

Improving Students' Study Practices Through the Principled Design of Research Probes

Turadg Aleahmad

May 7, 2012
CMU-HCII-12-105

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

Kenneth Koedinger (Co-Chair)
John Zimmerman (Co-Chair)
Vincent Aleven
Mark Guzdial (Georgia Institute of Technology)

*Submitted in partial fulfillment of the requirements for the Degree of Doctor of
Philosophy.*

Copyright © 2012 Turadg Aleahmad, Some Rights Reserved.
Licensed under a Creative Commons BY-NC-SA 3.0 License.

This work was supported in part by a Graduate Training Grant awarded to
Carnegie Mellon University by the Department of Education (R305B040063), a
National Science Foundation award (OMA0836012) and a Spencer Foundation
grant (200900164).

Keywords

Research through design, design-based research, learning sciences, usable knowl-
edge, procrastination, time management, achievement goals, metacognition,
testing effect, cognitive load theory, worked examples, prompted self-explanation,
operant probe, in vivo experimentation, educational technology.

Abstract

A key challenge of the learning sciences is moving research results into practice. Educators on the front lines perceive little value in the outputs of education research and demand more “usable knowledge”. This work explores the potential instead of usable artifacts to translate knowledge into practice, adding scientists as stakeholders in an interaction design process. The contributions are two effective systems, the scientific and contextual principles in their design, and a research model for scientific research through interaction design.

College student study practices are the domain chosen for the development of these methods. Iterative ethnographic fieldwork identified two systems that would be likely to advance both learning in practice and knowledge for applying the employed theories in general. Nudge was designed to improve students’ study time management by regularly emailing students with explicit recommended study activities. It reconceptualizes the syllabus into an interactive guide that fits into modern students’ attention streams. Exemplify was designed to improve how students learn from worked example problems by modularizing them into steps and scaffolding their metacognitive behaviors through problem-solving and self-explanation prompts. It combines these techniques in a way that is exceedingly easy to author, using existing answer keys and students’ self-evaluations.

Nudge and Exemplify were evaluated experimentally over a full semester of a lecture-based introductory chemistry course. Nudge messages increased students’ sense of achievement and interacted with students’ existing time management skills to improve exam grades for poorer students. Among students who could choose whether to receive them, 80% did. Students with access to Exemplify had higher exam scores ($d = 0.26$), especially on delayed measures of learning ($d = 0.40$). A key design decision in Exemplify was not clearly resolvable by existing theory and so was tested experimentally by comparing two variants, one without prompts to solve the steps. The variant without problem solving was less effective ($d = 0.77$) and less used, while usage rates of the variant with problem solving increased over time.

These results support the use of the design methods employed and provide specific empirical recommendations for future designs of these and similar systems for implementing theory in practice.

Acknowledgements

This work is my attempt to make a dent in the boundary of human knowledge. As challenging as it has been, this page may be the most difficult of all to write. I feel deeply grateful to innumerable people who have shaped my desire and ability to reach this point, by inspiring me forward, helping light my path, and even those who have helped me to see where not to tread. To all those people: thank you.

When I started this PhD program at Carnegie Mellon, I was a passionate dilettante with little understanding of research, how to discover the boundaries of human knowledge and how to push them. I did not know yet how to see further by standing on the shoulders of giants. Instead, I was lost among them. To my early advisors, Vincent Aleven and Bob Kraut, thank you for your patience while I explored where to push and challenging me to find my place. Foremost, thank you for compassion and support in my late-stage transition to a new direction of research.

To my co-chair, Ken Koedinger, thank you for your openness and mentorship in many forms, always within my zone of proximal development. You are a master of the *assistance dilemma* and an inspiration for putting learning sciences out into practice. To my second co-chair, John Zimmerman, thank you for opening my eyes and ears to Design and how to secure a role for empathy in my research. Your perspectives, flip-side questions, bon mots and coffee chats have been invaluable. To my other committee members, Mark Guzdial and, again, Vincent, thank you both for your insightful analysis of the work and thorough suggestions that improved it.

Thanks to HCII, PIER and PSLC for vibrant intellectual communities. Sharon Carver, Scott Davidoff, Paul Karol and Marsha Lovett helped to shape the work described here. Thanks to many colleagues and friends who have supported my and its development along the way: Aruna Balakrishnan, Moira Burke, Matt Easterday, Chris Harrison, Matthew Lee, Min Kyung Lee, Derek Lomas, Ilya Goldin, Gahgene Gweon, Johnny Lee, Ian Li, Ryan Muller, Timothy Nokes, Amy Ogan, Nora Presson, Kelly Rivers, Ido Roll, Christian Schunn, Howard Seltman, Eliane Stampfer, Caitlin Tenison, Nathan VanHoudnos, Jason Wiese, Erin Walker, and Ruth Wylie.

Jo Bodnar, Queenie Kravitz, Melissa Carrozza, and Audrey Russo, thank you for helping with all the logistics and for doing so with a smile. Sally Wu, thank you for being the smartest, hardest working, and nicest intern I could hope for. Jim Slotta, Marcia Linn, Anthony Perritano and the rest of the TELS and Educoder crews, thanks for supporting me in ed-tech hacking and growing into a researcher.

Finally, thanks to my family. Linda, Iradg and Sima, I learned from you to always live passionately and towards the benefit of all. A very special thanks to my betrothed, Kristie Boyce, for walking into my in life that day and helping make this whole journey a thoroughly rewarding experience.

Contents

| | |
|-----------------|----------|
| Keywords | 1 |
| Abstract | 2 |

| | |
|--|-----------|
| Acknowledgements | 2 |
| 1 Introduction | 6 |
| 1.1 Motivation | 6 |
| 1.2 Scientific Research through Interaction Design | 7 |
| 1.3 Operant Probes | 8 |
| 1.4 Process | 9 |
| 2 Operant Probes for Scientific Research through Interaction Design | 11 |
| 2.1 Introduction | 11 |
| 2.2 Design-based Research in Education | 12 |
| 2.3 Scientific Research Through Interaction Design | 13 |
| 2.4 Operant probe as research artifact | 15 |
| 2.5 Value for Research | 17 |
| 2.6 Problems in education research | 18 |
| 2.7 Limitations | 22 |
| 2.8 Design process | 23 |
| 2.9 Evaluating an operant probe contribution | 24 |
| 2.10 Conclusion | 25 |
| 3 Discovering SRtID Opportunities in College Lecture Courses | 25 |
| 3.1 Introduction | 25 |
| 3.2 Methods for Discovering Opportunities | 26 |
| 3.3 Focus and context selection | 28 |
| 3.4 Semi-structured interviews | 28 |
| 3.5 Interpretation | 29 |
| 3.6 Theory in Context | 32 |
| 3.7 Ideation | 38 |
| 3.8 Needs Distillation | 40 |
| 3.9 Scientific Impact Evaluation | 40 |
| 3.10 Needs Validation | 41 |
| 3.11 Needs Selection | 43 |
| 3.12 Conclusion | 46 |
| 4 Nudge: Supporting Students' Study Time Allocation | 48 |
| 4.1 Introduction | 48 |
| 4.2 Background Theory | 48 |
| 4.3 Core Features | 49 |
| 4.4 Iteration | 50 |
| 4.5 Experimental Design | 55 |
| 4.6 Results | 58 |
| 4.7 Limitations and Opportunities | 65 |
| 4.8 Conclusion | 67 |
| 5 Exemplify: Enhancing Worked Examples for Better Learning | 67 |

| | | |
|-----------|---|------------|
| 5.1 | Introduction | 67 |
| 5.2 | Background Theory | 68 |
| 5.3 | Core Features | 70 |
| 5.4 | Experimental Design | 81 |
| 5.5 | Results | 84 |
| 5.6 | Discussion and Conclusion | 91 |
| 6 | Summary and Conclusions | 93 |
| 6.1 | Introduction | 93 |
| 6.2 | Process Overview | 94 |
| 6.3 | Nudge | 95 |
| 6.4 | Exemplify | 97 |
| 6.5 | Scientific Research through Interaction Design | 101 |
| 6.6 | Final Thoughts | 105 |
| 7 | References | 106 |
| 8 | Appendix A: Output of ideation | 115 |
| 9 | Appendix B: Early design areas for lecture courses | 118 |
| 10 | Appendix C: Scenarios developed for Needs Validation | 119 |
| 11 | Appendix D: Scenario sketches used in Needs Validation | 135 |
| 12 | Appendix E: All tasks reminded or polled by Nudge | 151 |

1 Introduction

Developments from the learning sciences move slowly, if at all, into educational practice. Consider the low adoption of spaced practice of learned material, a robust finding first observed by Hermann Ebbinghaus in 1885 and validated at scale in classrooms with thousands of students in 1939 (Ebbinghaus, 1913; Spitzer, 1939; 1939; Whitehurst, 2003). Students still cram. Teachers still march linearly through curriculum and rarely repeat assessments. Why is it so difficult to implement research findings into practice? Burkhardt and Schoenfeld argue that “part of the reason is that the traditions of educational research are not themselves strongly aligned with effective models linking research and practice” (2003).

The field of human computer interaction research, like many applied sciences, has also grappled with the gap between research and practice (Buie et al., 2010). In this dissertation I draw on methods common in HCI to describe an emerging model of linking research to practice in education: the operant probe. I adapt the HCI methods user experience design to an education-focused *learner experience design*, first exploring learner experiences through sketching to map out design opportunities. Then from this map I designed two operant probe systems to support the practice of studying: Nudge for allocating study time and Exemplify for better learning from example solutions. In the following chapters, I describe the potential benefits of operant probe development, the fieldwork that inspired the systems, the design iterations to create them, and the formal evaluation to rigorously validate the design decisions. I then return to examining the process. But first, I elaborate how research in education can benefit from new design processes.

1.1 Motivation

1.1.1 Education is important to improve

Quality education is critical to modern society. The cognitive skills of a population are powerfully related to their individual earnings, distribution of income, and economic growth (“The Role of School Improvement in Economic Development,” 2007). While continued economic growth requires growth in cognitive skills, the United States educational system is in decline. In 1983, the Reagan administration published A Nation at Risk

(United States National Commission on Excellence in Education, 1983) and in 2012 the Council on Foreign Relations reports that the poor state of American education threatens not just US prosperity but its national security (Klein & Rice, 2012). The comparative decline of US education is most acute in higher education. While America is ranked first in the world in college degrees per capita for people aged 55 to 64, for ages 45 to 54, it is third and in ages 25 to 34, it has fallen to 10th place globally. While this is a decline in rank, it is mostly due to other countries improving their education while the US has been stuck at

40% college completion for decades, despite large scale reforms such as the No Child Left Behind act (*No Child Left Behind Act*, 2002).

1.1.2 Education is difficult to improve reliably

Part of the challenge is how difficult it is to know what proposed changes work in practice. This is due in part to the history of education research in the US. Quoting from the U.S. Department of Education’s Strategic Plan for 2002–2007 (2002, cf. Burkhardt & Schoenfeld 2003):

Unlike medicine, agriculture and industrial production, the field of education operates largely based on ideology and professional consensus. As such, it is subject to fads and is incapable of the cumulative progress that follows from the application of the scientific method and from the systematic collection and use of objective information in policy making. We will change education to make it an evidence-based field. (p. 48)

Since then the federal Institute for Educational Science, established by the No Child Left Behind Act, has fostered an emphasis on determining “what works”. They have defined a gold standard for research, randomized controlled trials. Yet, many of the studies they fund fail with “no effects”. Researchers and other experts question the design of these studies, running up to \$14.4 million. Scholars worry that even when study results are positive, they do not carry over into other educational settings. Policy makers want to know, “What will work in my school?” (Viadero, 2009).

