5)	0	0	59.5 (n=3, sd=18.8)	68.7 (n=3, sd=22.1)	9.1 (n=3, sd=7.3)
- *Nonsolving*	68.2 (n=44, sd=12.5)	0.26 (n=46, sd=.77)	0.37 (n=46, sd=.77)	72.0 (n=45, sd=14.8)	62.3 (n=43, sd=20.0)	−10.5 (n=43, sd=19.5)
- *Solving*	72.7 (n=40, sd=11.1)	1.1 (n=45, sd=2.9)	0.6 (n=45, sd=1.2)	77.7 (n=43, sd=12.0)	75.7 (n=42, sd=17.6)	−1.2 (n=41, sd=15.1)

5.5.2 H-immediate

Students with Examplify with solving interaction score higher on immediate assessments.

The variables being compared are summarized in Table 5-4. The "Immediate exam scores" is the average of scores on the four non-cumulative exams (E1-4). The "Delayed exam scores on early topics" is the average score on the half of Exam 5 that was on earlier topics, scaled to 100. A regression model predicting the immediate exam scores takes into account the section (p=.111), whether they ever opened the tool (p=.037), the assigned Examplify variant (n.s.) and its interaction with having ever opened the tool (F(1,125.2)=4.3, p=.052). (Section differences, freshman status and mastery-avoidance, were not significant.) Students with the solving variant scored significantly higher across immediate assessments than the nonsolving control variant (F(1,147)=5.2, p=.024, d=.35) in a contrast test of tool variants among students who opened it. So while there was effect by merely which tool was assigned, there was an effect of which tool the student ever saw. A simpler model comparing immediate exam scores only among students who opened the tool (n=90) also shows that students seeing the solving variant scored higher than students seeing the nonsolving variant (F(1,87.3)=5.4, p=.023, d=.36).

Students with access to Examplify with solving scored marginally higher than BAU control section students on immediate assessments (F(1,145)=3.1, p=.082, d=.26) in a contrast test (Figure 5-10). There was no significant difference between students in the nonsolving control and the BAU control.

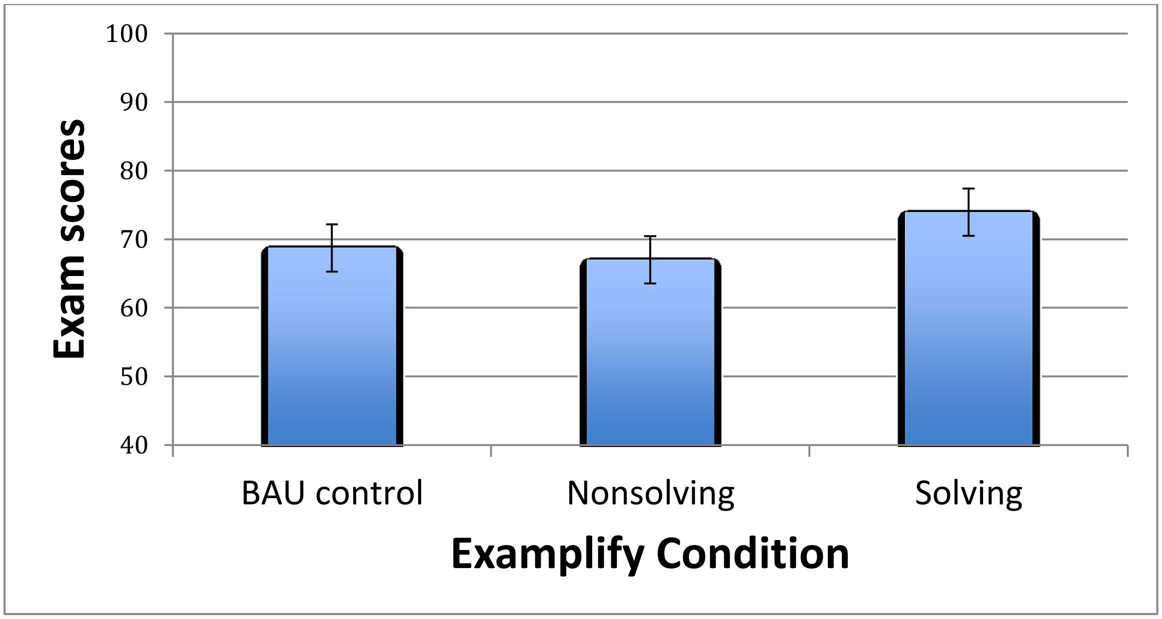


Figure 5-10 Average exam scores by condition showing solving variant of Examplify leading to better exam scores on immediate assessments than both the nonsolving variant and business-as-usual controls

5.5.3 H-delayed

The benefits of Examplify with solving will be greater on delayed assessments than immediate assessments.