Part of the difficulty in reforming education research is the tension between understanding how learning works, understanding how to improve learning, and actually improving learning. Let us return to the example of spaced repetition. Why isn’t it used more in practice? We understand learning enough to know that distributed (spaced) practice has better long term learn effects than massed practice (Committee on Developments in the Science of Learning, 2000; Karpicke & Blunt, 2011). We also know that we could improve learning by increasing the amount of distributed practice by students. However, the next step is the hard part. Students know massed practice by another name, cramming, and when they are told it is not the best strategy for studying, they continue to use it. The classroom environment makes that the easiest option for most students. It doesn’t require planning or self-regulation; it let’s them avoid confronting the limits of their knowledge; it isn’t facilitated by the instructor; and it doesn’t affect their grades much as they still do as well on the assessments they’re given.

1.2 Scientific Research through Interaction Design

How do you replace cramming with more effective study strategies? The answer is not a matter of just science but also of design. Science is powerful because of its ability to generalize, through nomothetic descriptions. In this case, the science falls short of solving the problem of shaping student learning. To

change a student's study strategies requires recognizing and fitting with the fuzzy factors that influence that student's behavior in her specific environment. Design is an idiographic tradition, which tends to specify and understand the meaning of contingent phenomena in order to change a current state of the world to a preferred state. To use the power of scientific theories of learning to improve actual learning requires creating working solutions that operationalize basic research into practice. To evaluate the operationalization requires then evaluating those artifacts and their features for their ability to effect the desired changes. Through many specific designs, generalizations can be developed for normalized solutions and theories of implementation. I describe this approach as Scientific Research through Interaction Design, building on the concept of Research through Design in the HCI literature (J. Zimmerman, Forlizzi, & Evenson, 2007).

1.3 Operant Probes

Further, we as a field need these artifacts to be designed to operate in the "real world", as Fishman et al. contend:

Why are cognitively oriented technology innovations not widely used in schools? Why aren't they scaleable or sustainable? We believe an underlying explanation to be that we, as a scholarly community, have not focused our research on the development and use of cognitively oriented technologies in a way that addresses the fundamental needs of school systems. Instead, research on cognitively oriented learning technologies has focused primarily on students, teachers, and classrooms as the primary unit(s) of analysis. Though we recognize the need to link technology and reform, the field lacks a bridge between focused research and development of learning technologies and the broad-based systemic use of these innovations in schools. Shepard (2000) recognized this as problem for the broader educational research community in her AERA Presidential Address, when she advised researchers to develop methodologies that embrace "dilemmas of practice." Such work "would advance fundamental understandings at the same time that they would work to solve practical problems in real-world settings" (p. 13). This focus would lead to the production of more readily "usable knowledge" (Lagemann, 2002). As researchers, we have developed rich understandings of how technology can foster learning in specialized situations; we now need to develop knowledge about widespread appropriation and use of cognitively oriented technologies by schools and school systems as part of real-world reform efforts. (Fishman, Marx, Blumenfeld, Krajcik, & Soloway, 2004)

In this dissertation I developed two cognitively oriented learning technologies that operationalize theory into real-world contexts. The resulting designs are not just "usable knowledge" but "usable systems". Moreover, their use serves to help

inform scientific theory. I classify them as “operant probes”, a term and type of research contribution that I motivate and define in Chapter 2. Operant probes form part of an emerging paradigm of research, using web technologies and scale to design research artifacts that operate *in vivo* and provide the controls and data collection needed for rigorous quantitative research.

1.4 Process

This work explores methods for designing operant probes in the approach of Scientific Research through Interaction Design, adapting HCI methods to education research. The process can be organized by the stages of Bannan-Ritland’s Integrative Learning Design Framework (Bannan-Ritland, 2003). In this framework, the design process begins with Informed Exploration of the design context, followed by Enactment and then Evaluation of Local Impact. Finally, the design may be evaluated for Broader Impact. The designs in this work have not yet been evaluated for Broader Impact but I will speak to the aspects that support confidence in their suitability to transfer into other contexts.

1.4.1 Informed Exploration

To develop systems that fit into real-world contexts requires a rich understanding of those contexts. Because my goal is to develop new types of probes to open new opportunities for research, I begin with an exploration of opportunities in a specific context. Finding opportunities does not require a formal method, but here I offer a reliable method, adapted from the validated best practices of HCI, to rapidly identify new opportunities. I extend HCI user experience design to the unique challenges of *learner experience design*, evaluating these opportunities by both learner impact and contributions to accretive education research.

I chose large college lecture courses as the context of inquiry. Lecture courses are presently the dominant way that the 20 million college students in the US are taught. As a centuries-old mode of instruction they are ripe for innovation. The goal here was to find opportunities to directly improve learning for a large number of students. This opportunity mapping approach to the Informed Exploration is fully described in Chapter 3. In this phase I identified two opportunities for which to design: 1) helping students to better allocate their study time and 2) provide students with more immediate feedback on their learning.

1.4.2 Enactment

With the two design goals, I iteratively developed two software systems, Nudge and Exemplify. Nudge (attempts to improve how student allocate their study time by decomposing the course syllabus and adding explicit tasks with due dates (such as “Study for the upcoming exam” one week before it takes place.) The small tasks are sent to the student by email when they are due and students can indicate their progress on the task. Exemplify enhances traditional answer keys with an interactive activity to scaffold how students learn from them.

From storyboards evaluated through user interviews to working prototypes evaluated in pilot classroom trials, each design decision was weighed between its ability to operate on the environment to achieve the desired outcome and to probe the environment and its use to advance the science of learning. As operant probes, part of the design process was to refine an understanding of the implications and limits of the evidence in general theory and local observations for each design decision. When design questions could not be satisfactorily resolved by general theory or local observations, they were identified as candidates for resolving empirically using probe variants. In this work, a tension was identified in the practical recommendations of the literature on the “worked example effect” in cognitive load theory versus the “testing effect” in memory theory. To help resolve this both for the design of this system and future similar systems, a variant of Exemplify was produced for each of the two competing predictions of the theories in practice. In one variant of the tool, the problem-solving prompt was removed so that it was strictly a worked example. The full iterative development process of each is described in the chapters on Nudge and Exemplify (Chapters 4 and 5).

1.4.3 Evaluation

With the two operant probes fully developed, the next phase was to evaluate their efficacy through an experiment. The experiment took place in a large college introductory chemistry course consisting of two similar sections. The larger section was chosen as the Experimental section and received both Nudge and Exemplify while the other, control section, had neither. Within the experimental section, students who opted into the study were randomly assigned to a Nudge condition (Nudge required or unavailable) and an Exemplify condition (including prompts to solve or not). Nudge and Exemplify both fit well to the context. Students used both systems voluntarily through the whole semester, including students not in the study.

Both Nudge and Exemplify affected student learning measures. Nudge interacted with students’ time management skills to better aid students with worse time management. Exemplify provided big gains on robust learning, supporting the testing and proceduralization over worked example effects in practice. On immediate measures, students with the variant that prompts to solve performed better than students both with the nonsolving control variant ($d = 0.35$) and business-as-usual control section ($d = 0.26$). On delayed measures, the effect was roughly a full letter grade over the nonsolving control variant ($d = 0.77$) and business-as-usual control section ($d = 0.40$).

On the data production measures, Nudge provided data on student activities that could be used to model student study practices. Exemplify logs provided data that helped explain the mechanisms of its effects. The full analyses of each system are detailed in their respective chapters (4 and 5).

1.4.4 Wrap-up

I do not set out in this work to answer a specific theoretical question. Instead, this work is to improve the practice of education by designing 1) technological artifacts that enact learning science principles to effect learning objectives in a specific natural context and 2) evaluations of the artifacts to inform future applications of said principles and the principles themselves.

In doing so, this work contributes to processes of design research in education and to specific design implications for two classes of technologies for education. The primary contribution to design research broadly is the articulation of the operant probe as a productive research artifact and the scholarship to situate it within existing design practices and research issues (Chapter 2). A related contribution is the reflection on the learner experience design methods suitable to designing operant probes, chiefly mapping of opportunities (detailed in Chapter 3). With these processes, I developed two systems that operationalize and thus inform the real-world application of theory. Nudge informs the potential for supporting student time use and the contextual utility of theoretical principles of motivation (Chapter 4). Exemplify informs the potential for supporting students learning from worked solutions and the contextual utility of theoretical principles of cognitive load, practice and proceduralization. Further, the evaluation of Exemplify experimentally measures the relative utility of competing theoretical principles when put to use (Chapter 5).

I argue that through an analysis of existing methods and reflection on my design processes and artifacts, I will demonstrate a new and effective approach to design research in education. Stated explicitly:

The Scientific Research through Interaction Design approach can enact preferred states in a manner that explains outcomes, informs the conditions for applying scientific theory, and generates new experimental hypotheses.

In Chapter 6, I evaluate the success and limitations of this work in supporting this thesis statement by reflecting on the cases of Nudge and Exemplify.

2 Operant Probes for Scientific Research through Interaction Design

2.1 Introduction

A persistent issue in education research is the question of what its relevance is to actual educational practice. Burkhardt and Schoenfeld argue that “part of the reason is that the traditions of educational research are not themselves strongly aligned with effective models linking research and practice” (2003). Shepard, in her AERA Presidential Address, called for the field to develop methodologies

that embrace “dilemmas of practice.” (Shepard, 2000) Leaders in the field call for “usable knowledge” (Lagemann, 2002) using these methods.

Knowledge is usually communicated through media (text and images), but knowledge also lives in and is communicated by designed functional artifacts (Cross, 1999). Modern computing makes it easy to share not just bits that represent words but also bits that represent interactive artifacts (i.e. software). Fishman et al. call for design knowledge to make software that is cognitively oriented, scalable and sustainable (Fishman et al., 2004). This work posits that developing this knowledge base is promoted by targeting the design of “usable artifacts”. These are research artifacts that can cross the chasm into practice or be adapted by practitioners. I explore the potential of such artifacts to promote research and illustrate the opportunity for a new type of situated research artifact, the operant probe.

2.2 Design-based Research in Education

Education is fundamentally a design endeavor. Adopting Simon’s definition of design, “transformation of existing conditions into preferred ones” (Simon, 1969), all facets of education are design: e.g. teaching improves conditions of learners’ minds; better instruction improves conditions of teaching; better technology and research improves instruction; better public policy improves all of the above.

To help close the gap between research and practice, education researchers in the early 90s began “design-based experiments” wherein they would iteratively and reflectively prototype interventions in classrooms (Brown, 1992; Collins, 1992). This grew into the Design-based Research methodology and movement, The Design-based Research Collective (Design-Based Research Collective, 2003). However there are numerous conceptions and splinter methodologies and terminologies. Two excellent surveys are those of Mor & Winters and Wang & Hannafin (Mor & Winters, 2007; Wang & Hannafin, 2005) and the book Educational Design Research compiles critical essays (van den Akker, Gravemeijer, McKenney, & Nieveen, 2006b). Howley contrasts these conceptions of design within the field of education with design more generally (Howley, 2010). To stay above the terminological morass, I will use her term DBRE to indicate the cluster of design-based research methods in education.

DBRE is motivated by the observation that the direct application of theory is not sufficient to solve the complicated problems of education (van den Akker, Gravemeijer, & Nieveen, 2006a). Instead researchers situate themselves within the context of use and iteratively intervene by reflecting on the situation. These well-documented reflections and iterations form the basis of “humble theories”, which are domain-specific and pragmatic for the activity of design (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003).

DBRE adds an important methodology that was missing from the toolbox of education research (Collins, Joseph, & Bielaczyc, 2004). However since the debut of DBRE in 1992, modern computing has opened new opportunities for

situated interventions. The field of Human Computer Interaction has made greater progress in understanding how to build usable computing systems. I will switch to the lens of HCI to articulate this emerging research paradigm and then return to how it can address some of the remaining gaps in the methodological toolbox of education researchers.

2.3 Scientific Research Through Interaction Design

Interactive computing systems have the potential to improve the quality of education while lowering its costs. The field of human computer interaction has developed to address the considerable complexities in making systems that people can and want to use. Further, the field of HCI, like many applied sciences, has also grappled with the gap between research and practice (Buie et al., 2010).

Because the term “design” is so frustratingly polysemous, I situate this work in a particular framework of “research through design” (J. Zimmerman et al., 2007). RtD describes much of HCI research and draws from Frayling’s distinctions between research *into*, *through*, and *for* art and design (Frayling, 1993). Research through Design defines (i) a model of interaction design research that benefits both research and practice communities and (ii) a set of criteria for evaluating the quality of an interaction design research contribution.

A distinction of RtD from other design methods is the conception of the artifact itself as a research outcome. The artifacts, in transforming the world from its current state to a preferred state, serve as exemplars for HCI design practitioners. While the artifacts themselves communicate design knowledge and facilitate extension of the ideas therein, RtD contributions also describe their process in detail to support practitioners in making similar insights in their own design work. As illustrated in Figure 2-1, through this process the interaction designer can synthesize the knowledge from multiple modes of inquiry into an artifact that passes easily into practice.

RtD distinguishes research activities from practice, in part, by the goal “to make the *right* thing: a product that transforms the world from its current state to a preferred state.” They differentiate research artifacts from design practice artifacts in two important ways:

First, the intent going into the research is to produce knowledge for the research and practice communities, not to make a commercially viable product. To this end, we expect research projects that take this research through design approach will ignore or de-emphasize perspectives in framing the problem, such as the detailed economics associated with manufacturability and distribution, the integration of the product into a product line, the effect of the product on a company’s identity, etc. In this way design researchers focus on making the right things, while design practitioners focus on making commercially successful things.

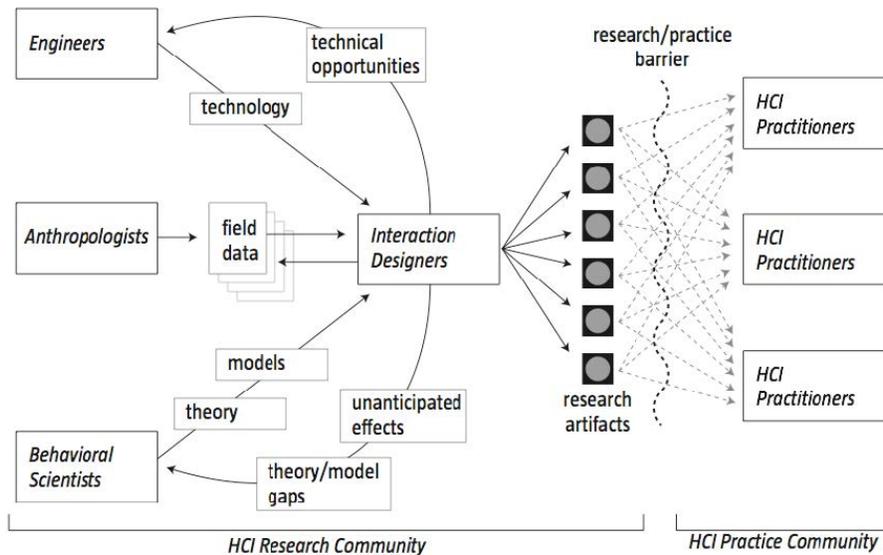


Figure 2-1: Pathways and deliverables between and among HCI researchers and practitioners (from Zimmerman et al. 2007)

This distinction between the *right* and the *commercially viable* highlights an important issue in education research, distinct from HCI practice. While researchers are generally concerned with what is right and good for learners, the communities that produce the products in the educational marketplace are concerned with what creates a perceived value for which the consumer will provide money (or attention for advertising, etc.). In education, this often means that commercial products are adopted that may be viable but not beneficial to learning outcomes, or at least not as beneficial as some *right* but less viable research artifact. However, the public and philanthropic funding of education provides an opportunity to make systems that are both *right* and also *viable* through non-commercial means of distribution and funding.

A second distinction of the education domain is that the most common practitioners, teachers, do not design technologies. While they design experiences of how technologies will be used in their courses, and may re-appropriate technology in creative and inventive ways, they do not and can not be expected to design the technologies themselves. Participatory design methods like co-design draw in education practitioners as actors in the design process (Roschelle & Penuel, 2006), but at some point the system is made and deployed to practitioners who will have nothing to do with its design.

The difficulty in translating research into practice is a great challenge to the education research enterprise. There is a growing literature of “usable knowledge” in the form of practice guides, etc. However not all basic knowledge can take these forms. We also need “usable artifacts” that operationalize this knowledge

into a usable form. By designing interactive usable artifacts, we can bring that knowledge into practice and help to inform the practical constraints of existing scientific knowledge and opportunities to advance it. Within the frame of Research through Design, I call this approach Scientific Research through Interaction Design.

2.4 Operant probe as research artifact

Interactive software systems are an ideal form for operationalizing knowledge in education. They can shape the behaviors of learners and mediate their interactions with learning materials, peers and facilitators. Further, recent advances in computing afford software applications that (i) cost little to develop, deploy and scale; and (ii) provide instrumentation to collect data and run controlled experiments on live systems in natural contexts.

The costs of building web-based software systems are lower than ever. Software standards like HTML5 have driven down the costs of developing for a wide audience. Open-source operating systems and application stacks have driven down the costs of software infrastructure. Commoditization of computing has driven down the costs of hardware infrastructure. Today, one lone developer can make a web-based application, integrated with other services, and serve it to millions of users. The costs to develop and run such systems are miniscule compared to the value they can create. For example, Instagram was recently purchased for \$1 billion and had only 13 employees. Projecting falling prices of server resources, a researcher could leave all their software systems running online the rest of their careers for less than the cost of a conference trip.

Further, these systems can be used by people in real natural settings. Today software is constantly adapting to users and the objectives of its designers. For example, Google monitors everything its users do and make inferences to update their designs. These are small hypotheses about how a change to the product (e.g. autocompletion of search queries) can improve some desired outcome (such as speed to find a satisfactory search result). Often a hypothesis is tested through an A/B test which randomly assigns some users to one variant of the system. In this way Google can develop its theories of its specific product and general theories of user behavior.

The software then is an artifact which operates on the user's environment (e.g. web browser) to achieve an outcome (e.g. fast search) and also probes that environment for data to improve the product's design and more general theories. Let us name this type of artifact an "operant probe". As an outcome of Research Through Interaction Design, an operant probe creates a new space that is both research and practice (Figure 2-2). Research from different modes of inquiry can be brought together to influence the design of an operant probe. The probe, operating in real world contexts, can then influence the learning experience. This influence can be experimentally manipulated to test theories, both humble and

robust. Finally, all the data from its use can be used to model the outcomes and mechanisms. With these features, I offer a formal definition:

Operant probe (n): an in vivo research apparatus that operationalizes theoretical constructs and collects data by which to both evaluate its effects and model the mechanisms.

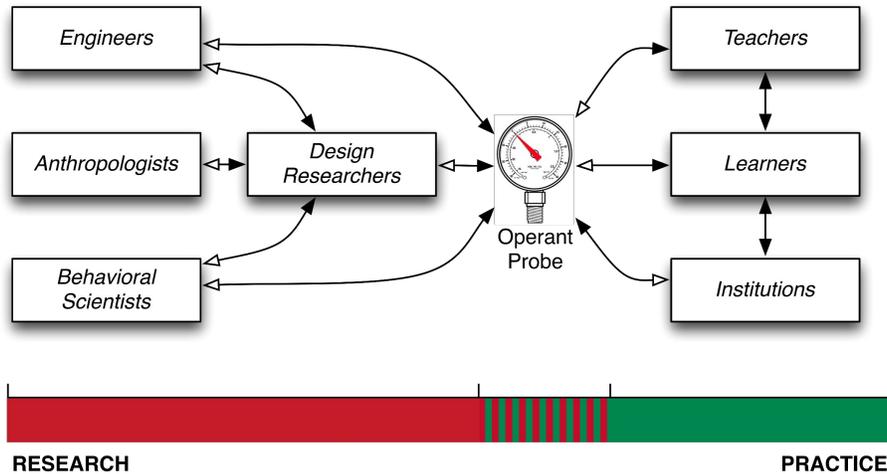


Figure 2-2: Relationship of Operant Probes to communities and contexts

Operant probes are not new to education but there should be more. Intelligent Tutoring Systems make a large class of operant probes. The Cognitive Tutor product from Carnegie Learning could be described as the most successful operant probe in education, operating in thousands of schools and providing millions of data points to improve the product and scientific theories of learning with intelligent tutors. From cognitive tutors a whole field has emerged with variations and enabled exploration and quantitative evaluation of new designs and scientific ideas. Games for learning are a growing class of operant probes, using interaction data to improve the design of the game and sometimes to contribute back to theory. Khan Academy is a popular system that uses interaction data extensively to drive their design, though they haven't yet engaged any scientific research community.

The operant probe concept fits well into the “iterated in vivo experimentation” methodology (E. Walker, 2010). Walker explains that such experiments, “use a design-based research process to create an intervention, deploy the intervention using an in vivo experiment, and then interpret the effects through a design-based lens, may be a more effective way of theory building than executing an in vivo experiment in isolation.” In such experiments, the intervention may be chiefly or entirely an operant probe. So in some sense, “operant probe” is a name for the artifact used in this methodology. But moreover, the design of operant probes is to produce systems that function as well outside the research activity because

they are designed to have apparent value to their users and fit easily into their existing behaviors and constraints. Once in place, a well conceived and designed operant probe needs only minimal intervention in order to provide value to users and researchers.

2.5 Value for Research

Operant probes have specific affordances to the practice of research through design. The requirements of a software system to be considered an operant probe are listed in the first column of Table 2-1, with their benefits to research practice and validity.

Table 2-1: Benefits of operant probes to research practice and validity

| Requirement | Practical Benefit | Relevant Validity |
|--|--------------------------|----------------------------------|
| Low cost and high fidelity of distribution in lab and real-world | high deployability | External and Ecological validity |
| Consistency of intervention, ease of replication | high replicability | External validity |
| Operationalization of theory | high specificity | External validity |
| Instrumentation to provide data to model its use and context | high resolution of data | Internal and External validity |
| Controlled manipulation of its design within and between deployments | high manipulability | Internal validity |

Research framed with operant probes helps fill the gap between descriptive accretive science and interventionist case-based design by supporting endeavors that are both interventionists and accretive. New classes of operant probes can provide learning scientists with new opportunities for real-world impact and new funding sources. They can provide design-based researchers with new tools for building on scientific approaches and rigorously evaluating their designs in many contexts without their active participation. I argue that the design and deployment of new operant probes can improve outcomes, validate the practical import of basic theory, and generate new research questions.

To advance research, probes must fit into their targeted contexts, provide conditions to validate or invalidate research hypotheses, and provide data to inform the selection of competing models explaining the outcomes. Below I survey the prevailing challenges for design-based research in education to highlight the opportunity for operant probes.

2.6 Problems in education research

In the couple decades since the introduction of design-based research in education, challenges in the paradigm have been articulated by an array of researchers. Operant probes provide an opportunity to address some of these problems in education research. This section quotes heavily from other authors to convey the tone of the critiques and variety of voices.

2.6.1 Ill-defined methodologies

One challenge in design-based research is that the methodologies are so shaky. This is not due to any lack of definitions and procedures; in fact there exists a surplus of distinct frameworks and terms (Wang & Hannafin, 2005). This focus on process has led to research that is over-methodologized and under-conceptualized (Dede, 2004).

Operant probes as a design and research activity fit with existing methodologies instead of seeking to supplant them. As an interface between different communities, they have a face in each that can be engaged with and evaluated per the existing norms of those communities without having to invent new evaluation methods.

2.6.2 Design principles have little traction

Design-based researchers extract principles from their design activities. These are modeled after the design patterns of architecture and software development (Mor & Winters, 2007). Some frameworks advocate a hierarchy of patterns as design principles: General Cognitive Principles such as *self-regulation*, Metaprinciples such as “promote autonomy and lifelong learning”, Pragmatic Pedagogical Principles such as “Encourage monitoring”, and Specific Principles such as “Multiple, diverse opportunities for students to reflect on their ideas and create representations of their views” (Bell, Hoadley, & Linn, 2004; Linn, Davis, & Eylon, 2004).