To assess robust learning, I predict each student's score on the delayed (final) exam with their earlier average score on those same topics as a covariate (p<.0001) and whether they had access to solving interactive examples, nonsolving interactive examples, or BAU static examples. Students with access to solving interactive examples in Examplify scored higher on the delayed assessment (p=.012) than students with nonsolving examples in the same section (d=0.48) and students with access to only traditional examples in the BAU section (d=0.44). The nonsolving and BAU were so similar that their regression lines practically overlap (Figure 5-11).

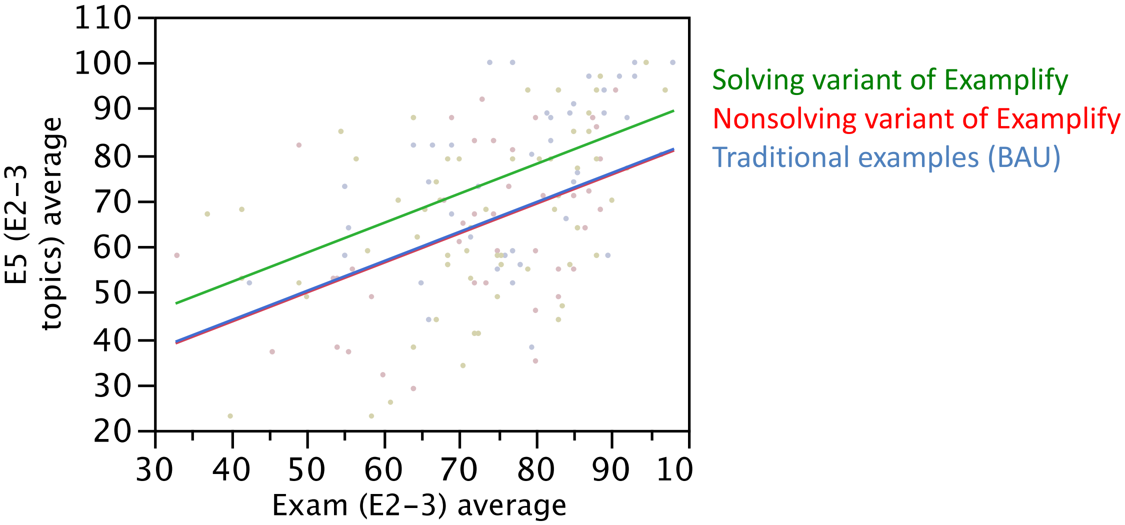


Figure 5-11 Prediction of delayed exam scores by type of study examples available with earlier scores on same topics as covariate. Students with access to solving variant of Examplify performed significantly better on delayed assessments, taking into account their earlier (superior) performance.

It is not clear whether the differences between solving and nonsolving are due to how students studied from examples before the early assessment or in the interim between the early and delayed assessment. To help answer this, I restrict the analysis to students in the Examplify section who ever opened Examplify, meaning they could have been affected by the variant offered, and look at the interaction of condition with whether they accessed the tool before taking the earlier (Exam 3) assessment. In this model I predict delayed assessment score from early assessment (p<.0001), the assigned Examplify variant (p=.003), whether they opened the Examplify tool before Exam 3 (p=.197) and its interaction with variant (p=.064) and freshman status (p=.007). Among students who ever opened the tool, opening it before Exam 3 was linked to better performance if they were in the solving condition (F(1,77)=6.5, p=.013), predicting a score over a letter grade higher that students in the solving condition who didn't open it until later (d=0.68). Opening it earlier or not made no difference among students in the nonsolving control condition (p=.73).

One complication is that some scores are missing. A visual examination of the data for students who missed some exams does not reveal any outliers or differential attrition. I tested whether having missed an earlier exam was motivation for performance on the delayed exam (which can replace an earlier grade) but it made no difference.

5.5.4 Post-hoc: Mechanisms

What explains the higher scores on immediate and much higher scores on delayed assessments with the solving variant of Examplify? Are the mechanisms cognitive or motivational? In medical terms, is the medicine more effective or just better tasting? It may be both.

As described in "Examplify usage" above, solving students opened more examples and the gap widened over time. This suggests students perceive a greater utility in using it. But is that because it's more pleasant than the alternatives ("better tasting") or they perceive it more as a good use of their time ("more effective per dose")?

I would like to test the effectiveness per dose by looking at the performance outcomes from usage, but greater usage can be an indication of greater need because it is a self-allotted dose (i.e., medicine consumed more to treat more severe symptoms). To control for need, I again use the early topics to compare immediate and delayed scores.

Solving led Examplify users to open marginally more examples on early topics before those exams (F(1,88)=3.5, p=.064) and more opens are correlated with higher scores on those exams (r=.25) for both solving (r=.31) and nonsolving (r=.15).

As a loose measure of robustness I consider the delayed measure minus the immediate. For solving, this difference is more correlated with opens before the immediate exam (r=.17) than after (r=.07). For nonsolving, opens before do not demonstrate robustness (r=.017) while opens after do (r=.28). This suggests that for nonsolving the difference is accounted for by studying the topics later, while for solving the difference is better accounted for by studying the topics earlier. This lends support to the interpretation that the gains of Examplify with solving on delayed assessments are due in part to a better effectiveness per dose. In other words, compared to business-as-usual or nonsolving, studying using Examplify with solving improves retention of the studied material.

5.5.5 Student perceptions

In the final questionnaire, students were asked, "If the email reminders were a person, what kind of person would it be?" One positive theme was that of a close and helpful friend. E.g. "my boyfriend" and "studious friend". More often students describe Examplify as someone expert in chemistry. E.g. "A helpful, sympathetic older student"; "Someone who's been doing chemistry for a long time and understands how to look at and start any given chemistry problem"; and "A very knowledgeable and helpful person. I would love them forever. I would want to study with them all the time." While they found the system to be helpful, many were frustrated by its limitations,

If the past exams archive were a person they would be a know-it-all who was always willing to answer questions. They would be there to show you what to do, but they could not really explain it - they just knew the answer. Helpful, but sometimes frustrating.

One particularly colorful account portrayed the solutions without explanations as snake-oil:

The exams archive would be a morbidly obese Wild West snake oil salesman. Horrendously bloated with year after year of exam, you have to wonder how much that guy ate (or maybe it was hormonal?). Sometimes his solutions work, sometimes they're even what you expected, but they come with absolutely no explanation and you're left with an unending suspicion that you're being bullshitted. But he's the only medicine man in town, and if you take enough of his treatments they seem to work, so you just keep investing in them more heavily. Sure, you could go to the local Wild West barber ([the course professor]) for an operation, but his explanations don't make any more sense.

That was a really weird metaphor but I think you get the point.

It seems the explanations were particularly desired for the multiple-choice or fill-in-the-blank responses for which the process of arriving at the answer wasn't apparent:

they are a person with not many friends because very few people have written anything in the hints section so that part is not very helpful even though I wish it were because at times I didn't understand a multiple-choice answer and wanted an explanation but there were none.

This last quotation points to the problem with the explanations. Very few were written. I discuss this limitation and possible remedies below. One encouraging repeated sentiment was that despite the system's flaws, they did value it and had patience for "a child who is still constantly growing". Another student wrote:

I'm not sure what this question is looking for, but I'd say it's a very nice, clean cut person who would go out of their way to help you more often than not. Everyone slips up occasionally and is wrong about one thing or another, but overall, I think I love this person.

5.6 Discussion and Conclusion

Examplify was designed to improve learning outcomes in university lecture courses using observations from the field and theories from existing Cognitive Science literature. In a large introductory chemistry course, students with the solving variant of Examplify performed better on immediate assessments than both the nonsolving control variant and the business-as-usual control section. The benefits on delayed assessments were even greater, about a full letter grade. What can explain these effects? First I contrast the two variants of Examplify.

Students with the solving variant opened more examples, pointing to a motivational effect. They used Examplify more throughout the term and their usage increased by more in the 4th quarter when the system became easier to navigate. So, part of the explanation is that Examplify with solving motivated more studying than the nonsolving variant.

However, usage factors do not entirely account for the differences in performance. The solving variant may also help students to get more out of opening each example, a cognitive effect. Rigorously determining whether this is true is not possible with this study design because a student's open rate is confounded with their study beliefs and self-assessment. However, the regression model predicting delayed scores from their condition and its interaction with whether students opened the tool early also lends support to this interpretation. In the nonsolving variant, opening the tool early made no difference and for the solving variant it related to better delayed performance. Additionally, the correlations in the ad-hoc analysis suggest that Examplify with solving leads to better retention than how students otherwise study.