While these principles could be useful to a designer employing them, there is little evidence that education technology designers refer to such principles. Technology designers, like other designers, generally learn from examples, not abstract principles (Cross, 1999). Fortunately, unlike in architecture, in education technology the design artifacts are increasingly easy to share. Operant probes serve as highly visible design examples to learn from and iterate upon.

2.6.3 Split competencies and interests

Designing systems that are 1) rigorously grounded in theory and 2) are appealing in messy real world contexts requires two distinct competencies. Dede describes the consequences (Dede, 2004):

[...] much DBR lacks a strong theoretical foundation and does not attempt to generate findings important for the refinement and

evolution of theory. Part of this shortfall may be that the skills of creative designers and the attributes of rigorous scholars have limited overlap. Effective DBR groups have a complex “cognitive ecology” with contradictory tensions: freewheeling, “whatever works” innovation versus controlled, principled variation. People fascinated by artifacts also are often tempted to start with a predetermined “solution” and seek educational problems to which it can be applied, a strategy that frequently leads to under-conceptualized research.

In my experience in education technology for design, there are more than just two categories. Psychologists, technologists, interaction designers, visual designers, ethnographers, teacher liaisons, and others all play a role in the development of successful technologies for learning. While there have been attempts to engage technologists as collaborators in research (Slotta & Aleahmad, 2009), it is a difficult social challenge. Operant probes provide a productive interface and artifact with which to engage.

2.6.4 Difficulty of modeling across layers of complexity

The design of technology for education requires multiple distinct competencies, and research to support these different aspects requires different models. DiSessa and Cobb explain the gap between these layers (diSessa & Cobb, 2004):

We introduce the phrase “managing the gap” to name the issue that is behind the failure of most frameworks for action to achieve what we would like accomplished. The “gap” arises from the fact that instruction is the result of many sorts of complex, interacting elements. Instruction depends on the values of the participants; it depends on technological infrastructure; it depends on the nature of classroom discourse; it depends on practicalities such as available time. We also want to make instruction both depend on and serve to test theory. And yet, in order to see and assess the impact of underlying theory, we must cleanly separate it from the myriad of other issues that we handle, as best we can, in the management of trade-offs among the multiple constraints impinging on instruction. In the ideal case, then, pedagogical strategies and conjectures are separated by a carefully considered and articulated gap from the theory or theories that explain or motivate them. A well-managed gap separates the implications of a particular theoretical claim from other claims and also from atheoretical aspects of design. Attention and effort are necessary to perform this management.

To take a noneducational example, there can be no doubt that there is science in the design of airplanes. However, the shape of a Boeing 747 aircraft does not follow in a direct and simple way from any of this science. Neither does the shape of the aircraft, as a whole, directly test elements of the underlying theory. With sufficient care

(corresponding to managing the gap between design and theory), however, a failure attributable to the shape of the aircraft might implicate a failure of a theory of strength of materials, not just to a careless mistake, a failure to anticipate transient loads, or a poor choice of materials.

An operant probe is like the aircraft. Similarly, it is a specific operationalization of an underlying theory (or theories), and a failure to affect the desired outcome could be due to any number of factors that do not directly test the theory. However, those success conditions and effect sizes do inform the relative significance of theoretical predictions, ease of operationalization, etc. The pursuit of successful operant probes helps illuminate these factors. The iteration of probe designs by different communities can help isolate the features that contribute to that success.

2.6.5 Expense of collection, management and analysis of data from context

Much of the *in situ* data for design-based research comes from observing the physical environment. While these methods can facilitate the discovery of important subtle issues, they are expensive in time and resources. Collins et al. describe the prevalence of unmanageable data (Collins et al., 2004):

Design researchers usually end up collecting large amounts of data, such as video records of the intervention and outputs of the students' work, in order to understand what is happening in detail. Hence, they usually are swamped with data, and given the data reduction problems, there is usually not enough time or resources to analyze much of the data collected. It also takes resources to collect so much data, and so design experiments tend to be large endeavors with many different participants, all of whose work needs to be coordinated. All these factors make design experiments difficult to carry out and the conclusions uncertain.

Operant probes automatically collect their own data through logging of their operation. Additional data can be cheaply and reliably related to the recorded data. Keeping with the airplane metaphor, operant probes each have a black box recorder.

2.6.6 Failure to scale

Educational innovations can be difficult to scale to more users. This is often because the innovation is transformed in new contexts, sometimes losing the essential productive aspect or even becoming “lethal mutations” (Brown, 1992). Traditional design-based research requires working intimately with these contexts to manifest the innovation as intended. Operant probes, as Web-enabled software artifacts, have can be manifest in new environments with high fidelity to their original designs.

When innovations that worked in a limited context are scaled up to more contexts, they often fail. Large randomized controlled trials, such as those advocated by the IES What Works Clearinghouse, find null results so often that some researchers think that What *Doesn't* Work would be a more accurate description (Schoenfeld, 2006). A common reason for the failure to find effects is that the intervention is not implemented correctly in the schools. Teachers often resist these top-down changes to their practice and reject them, even those who had signed up for the study. Instead, operant probes are designed for adoption by practitioners and randomized controlled trials can take place within systems that are already integrated into practice. Ideal operant probes fit easily into new contexts without much researcher intervention. Imagine an airplane that can be duplicated without cost and modified like clay.

2.6.7 Difficulty of reproducing studies

Due in part to scaling issues, experiments in technology design are difficult to replicate. This limits their contributions to generalizable knowledge.

DiSessa & Cobb content that “design research will not be particularly progressive in the long run if the motivation for conducting experiments is restricted to that of producing domain specific instructional theories” (diSessa & Cobb, 2004). To test the generality of principles and design decisions requires testing concrete enactments of them in other contexts.

The story of the designing the first supersonic jet (Phillips, 2006) helps to illustrate how operant probes can help:

In the period immediately following the end of World War 2 the US military, in conjunction with the agency that was the forerunner of NASA, set out on a project to design a plane that could regularly (and safely) fly faster than sound. Not only was it desired to produce a workable product (the X-1 plane), it was also desired to understand the physics – the aerodynamic principles – of flying at speeds greater than Mach 1. In essence, then, the participants were involved in an early piece of design research. The situation was alleviated by the use of two planes, one for pushing as hard as possible, the other for slower testing. Maybe there is a moral here for design researchers, and their funders.

The operant probe, as an easily replicable and deployable artifact, can be used in multiple settings for multiple purposes. In one, designers can move fast to optimize its operation (the plane pushing as hard as possible). With copies (forks) of the probe, researchers can move slowly to maximize utility of the probe to understand what is going on and what can be learned more generally with confidence before changing it again. When the designers have innovated something interesting to the researchers, they can move on to that version to investigate.

2.6.8 Limited duration of studies

Design-based research in education often requires the active participation of researchers in the context. By putting this design effort into developing systems that can be deployed cheaply to many other contexts, the marginal cost of additional time for a study is greatly reduced. A researcher can let the system persist and check in on it just periodically. This facilitates more longitudinal research that looks at changes over months and years. The systems can even follow participants through multiple learning contexts.

2.6.9 Control of variables

Perhaps the foremost challenge to design-based research in education is the lack of control of variables. Dede describes it poetically (Dede, 2004):

The queasiness about DBR felt by many scholars conservative in their research methods stems from the realization that in DBR studies many variables are deliberately and appropriately not controlled, the “treatment” may evolve considerably over time, and even the research methodologies utilized may shift to fit the morphing intervention. Further, to aid with interpretation under these difficult circumstances, in DBR large qualitative and quantitative datasets of various types are often collected by many different participants, introducing substantial problems of alignment, coordination, and analysis. To a methodologist steeped in traditional Campbell and Stanley research strategies, this combination of challenges may seem less a promising new approach to scholarship than a type of study conceived in hell as Sisyphus-like torture for investigators.

With operant probes, the system itself is a well-controlled variable, due to its reliability of replication. Evaluating a design in whole however confounds many variables that make up the design. As a way around this, features of the system can be selectively ablated to determine which are important to the outcome variables. The scale of operation allows these sorts of experiments.

2.7 Limitations

Operant probes offer many benefits, but only for situations in which they fit. Not all settings would allow an operant probe study. One strong reason would be the privacy of the subjects. The benefits of operant probes are largely in the remote collection of data. The limitations that different contexts pose on data collection can limit the viability of a probe. For example, federal law places strict requirements on how schools store and release grade information. In light of the benefits of mining these data, these restrictions may be lessened.

Another limitation of the operant probe is in the requirement that it be desirable in the context. This is much easier said than done. Educational games offer a

clear example of a desirable probe. What other systems can be made that are desirable? That question will drive research in the design of probes.

Finally, there are new types of limits to the control and data collection. Operant probes take some features of the lab out into real world, such as control of design and collection of data. However the remote nature of use, while providing other benefits, limits the data and control to only what takes place inside or with the system. There is no way of controlling or even knowing exactly where and under what circumstances users are interacting with the system. The system can also them such questions or operate a virtual laboratory though a web camera, but such solutions trade off on the authenticity of the learning experience.

2.8 Design process

There are many methods for developing such probes, whether called probes or intentionally following any method at all. Because probes must fit into a user's real world experiences, the most productive design framework would be user experience design. User-centered design is a perspective and set of design methods to help designers products that their target population *can use*. User experience design builds on user-centered design and expands the scope to methods for designing products that the target population *wants to use*.

A necessary step in designing an operant probe is selecting a goal for which to design. Researchers often approach the design process with a problem frame in mind, for which they are designing a solution. Without questioning this frame, they proceed to iterate towards better solutions within that frame. Many other problems and many other frames to the same problem are often ignored. Mapping out the opportunities can lead the researcher to discover new ways of looking at the context and where operant probes could be most successful. As part of user experience design, sketching user experiences can help to create this map (*Sketching User Experiences: Getting the Design Right and the Right Design*, 2007).

User experience design has some limitations for education research, principally that it looks to the user as the source of data for design decisions. It does not always ask the user directly, for they might not know what they really want, but it does try to extract the design knowledge from the user. However in education, many important design principles are secret from both the designer and the user. For example, there is a whole class of “desirable difficulties” in the learning process that improve long-term retention of the learned material (Bjork, 1994). For example, running counter to the training and intuitions of graphic designers, making fonts hard to read can be desirable because the difficulty leads learners to better retain the information (Diemand-Yauman, Oppenheimer, & Vaughan, 2011). Decades of education psychology research have illuminated the processes of learning and many results counter our human intuitions. Indeed, these counterintuitive findings are perhaps the most important to implement in usable artifacts because they run so counter to the dominant practices. In the

following chapter I discuss how to integrate theoretical knowledge of learning into the user experience design process.

2.9 Evaluating an operant probe contribution

For an operant probe to be a research contribution, there must be clear criteria by which to evaluate it. Like other Research Through Interaction Design artifacts, operant probes should be judged by Process, Invention, Relevance and Extensibility. Quoting from Zimmerman, Forlizzi and Evenson: For process, “interaction design researchers must provide enough detail that the process they employed can be reproduced. In addition, they must provide a rationale for their selection of the specific methods they employed.” For invention, “Interaction design researchers must demonstrate that they have produced a novel integration of various subject matters to address a specific situation. In doing so, an extensive literature review must be performed that situates the work and details the aspects that demonstrate how their contribution advances the current state of the art in the research community.” For relevance, interaction designers “must articulate the preferred state their design attempts to achieve and provide support for why the community should consider this state to be preferred.” “The final criterion for judging successful design research is extensibility. Extensibility is defined as the ability to build on the resulting outcomes of the interaction design research: either employing the process in a future design problem, or understanding and leveraging the knowledge created by the resulting artifacts. Extensibility means that the design research has been described and documented in a way that the community can leverage the knowledge derived from the work.”

In addition to the above criteria for all RtD, operant probes must satisfy three more criteria: Acceptance, Insight, Scalability and Effectiveness. Acceptance is the evidence that the operant probe artifact is desired in its target context, such that it will be accepted by the stakeholders and would be used independent of a research activity. Insight is the production of some generalizable knowledge through the operation and evaluation of the probe. For example, practical limitations to the operationalization of some theoretical construct or lab-based results. Insights can also lead to generalizable knowledge; for example, by raising new testable hypotheses. Scalability is evidence that the system can easily scale to more users and contexts. For example, a highly desirable system might be one where the researcher pays to have all student essays graded with detailed feedback and annotations. The costs of such a design prohibit scaling up. This can also be conceived as the marginal cost of additional use within and between contexts. Effectiveness is evidence that the operant probe operates on the context to achieve a desired outcome. This includes not just whether there is an effect but how great it is (e.g. effect size measure such as Cohen’s d). This allows both comparing the effects of a probe in multiple settings and assessing whether the probe is worthy of the resources it requires.

2.10 Conclusion

The inability to translate research into practice threatens the enterprise of education research and stagnates the practice of education. The methods of Design-Based Research in Education help to develop this *usable knowledge* for practice. Modern computing is creating a new opportunity to create *usable artifacts* that carry over research knowledge and engage researchers directly in the real world environments of practice.

Operant probes are a type of research artifact that are usable directly in practice and allow researchers to test hypotheses and models in vivo with relatively high experimental rigor. Research through Design provides part of a design frame by which to create and evaluate usable artifacts. I've extended RtD into Scientific Research through Interaction Design as a frame by which to create and evaluate operant probes in particular. In the next chapter I will explore methods for mapping opportunities for operant probes and in the following two chapters I document the design and evaluation of two probes designed to exploit found opportunities.

3 Discovering SRtID Opportunities in College Lecture Courses

3.1 Introduction

In chapters one and two I argue that design research in education can be more productive by designing, studying and iterating upon systems that operate in real-world, authentic contexts. To design a system to operate successfully in authentic contexts requires a rich understanding of that context. Further, for the system to be adopted requires designing for the factors that drive stakeholders to consider and use the system. Developing this understanding of stakeholders helps to identify opportunities for which to design that will be adopted easily, advancing the practice of education and the resources with which to advance the theory. I describe this as a sketching phase in a broader learner experience design process.

In this chapter, I describe a case of adapting the design sketching methods common in HCI to the theories and methods of learning science. The outcome is a lightweight local theory of the target context. This local theory makes predictions about how different design interventions would affect different desired outcomes in the context. These design intervention hypotheses can, and often are, developed directly from literature or casual observation, but I will argue that the rich tacit knowledge gained through this process improves the designs and thus the interventions and their assessment. Further, I describe a subsequent process for combining local theory with general theory to evaluate contextual factors in the implementation of general principles and opportunities for general principles to address issues observed in context. With this map of the opportunity space,

I describe a process of ideating solutions to the most opportune issues and sketching out designable experiences that highlight those issues. Finally, I return to the stakeholders to discuss the sketched learning experiences, refining the opportunity map and my understanding of the learner needs within it.

3.2 Methods for Discovering Opportunities

Design requires first articulating a current state and a preferred state (Simon, 1969). Education researchers often have current and preferred states in mind, as they are well familiar with the challenges and goals of education. Successfully designing systems that change the world from the current to preferred state also requires understanding the context.

For a technology to affect change in a context, it must work through the culture of that context. (Culture is meant here broadly as the influences on a person's behaviors, e.g. from institutions, authorities, peers and self.) A technology artifact with a perfect cultural fit would be adopted quickly and effortlessly. Consider the iPhone. What prospects would an entirely new computer platform have in a saturated marketplace? The iPhone was adopted because, although it is a general-purpose computer, it fit the cultural role of phone. There are contexts it did not fit to initially, such as enterprise environments, which often require top-down administration. Once it met those contextual needs, the iPhone was rapidly adopted. Research in education must also have cultural fit at some level. If an artifact does not fit the culture now, why expect that it would later? If the goal is to study a principle and not an artifact, why expect that the knowledge gained could be applied? There can be valid answers to these questions but they must be asked. A further advantage of cultural fit is it increases the rate at which research questions can be asked and answered. A widely adopted artifact can provide a large and varied data source in which to compare quasi-experimentally across differences of learners or context, or to compare experimentally across theoretically driven variations.

Design researchers in education have developed methods for understanding context, such as the Informed Exploration phase of the Integrative Learning Design Framework (Bannan-Ritland, 2003). Many variants of design-based research practices are used in research on technology-enhanced learning. Figure 3-1 presents a simple framework, orthogonal to those, to describe the process of producing software artifacts used in research and practice. The vertical axis indicates the steps in developing a software artifact from initial motivations in a context all the way to a working system. The horizontal axis indicates the setting of the research activity and progresses from understanding the present state to imagining preferred states to testing them through simulation and finally to enactment in the naturalistic context.

To design systems for user-driven adoption requires one more consideration: what are the current and preferred states as perceived by the stakeholders in the context? Sometimes you can simply ask them, but as Henry Ford is

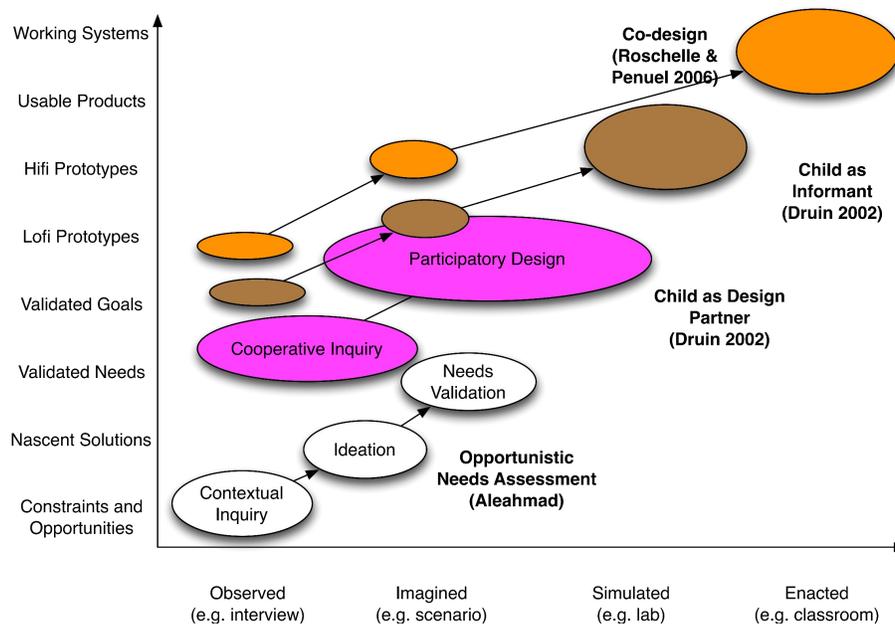


Figure 3-1: Methods in education technology design to support contextual fit

(apocryphally) alleged to have said, “If I had asked people what they wanted, they would have said faster horses”

(O’Toole, 2011). Co-design (Roschelle & Penuel, 2006) and Child as Informant and Design Partner (Druin, 2002) are two examples of working with the stakeholders to design for their needs. The field of HCI, as a discipline largely dedicated to designing new technological artifacts that are complex and fit to people (Fallman, 2004), offers methods that are more widely practiced and validated. I contend that an earlier inquiry into context, by adapting HCI methods to an opportunistic assessment of needs, can help design researchers in education to develop systems that fit better into real world use. Systems that tap into a need deeply felt by stakeholders may be adopted easily and become more productive tools for research and design.

The goals of this inquiry were two-fold. One was to discover opportunities for which I could build an operant probe and conduct a scientific evaluation within the time constraints of this dissertation (1 year). The design process steps to select the needs for which to design, i.e. the opportunities, are depicted in Figure 3-2. An additional goal of this inquiry was to describe the culture of undergraduate lecture courses in such a way that other researchers and technologists can find new opportunities for their own work. The *Theory in Context* section below offers this type of description.

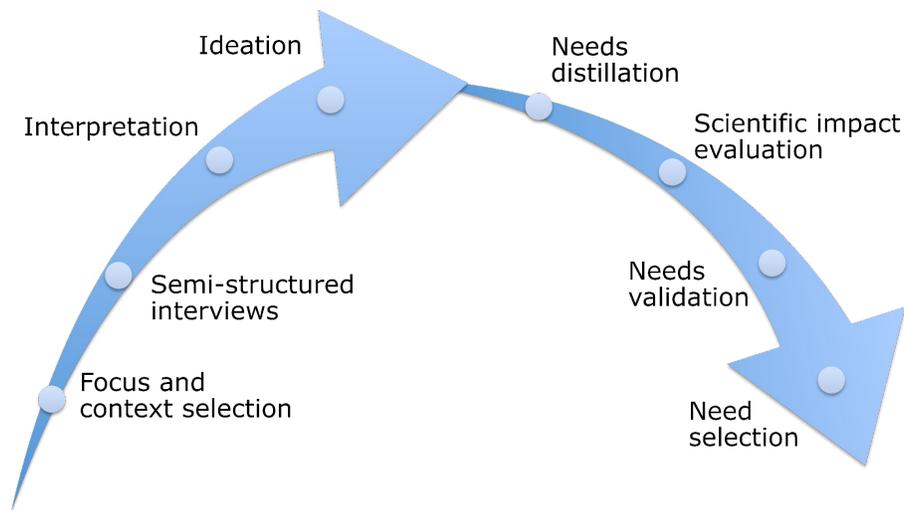


Figure 3-2: Steps of design inquiry process

3.3 Focus and context selection

For this design process, I began by selecting a context that could benefit from the scalability afforded by computing technology. Large college-level lecture courses were chosen because they fit this criterion, are relatively convenient to study in a university, and college instructors have more freedom to try experimental technologies and deploy ones demonstrated to work.

A successful design that fits easily into college lecture courses could improve the learning outcomes of tens of millions of students around the world. In 2007, over 150 million students were enrolled in college (“Data Points: More College Students Around the World,” 2009) and 18 million in the US alone (Snyder & Dillow, 2011). The numbers of enrollments is growing rapidly, and most of these students are in the familiar lecture type course. The learner experience in these courses is based in centuries of tradition and ripe for innovation.

3.4 Semi-structured interviews

To investigate the context I use HCI methods and adapt them to the values of the operant probe paradigm. The first set of methods come from Contextual Design, a methodology that has been used in hundreds of software products (Beyer & Holtzblatt, 1997). The first step of contextual design is Contextual Inquiry, collecting data through interviews and observations in the actual context of potential use. The second is Interpretation of the data into models of perspectives on the context (e.g. cultural, physical) for each informant. The third step is Data Consolidation of the models across informants to create a single comprehensive model for each perspective. I also used Affinity grouping, another approach to Data Consolidation, to find themes among the side notes I took along the way.

To conduct the contextual inquiry interviews, I first began by focus setting, a process for making explicit key goals of the inquiry. Aspects of the context outside the focus can come up in the interviews but the focus helps keep the dialogue on a productive course. To choose the foci, I first wrote on sticky notes a few dozen ideas I had about aspects of the context that technology could help improve. I then clustered these together on a whiteboard in an affinity diagram to find distinct themes. From these several themes, I chose two. The “material use” focus encompassed the production and management of materials such as syllabi, lessons, quizzes, class notes, etc. The “social context” focus encompassed how student peer interactions, discussion boards, sharing of materials among instructors, asking questions, etc.

3.4.1 Participants

Contextual Inquiry requires interviews with participants in each role that will be impacted by the technology. For our chosen context of the large lecture course, I included students (10), faculty instructors (6) and graduate teaching assistants (3). The faculty instructors were drawn from two universities in the area by emailing a list of candidates. The list of candidates was made from personal recommendations and the course schedule. All faculty were from psychology except one from biology. The teaching assistants were drawn from the course schedule for the summer term in which the research was conducted and were all from psychology. Students were eligible by virtue of a participating instructor and were solicited by announcements in class or by email.