The study did not have measures of student activity outside of Examplify, where students likely spent the majority of their study time. Another explanation for the benefits of solving over nonsolving are their relative impact on students study beliefs and dispositions, a metacognitive effect. For example, the solving interaction may have increased students' awareness of their readiness for the exam and the nonsolving interaction may have induced a false sense of readiness.

The solving variant of Examplify was better than the nonsolving version and the business-as-usual control, but explaining the differences with BAU is harder because there are more differences and less data to explain them. Examplify may have been more engaging and motivating than the BAU static examples, reduced cognitive load through the step-wise modularization, or any of the solving/nonsolving possibilities. This study was not designed to discern between these and I encourage future work on these questions.

The data I do have comparing Examplify and BAU are partially confounded. Because the two section conditions were not randomly assigned, I cannot be certain that the differences were due to the Examplify treatment but the analyses did factor in key differences between the sections when they were significant. The differences were only marginally significant and were not significant on the delayed measure. One interpretation of the data is that the BAU condition was much like the nonsolving condition. The regression model of the delayed assessment supports this interpretation, through the almost identical parameter estimates for those two conditions. Only students with the solving version of Examplify scored differently on the delayed assessment when accounting for their scores on the earlier one.

Overall the immediate and delayed measures seen against both controls provide evidence for the positive benefits of Examplify with solving. It appears that the solving variant of the tool was motivating to students and led them to learn in more robust ways from their studying with it. The nonsolving variant, having similar outcomes to the BAU section, may closely match how most students study from the worked example problems without the tool. That is, by examining the expert's worked solution and explaining it to themselves rather than first attempting to do the work. In this interpretation, studying with Examplify with solving could be improving their awareness of the skills they need to develop or directly increasing their fluency through practice.

How to reconcile these results with Cognitive Load Theory? The solving variant prompted students to solve before they ever studied an example, which would be predicted to increase cognitive load and yield poorer learning, instead of the very positive effects observed. There are a number of possible explanations for this apparent contradiction with Cognitive Load Theory. The first could be the expertise reversal effect. In this interpretation, with the nonsolving variant of Examplify, students continued to study worked examples even when they had such expertise that problem solving was more appropriate. While this may be true, it doesn't appear to be an effect of the tool as there were no significant differences between the nonsolving variant and business-as-usual. To validate this explanation, future work should test in isolation supporting students' transition to problem solving.

A second, and compatible, explanation, is that the solving variant of Examplify requires much less extraneous cognitive load than the problem solving conditions to which worked examples have been compared in previous studies. The problem solution feedback in most worked example studies is presented only after a student attempts the whole solution, subjecting them to possible floundering and, indeed, extraneous cognitive load. In the Examplify solving variant, the feedback is given after each step. Experiments with Cognitive Tutors, which share this step-wise feedback feature, have found a reduction in the worked example benefit (Salden et al., 2009). Potential extraneous load is further reduced in the Solving variant because instead of having a succession of hints about the next step as in Cognitive Tutors (which start off quite vague and may invoke extraneous load before finally getting to a worked-out example of the next step), in the solving variant students can go directly and quickly to the worked step if they choose. An analysis of how quickly students reveal the next step in the solving variant should be pursued in future work.

Examplify not only has strong positive impacts on learning, but is easy to adopt. The benefits to student learning required very little time from the instructor and no changes to his curriculum. All that was needed was spending one minute per page marking the static images for the interactive activities. In a course with similar exams, a teaching assistant could prepare 15 old exams in one hour, or about how long they have office hours each week. Because the markup does not require domain knowledge, it could be done by a work-study student or even outsourced to a micro-labor market such as Amazon Mechanical Turk. For other courses, getting old exams into a digital form may be a bigger task. Scanning a stack of papers is fast, but for instructors without a scanner a phone camera is an increasingly practical option.

For instructors who do not have high quality camera phones, their students may and could be incentivized to both snap photos of old exams and mark them up for interactivity.

Examplify is a simple technology that can provide big gains to learning. In a full semester evaluation in a real-world college course, Examplify with solving improved exam scores and had even greater gains on the delayed measure, suggesting benefits on longer-term learning. As a benefit to future related work, the techniques used by Examplify are drawn from cognitive psychology and are simple to implement and iterate upon.