3.4.2 Data Collection

The interviews were conducted in the contexts in which the subjects engage in activities for the course. For the faculty and teaching assistants these were their offices. For the students, these were the library, their dorm rooms, and cafés. Each interview spanned 1-½ hours and was audio recorded. For interviews in which the subject mentioned a tangible artifact, such as their class notes or a study guide, these were documented by taking a photo.

3.5 Interpretation

After the interviews, I listened to all the recordings and recorded observations relevant to the models specified by the Contextual Design methodology: Cultural, Flow and Sequence. I also noted “Design idea” observations that subjects had offered, that I had thought of out loud in the course of the interview, and that occurred to me while listening. For the affinity grouping technique of Contextual Design, I also noted “Affinity” observations, a catchall for observations to continue to process but which weren’t design ideas or informative to the models. The design idea and affinity type were used later for the ideation phase.

Rather than a strict transcription, I typed paraphrases that were time-stamped to the audio with Transcriba, a Mac application. I exported the observations from

each interview and concatenated them all in a spreadsheet of 1,014 observations. Each row contained the type, the subject, the time in the audio recording, and a note with the observation. Figure 3-3 shows the distribution of observation types. Note that the Cultural model observations were much more numerous than the Flow and Sequence observations.

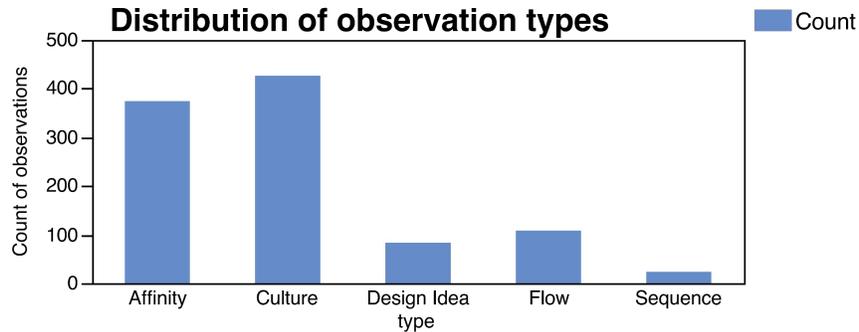


Figure 3-3: Counts of each type of observation recorded

To visualize the cultural forces influencing each stakeholder’s behaviors, I created graphical models of the data in each interview. I used a subset of the models in Contextual Design, including a cultural model for every interview, flow models for most of them, and a few sequence, artifact and physical models. Only the cultural model revealed insights not obvious in this context, and it is the only type for which I generated a consolidated model.

3.5.1 Consolidation

To reveal patterns across the interviews, I combined all 19 models into a single consolidated model. Because I had come to focus on the cultural aspects, I did this only for the cultural model. The observed cultural factors that influence behavior of students and teachers are in Tables 3-1 and 3-2. Each *Factor* is an abstraction of multiple statements recorded in the interviews. The *Influencer* is either the actor themselves (e.g. students behaviors are influenced by their own goal of doing the minimum to get the grade they want) or another actor in the context (e.g. teachers influence student behaviors through culturally communicating the idea that they reward students for short-term, not long-term learning.) This consolidated cultural model presented a coherent view of all the roles in the cultural context and the cultural forces acting internally and between the different roles. Because it combines the models from each informant, it reveals tensions, breakdowns, and design opportunities among the cultural elements.

Table 3-1: Table 3-1 Cultural influences on student behavior

| Influencer | Factor |
|------------|--|
| Self | I do the minimum I must to get the grade I want |
| Self | I procrastinate on everything |
| Self | I need to do well in my classes to feel good (otherwise feel stressed) |
| Self | I am distracted easily |
| Self | Organization reduces my stress |
| Self | I don't retain information |
| Self | I feel bad about my poor study habits |
| Self | My effort is based on how interested I am |
| Self | I want to know what I know and what I don't know |
| Self | I am unmotivated without personal interaction |
| Self | I can never be good at certain parts of this class |
| Self | I doubt my value to my peers' learning |
| Self | I enjoy having room for my own perspective |
| Self | I want immediate feedback on my reasoning |
| Self | It's painful to put in effort and not see results |
| Self | I am motivated by work that will help me succeed later |
| Self | I can't bear the thought of being wrong |
| Instructor | When we care, you care more |
| Instructor | I reward you for short-term, not long term retention |
| Instructor | I let you get away with bullsh*tting |
| Instructor | I grade you on things that you don't care about |

Table 3-2: Table 3-2 Cultural influences on instructor behavior

| Influencer | Factor |
|------------|---|
| Self | If students don't do well, I haven't done a good job |
| Self | It's important to connect class to students' lives (prior knowledge and real world experiences) |
| Self | I enjoy teaching depth more than breadth |
| Self | I care deeply about the quality of my assessments |
| Self | Stimulating learning is more important than accurately measuring it |
| Self | I must train you to learn on your own |
| Self | It's my responsibility that you value understanding this domain |
| Self | I want to install ideas that will shape your life |
| Self | The best learning occurs when it's fun |
| Self | I want each student to succeed in their goals |
| Publisher | Use shallow instruction and assessment we are selling |
| Department | You must turn out creative smart people |

| Influencer | Factor |
|------------|---|
| Department | We reward you for time spent researching, not teaching better |
| Students | When we care, you care more |
| Students | You should be able to make learning easy |

3.5.2 Results of cultural modeling

The consolidated cultural model brought to light several conflicts between the cultural forces shaping the behavior of participants in different roles.

Misaligned goals: Student comments indicate an almost exclusive focus on getting a good enough course grade as their primary goal. Faculty, on the other hand, talk almost exclusively about learning. They lament that high quality evidence of learning cannot be measured in class. What can be measured easily, such as recall, is not important. What is important, such as application of knowledge, is laborious. What is most important, that the learning help students achieve their life goals, is intractable. (e.g., a professor quoting Herb Simon: “I never gave an exam I liked that was gradable.”)

Misaligned motivations: Students are motivated to learn things they care about, but feel they are graded on content they do not care about. Consequently, students feel the system rewards those who focus on grades and not those focused on their interests or what is useful to know.

Tension in scope: Students want to know exactly what they have to learn. Professors want students to learn things that they do not specify or measure. Grades are the strongest signal to students and what is not assessed they will ignore. “Is it on the test?”

Depth versus breadth: Instructors enjoy teaching depth more than breadth. Students are more comfortable with breadth. Textbooks are broad and shallow to lower costs and meet majority preferences of the market. Publishers perpetuate a feedback cycle incentivizing breadth over depth.

Tension in support: Students have less stress when the class is organized. Faculty believe students should be learning how to learn so do not help “too much”. Faculty sees their role as teaching a domain and do not provide instruction or assessment of how to learn.

3.6 Theory in Context

3.6.1 Method

To find the challenges and opportunities posed by the context to operationalize and contribute to educational theory, I identified how different theoretical models and recommendations were active in the context and barriers to activating them. To begin, I selected a scope of theory and practice recommendations

to consult. For the recommendations I draw from three theory-based Practice Guides published by the IES What Works Clearinghouse: Organizing Instruction and Study to Improve Student Learning (Pashler et al., 2007), Using Student Achievement Data to Support Instructional Decision Making (Hamilton et al., 2009), and Structuring Out-of-School Time to Improve Academic Achievement (Beckett, Borman, & Capizzano, 2009). I refer to these, respectively, as *OIES*, *USAD* and *SOST*.

To orient to how these theory-based recommendations are active in the context, I coded each theory-based recommendation by the evidence for its use in the context. For each recommendation, I coded what data were presented in the interviews for its application and perceived need in the context. Levels of None, Low, Medium and High emerged, which I coded 0-3. When there was evidence against the application or perceived need, it was coded as negative. The two factors were also split by the two key stakeholders, students and instructors. When there was a range among informants within the role, the value in the range was chosen that most highlighted an opportunity for design.

In the spreadsheet I then added calculated columns: *Needs Spread* for the difference between the perception of students and teachers (three were ≥ 3), *SN vs. TP* for the difference in student perceived need versus the teacher practice (two were ≥ 3), *TN vs. SP* conversely (five were ≥ 3), and *Sum Practice* adding student practice and teacher practice codes (three at -4 and four at -3). I discuss these below situated in the cultural model.

3.6.2 Cultural barriers to implementing educational theory

The consolidated cultural model reveals several cultural barriers that may hamper the use of evidence-based recommendations. Below I begin with the recommendations highlighted by the above analysis and situate them among the cultural forces of the actors in the context.

1. Quality questions are scarce Because of the cultural practices of question production and evaluation, several of the recommendations are difficult to practice. One of the recommendations for which there is the most evidence, “Help students build explanations by asking and answering deep questions” (*OIES* #7) also scored highest on teacher perceived need versus student practice (5). Teachers value deeper conceptual reasoning and particularly enjoy teaching the deeper concepts (e.g., “I’d like to teach a higher level”). Students tire of rote questioning and enjoy questions for which there is not a single “right” answer that they either hit or miss. A strong opposing force is that deep questioning requires more effort on the part of all participants. These efforts can be divided into question generation and answer assessment. For professors, there is a trade-off in time and difficulty between their generation of questions and grading assessment. An open-ended question can be easier to produce than a set of multiple choice questions that measure the same understanding, but the price is paid at grading time when each answer has to be read and fine distinctions have to be inferred as

to what levels of understanding are demonstrated. Grading time can be reduced by multiple choice or short answer questions, but to make these “deep” questions requires much more thought and, often, more time pilot testing to produce good foils that represent common student misconceptions.

The instructors in our study often turn to textbooks for their questions because they are plentiful and no additional cost. Unfortunately, the questions in texts are often shallow and rarely valued by the instructor. The textbook publishers sell primarily the textbook and a book’s question bank is offered to help sell that book, rather than being a revenue builder itself. An associated design idea is a technology in which publishers’ create question banks as a living online resource that instructors can review, collaboratively filter and improve upon.

The efforts of professors to produce questions can also be amplified. They work so hard to make these that they are cautious in sharing them and they are typically shared with only a few colleagues at best. This observation supports a design constraint that there are sufficient security assurances of access only by qualified instructors. With such, there might be a peer market so that faculty can exchange their valuable questions, increasing the supply and thus reducing the cost.

1a. Secrecy of questions limits formative assessment The scarcity of questions compels the teachers to re-use the good questions they do have. The re-use leads to careful guarding of questions limited in use to summative assessments and accessed only by trusted colleagues. The hampers *OIES* recommendation #5a, “Use pre-questions to introduce a new topic.” Practicing this would help address students’ expressed needs to be aware of what they do not know and have evidence of their learning. It would also help prevent the perception of some students that teacher’s questions are meant to trick them. If there were a sufficient supply of exam-quality questions, they could be used to help guide students’ attention to a new topic and familiarize them with the content and style of questions their grade will later depend on.

Another formative use of questions is after they are answered. The OSAD guide recommends to “Teach students to examine their own data and set learning goals” (#2). The guide explains that “instructional strategies such as having students rework incorrect problems can enhance student learning.” (Clymer & Wiliam, 2006) Because teachers are compelled to reuse their questions, they cannot allow students to keep their graded exams. (Students would certainly share the questions with other students who had not yet taken the course.) One technology to address this without revealing the exact questions would be to provide detailed formative feedback with each summative assessment. Instead of the single percentage grade from the Scantron™ most students receive now, they could be provided with a report of their performance on different knowledge components (of large grain size if need be) and supplemental instruction to help address their deficiencies. This could help promote a Mastery style of learning and explicit goal setting.

Another technology opportunity is to address the scarcity of questions by the recommended practice of students asking deep questions. With the proper supports for quality in production and filtering, students' questions can be used for formative or even summative assessment. Because students' questions are often shallow, they could also be used to free up the teacher's efforts for deeper questions or as a starting point for a guided inquiry into deeper questioning. This crowd-sourcing technique could be extended to students that are more senior or the entire Internet. With sufficient quality assurance mechanisms and participation, instructors or tutoring systems could draw randomly from large pools without risking repetition.

2. Specifying learning goals conflicts with flexibility and adaptivity of the course Several recommendations and perceived needs hinge on specifying learning goals. The OSAD guide recommends to "Teach students to examine their own data and set learning goals" (#2). "Tools such as rubrics provide students with a clear sense of learning objectives, and data presented in an accessible and descriptive format can illuminate students' strengths and weaknesses (see recommendation 5 for more information on reporting formats) (*Assessment for learning: putting it into practice*, 2003) Students want to know what they know and don't know, which depends on a taxonomy of what they are expected to know. They also want to feel a sense of progress for their efforts, which is difficult without clear targets. Finally, students expressed that organization in the curriculum lowers their stress and some instructors expressed a desire to minimize anxiety for students.

The responsibility for specification of learning goals falls on the instructor, and in them are two forces pushing back. One, it is a lot of work that they see little point in. They believe they know what the goals are even if they have not articulated them. They will teach them so why spend the effort to write them down? This could be addressed by outsourcing that labor somehow. Students or assistants could be involved in writing the specifications. One technology opportunity might be a shared workspace in which the students collaboratively develop and revise what they believe to be the learning goals of their lessons and activities. This would certainly yield gains in metacognition and the instructor could give students the feedback necessary to reach a working set of goals.

The second conflicting force is deeper and a part of the instructor's conception of their role. The instructors see themselves as providing a unique experience based in their particular domain and pedagogical expertise. To follow a highly specified curriculum, especially one they did not create, reduces their creative role to one of executing a program. While detailed rubrics help students plan their learning and understand their progress, they also limit the instructor's sense of freedom in adapting to his students or seizing upon spontaneous opportunities that "will often pop into mind during lecture". In other words, instructors feel that learning goals can be slippery and shift throughout the course. Accordingly,

assessments are crafted based on the course as taught, not as planned, with redistributed emphasis based on how much a topic was or was not covered.

These observations lead to a design constraint to preserve instructors' needs for creative input and flexibility with students' desire for structure and predictability. Can we create a design to support explicit goal setting in a fluid manner? Students might be involved in articulating the learning goals based on the professor's teaching and revise the set as it evolves in class. The professor can refine the specification and use it with software to dynamically generate assessments that align with the course.

3. Faculty ostensibly teach how to learn but don't assess or instruct it directly The professors in our study see their most important role as teaching students how to think, not what to think. While they carry the responsibility that students value understanding of the domain, their larger goal is to instill ideas that will shape each student's life. These skills for learning range from the domain-specific metacognitive support to study strategies to general time management. Yet, none of these is taught explicitly, even though the largest courses are introductory courses with students unaccustomed to the rigors of post-secondary education. (e.g., ““by second semester freshman year I was trying to learn how to study, pretty much teaching myself.”)

Teaching study strategies has field-based evidence (*OIES* #6) and little cultural resistance. The guide recommends teaching students techniques to break their “illusion of knowing”. One way it recommends is through tests and quizzes, which is widely practiced and does not teach any strategies to the student. The other recommendation is to teach students how to create “judgments of learning” themselves while studying, transitioning from demonstrating in class to using the techniques on their own. Some participants, particularly in cognitive psychology courses, do share techniques like this. What they do not do is assess it. Students make clear that they do the minimum they can to get a desired grade. Whether a technique works or not, students are unlikely to find out because they lack the motivation to try it. For those students, it is ineffectual instruction. Formative assessment helps detect ineffectual instruction, but the instructors avoid explicit assessments of strategies because while they wish students in their course learn how to learn, they are not teaching a course on “how to learn” per se. They do not want to award points in a psychology course for general study skills. Moreover, these skills are implicitly assessed through the student's achievements in the domain.

The technological opportunity here is to develop a study tutor that motivates use either intrinsically or extrinsically besides a grade. The system might scaffold students in acquiring the skills and dispositions of effective study strategies and fade as they internalize them. Ideally, it would assess student use of target strategies and provide formative feedback to the instructor on how the students are learning to learn. Moreover, it would provide more data by which instructors can understand students' achievement in the domain per se.

With sufficient data and interactivity, computer-based systems could support mastery of fine-grained metacognitive skills. An intelligent tutor for metacognitive skills might help students to see their progress and provide the immediate feedback that they desire. A fine-grained model of metacognitive skills in the domain would support a Mastery orientation to the skills, motivating students who believe their abilities are fixed. Of course, this is a grand challenge but the results of our inquiry suggest it is a worthy one.

Another decomposition that technology can support is at the large grain size: completing the course. Students have difficulty breaking their goals into actionable chunks and scheduling them optimally. To the extent that the course is not meant to assess this skill, the entire course could be broken down to a hierarchical checklist with target dates. Software guides through the checklist would provide students the organization they seek and lower their stress, allowing them to focus on learning. To the extent that instructors wish to train or assess these skills, software could support student decomposition of the course syllabus and provide feedback on the chunks and on the schedule students have set. By making planning explicit, instructors can choose and convey how much students should be learning and internalizing “how to learn” versus the domain alone. Such approaches would ideally be implemented and applied across courses.

4. Improving instruction through student data requires infrastructure and roles that don’t exist Instructors care very much about the achievements of their students. They pay attention to class discussion, assignment submissions and exam scores to form their course, and to the apparent needs of the students. (E.g., “if people are not scoring very well, my presumption is that I didn’t do a good job.”) There is much evidence for the importance of these practices, so much that a whole IES practice guide is dedicated to using student achievement data to form instruction. The guide emphasizes that data practices require a culture and infrastructure. One recommendation “Provide supports that foster a data-driven culture within the school”, includes steps: 1) Designate a school-based facilitator who meets with teacher teams to discuss data, 2) Dedicate structured time for staff collaboration, 3) Provide targeted professional development regularly. I did not find any evidence of these practices.

An opportunity this presents to technologists is to support a bottom-up culture of data-driven instruction. For several professors in our study, exam scoring is outsourced to an office of testing, which returns the student answers and simple percentages. One instructor checks each question and each student test for irregularities. It requires hours of manual work that could be automated and other instructors do not have the time. If various reports were instantaneous, instructors could learn more about the quality of their assessments. If they were linked to a question bank, instructors could compare student performance, and their instruction, from semester to semester. One constraint would be to include the testing office in the design process and solution. Otherwise, they may fight what they perceive as a threat.

There are many other data relevant to student achievement that are not available in a data analytic form. For example, textual responses are often hand-written, preventing use of computer-based language analysis. Other modes of data could be collected, such as non-traditional assessments like question authoring, the metacognitive activities above, how students use their time, or simply class questionnaires. Using technology to provide new measures of student activity would allow instructors to explore and potentially develop their own hypotheses about how to improve their courses. Detailed student data could also help instructors differentiate instruction. “One of my biggest challenges is the range of abilities of students coming into the class,” said one instructor.

5. Instructors have precious little time to integrate new technologies A barrier impedes the implementation of all any technological system for implementing the practice recommendations is that instructors generally have very little time to spend integrating them. Faculty #3 lamented on the use of PowerPoint, “I spend an amazing amount of time with the apparatus of instruction rather than what I really want to be doing, which is thinking.” For most professors and teaching assistants, student learning is not their primary responsibility. Once a curriculum is developed, they have little incentive to change it. No wonder that many instructors reported beginning new courses with a colleague’s or publisher’s materials and adapting them over time.

Information technologies that are meant to help can end up taking more time. Faculty #2 tried using online discussions and retreated. “I did feel like more people participated, but it was easy to consume a much larger number of hours [of] everybody's [attention]. [...] I remember thinking, wow, if I was doing this all the time I would end up spending an inordinate amount of time.”

The instructors I spoke to were interested though in trying new ways to improve their courses. For example faculty #2 continued, “The things that make me change things are more... sort of new things that come my way, because I get bored. You know what I mean? Doing the same thing.” To be accepted, though, new technologies must be easy to integrate into courses and not require much time or curricular changes from the instructor.

3.7 Ideation

The aim of the fieldwork was not merely to describe the present but to look to the future to what can be designed to improve on the present. To take a step towards the concrete, in the next phase I generated many distinct ideas for systems that could be built. I began this process by reviewing the observations and organically grouping them per the affinity grouping method of Contextual Design.

The numerous observations and ideas required a different approach than the traditional whiteboard of sticky notes. To create the paper notes, I wrote a Python application in the Django web framework to format the 457 affinity and

design observations into a page template that I then print and cut into small squares. Instead of a whiteboard, I laid them out on the floor. Together with a design assistant, we put related observations closer together and gradually honed in on 14 parts of the context for which to design (Table 3-3).

Table 3-3: Clusters of design ideas and observations

| | | | |
|-----------------------|---------------------|--------------|--------------------------------|
| Class logistics | Knowing what I know | Study habits | Professor knowing each student |
| Assessment production | Help seeking | Motivation | Learning goals |
| Attention in lecture | Encoding lectures | Community | Instruction production |
| Reencoding | Pear learning | | |

Using the data on the context, its interface with theory, the tacit understanding developed through extensive interviews, and the organization of the affinity groups and process, I began to generate ideas for systems I could build. This ideation was a relatively fast process and after a few days I had generated 64 distinct ideas, some of which are listed in Table 3-4 below and all of which are listed in Appendix B. Many of the ideas were carried over from my spontaneous thoughts while interviewing and later listening to the interviews.

Table 3-4: Sample of ideation

| |
|---|
| Class study partner pairing system. |
| TA review session voting system (submit questions, everyone votes and popular ones first) |
| Contributions that do not require being right or wrong. (E.g. cog psych scenarios, provocative questions) |
| Analytics on how much students are working and how. (Anonymous logging and reporting.) |
| Study behaviors tutor, tied to real data from learning activities and outcomes. |
| Big ideas database to find concepts that cut across findings |
| Versioning system for teaching materials with in-class annotations on each version. |
| Micro-experiment tracking system for educators. Quick pre/post assessments around a small treatment. |
| Recording questions keyed to time code and point in the slides. Embed student experience in materials (for future self, and others) |
| Crowd-source the content of a learning game (authoring/use class/library/bus/home) |

Algorithm to distribute students throughout the hall for break-out with different groups

3.8 Needs Distillation

Each of the ideas was potentially good, but which were worth trying to develop into successful operant probes? Which ideas would be accepted by users and also contribute to scientific understanding? Instead of selecting directly from the ideas, I used them as part of a process to better understand the user needs and constraints, and the relevant science. The work following Ideation contracted the search space and helped to frame the problem, articulating the current and preferred states for which to design.

Many different needs were evident in the observations from the interviews, but for which needs would designing lead to a successful operant probe? The ideation phase helped to think creatively about the ways that technology could help with those needs. Working back to the user, I distilled these numerous technological ideas into the distinct user needs that motivated them. In doing so I focused on the subset of all observed needs that technology could be used to address. E.g. “Motivating interest, nourishing curiosity”. The full set with initial solution ideas is list in Appendix B.

3.9 Scientific Impact Evaluation

These distilled needs helped me navigate the design space to technological systems that would likely be accepted, but for scientific research through interaction design, I had a second goal of operationalizing scientific theory through designing for the need. This is where the design of operant probes deviates from traditional HCI practice. I conducted a literature review and for each for each candidate need, I cited lab-based results relevant to that need, and annotated it with three factors in its potential for impact: 1) That learning effects are predicted; 2) That the design and use of the artifact would inform other applications of those results; 3) Its fit to the research team. Could we capitalize on this research opportunity?

Through this new design process technique, Scientific Impact Evaluation, I identified which needs would more likely have scientific impact if designed for. Appendix C shows the evolved description of needs and the relevant learning science principles. The principles were drawn mostly from the IES Practice Guide for “organizing instruction and study to improve student learning” (Pashler et al., 2007). For example both students and teachers felt the need to support students’ sense of what they’ve learned. The space of solutions to this need related to the lab-based research on “help[ing] students to allocate study time efficiently” (#6) and the studies behind the two sub-principles to “Teach students how to use delayed judgment of learning techniques to identify concepts that

need further study” (#6a) and “Use tests and quizzes to identify content that needs to be learned” (#6b). Further it was expected that a successful operant probe designed for this need would contribute to the scientific understanding of students’ motivations.

3.10 Needs Validation

After the distillation and scientific impact evaluation, I had 17 needs that I had observed. Because my goal was to design for technologies that would be accepted, I also had to validate whether the stakeholders themselves felt these. Needs Validation is a design method to assess whether what needs the stakeholders themselves perceive and what solutions they are likely accept (Davidoff, Lee, Dey, & Zimmerman, 2007). This is no guarantee that the system built would be accepted, but it helps point the designer in the direction of acceptable systems and improves understanding of each need in order to properly frame the problem.

I elaborated the 17 needs for use in stakeholder interviews. Each scenario was refined and illustrated into a storyboard, to activate the user’s memories and feelings in a situation and draw out their perceptions of the plausibility of the situation, the character’s behavior, and their perceptions of the technological artifact. This can reveal the needs that a stakeholder perceives for their own role in the context. Appendix C shows all the needs annotated with learning science principles and descriptive scenarios to probe on the need. Figure 3-4 shows an example illustration for the scenario in need #10 and Appendix D shows them all.

Scenario 10



Figure 3-4: Sample scenario for needs validation interviews

With these elaborated scenarios and storyboards in hand, I conducted interviews. For each scenario I articulated my idea of what need the scenarios would recall for the student and teacher, along with a lead sentence for the interview to frame the conversation. I interview 7 students (2 as a pair), a group of 3 TAs, and 3 faculty (2 as a pair).

The needs validation process informed both whether stakeholders perceived the hypothesized needs and the cultural and practical constraints on my leading solutions to those needs. For example, I observed that some students resented school for reducing their connections with their friends. I perceived this as evidence of a need to involve students' wider social worlds in their scholastic activities and presented a scenario in which students share their grades and study activities with their friends and family (need #6, LearnShare). Both students and faculty rejected this. One faculty pointed out that during college young people are trying to develop their individual identity, breaking such support ties and finding their own way. While such a system may be desired and helpful in secondary education or even adult learners, the college student context would not accept such a design.

Another notable rejected need (or want) was to be more motivated in class through competition. This was based in my observations that some students already perceived class as a competition for good grades and wanted to be recognized for other achievements, such as writing the best study questions for one's peers (need #5, PeerQuiz). Students and faculty had no interest in promoting competition. While some students may thrive with a competitive technology, such a system would not be accepted by the faculty who are the gatekeepers to their courses. This is an important constraint on the move towards gamification in education.

As for the idea of students writing exam questions, faculty did like it. One instructor said, "This I would use. I'm going to do this actually." However students did not. They didn't trust that the questions would be helpful for the exams and might even lead them astray. Getting this design right would entangle complex social issues and require the active participation of the instructor. Indeed, systems such as PeerWise do implement such a system and do. Considering this system helped me to hone one of my design constraints: that the systems be scalable broadly without requiring the instructor to spend much time or to change their course instruction or assessment.

An interesting constraint of the traditional institution of higher education as a whole was evident in the rejection of the long-term retention scenario. In interviews students expressed frustration at forgetting what they had learned. This could be triggered, for example, by trying to help a friend through a class they had taken earlier and no longer being able to answer the questions, or simply by taking a subsequent course in a series and forgetting the prerequisite knowledge. While students and faculty both agreed it would be best for students to retain what they learn, and the OI&S first principle is spacing of practice, no one was willing to accept OlderCheck, a system in which receive an electronic follow-up quiz months after they complete a course. While an instructor stated plainly, "They forget most of what they learned after the exam," he could not imagine a scenario in which students would do this voluntarily or could be incentivized to do this. This design constraint is imposed by the current structure of university courses and credits. Competitors to the university model

may break free from this constraint. However the goal of this fieldwork was to identify improvements to the college courses of today. This issue highlights the potential for new institutional structures in the future and through this I adopted the design goal to design for back-porting the features from this better future into the structures of today.

Some of the needs were strongly felt but did not provide much opportunity to engage scientific research. An example of this is the students need to ask questions in lecture when they are afraid of seeming ignorant or slow. Need 13 in Appendix C and illustration 13 in Appendix D present a scenario for a system whereby students can ask questions asynchronously on their mobile phones during lecture? Students were eager for a system like this and teachers also supported it (if it didn't disrupt the flow of their lectures or take too much time). However, based on the scientific impact evaluation, I saw little opportunity to advance science through designing for this need. Of course, other researchers or teams may have more relevant expertise towards a scientific understanding of this need and acceptable solutions.

3.11 Needs Selection

The ultimate goal of the field study was to identify opportunities for new designs that would be accepted in college lecture courses. The needs validation affirmed two needs that pointed to solutions that would be informed by the science-based practice recommendations. In other words, two needs were felt (without satisfaction) and had solutions that were acceptable and had scientific evidence in to guide their designs. These features support the opportunity for an operant probe that could test the contextual design knowledge, the general scientific knowledge, and the effects of integrating them in a concrete scalable form.

3.11.1 Time Management

The first need was to support students' time management. All faculty and teaching assistants perceived this as a problem. Many students personally grappled with this and those who did not have this problem affirmed that many of their peers did. Students conceived of this primarily as procrastination. Probed on her study habits, Student 3 said sheepishly "I'm kind of a procrastinator. It's not good!" This student was ashamed of her time use for school ("I should put a lot more effort in all my classes. I've been a real slacker.") even though she falls asleep studying late into the night to wake up for work in the morning. Student 5 felt similarly guilty, "I have very bad study habit. I don't prereading." What students conceptualize as their lack of willpower or discipline could instead be the failure of the educational offering to support them. For example Student 7 reported,

When I got to college, my first semester freshman year was not the best. I didn't know how to study, and still don't study well. And I openly know that. Um, didn't know how to study, didn't know

how to take notes. Didn't really have to in high school, you know? You can kind of get through high school without doing anything too... extensive. So that was tough. By second semester freshman year, I was then learning out to take notes and how to study, pretty much teaching myself.

Faculty and TAs are sympathetic but don't address this is in their primary instruction. Faculty 3 said, "Students that exercise time control, planfulness and stuff, really are ahead of the game." The instructor subjects take some time to help, but students come in with a huge variance in time management skills and they mostly help with such issues in office hours. However many of the most needy students don't come. TA 3 explained,

My experience as a TA is that most students don't come to office hours. Ones who do come to office hours usually are the ones who are already pretty high achieving students, that have, that probably are high achieving students because they have a lot of good skills. How to attain help appropriately. How to know when you need help. My experience teaching [...] has been really eye opening about the varying levels of preparation for college.

Synthesizing from the contextual inquiry, recommendation literature ("help students allocate study time efficiently", OI&S #6), ideation and needs validation, I settled upon a design goal for which a solution would likely both be accepted and contribute to learning science,

- Computer support for students to use their limited time most effectively

To design such a system, the field work made apparent these constraints:

- Require no upfront action by the student in order to benefit
- Require no changes to the instructor's curriculum or schedule
- Require little or no time from the instructor to offer in her course

This goal and set of constraints formed the core design principles for Nudge.

3.11.2 Studying More Effectively

The other evident need was to help students study more effectively. Students validated the need for support in the study process but both students and faculty described the constraints on achieving this. Many students have trouble focusing on studying without action. Student 7 explained, "If I'm not actively doing something then I'm gone, that's it. ... If I stop writing, within 5 minutes that's it. You're done. It's gone." Some expressed wanting to study using more conceptual questions and not merely recall, but believed multiple-choice questions were all teachers had time for. When asked, "What's your ideal test?" student 10 rejected the question. "They have to be multiple choice questions these days. You can't expect a teacher to hand grade 100 tests." Faculty 5 hears from students that they want more multiple-choice questions ("One thing

they always ask is can we get more multiple choice questions”) but she can only provide so many. “The problem is if it's a question I think is really good, I'm saving it for the exam. And if it's a question I don't think is good, how much does that help you?”

While students may appear to crave multiple choice questions, my experience talking to students suggests that what they really want is to prepare for the test and receive immediate feedback on how well they are likely to do. Student 6-valued practice tests and said that in some domains, like math, you cannot study for the exam without practice. The paucity of practice currently may be due mostly to the production of good practice materials. Faculty 5 shared, “It gets more challenging for me every year to ask good, challenging questions.”

One of the students' goals in studying and practicing is to focus their time and attention only on what they don't yet know and need to. For example when asked, “How do you choose to study?” student 3 replied, “I don't know. I'll just like go over my notes and anything that I'm like unclear about, I'll study that like more closely.” While student 3 likes practice exams, she lamented that she doesn't have time to do them all. Feedback on the scenario for a progress tracking system affirmed that students definitely want help knowing what progress they're making. Faculty said they would love to do this for them but it means they have to write a lot more exam questions. They recognize though that many students “just don't know how to study”. For example, one faculty recounted that students come to office hours complaining, “I read the chapter and memorized it and I don't know why I didn't do well.” Students who don't do well perceive it as a limitation of themselves that cannot be overcome. Student 6 explained that science “just isn't [my] thing. [...] I won't do well no matter how hard I try.”

The field evidence points to a clear need to support students' study techniques. The IES practice guide provides some theory-driven guidance: to “help students ask deep question in order to build explanations,”(#7 (Pashler et al., 2007). Combining these points to a solution would likely both be accepted and contribute to learning science:

- Computer support for students to prepare for exams interactively by building explanations

The fieldwork also reveals some constraints for the design of such a system:

- Scaffold effective study techniques for students that work even for students who don't know them
- Be interactive enough that students are engaged
- Help students to accurately assess what they know and don't know
- Be self-paced so that students can go quickly over what they are already confident in

- Map well to course assessments so that students know when they are prepared

Again as for the other needs, there are general constraints of designing any operant probe for college lecture courses,

- Require no upfront action by the student in order to benefit
- Require no changes to the instructor's curriculum or schedule
- Require little or no time from the instructor to offer in her course

This goal and set of constraints formed the core design principles for Exemplify.

3.12 Conclusion

The goal of this fieldwork was to identify needs felt among stakeholders in college lecture courses and to understand the constraints on what offered computer-based solutions they would accept. The HCI practices of contextual design and needs validation helped draw a map of opportunities to design solutions that would be accepted. Learning, however, is a complex activity for which evaluation is much more difficult than mere adoption or perceived value. People are poor judges of their own competencies and learning experiences. To help ensure that the designed systems also achieved the goal of student learning, the design ideas were grounded in the scientifically vetted practical recommendations for educational experiences through the IES practice guides. In the next two chapters I detail the designs of the two proposed systems and ground them further in the scientific literature, both to improve their likelihood of achieving the desired outcomes and to provide an operant probe to better understand applications of the theoretical principles.

In closing this chapter, I'd like to take a step back to reflect on the methods used in this fieldwork. Could I have designed the same systems without conducting the fieldwork? Could someone else have who is smarter and better versed in the issues of college lecture courses? I believe this ethnographic-oriented work provided three essential benefits to this design process that would be difficult to arrive at simply through theory or intuition.

The first is the identification of needs and constraints felt by stakeholders. There is no literature (that I have found) saying that new technologies for studying education should not impose on the instructor. In fact, most design-based research in education calls for the active participation of the instructor. This is good for exploring how to improve teaching in the classroom, but there are many other needs students feel that teachers don't pay attention to or don't feel empowered or responsible to address. One of the results of this particular field study was that while some students have an urgent need for support with time management and it is critical to their learning, it is outside the domain the instructor is teaching (and on which they are so focused that they dedicated their life to teaching).

The second is to provide a wealth of tacit knowledge for the innumerable small decisions made throughout the design process, inquiry throughout the user-driven iterations, and the design of the classroom intervention by which the tools will be experimentally evaluated. So far I have only discussed the design space and constraints illuminated by the field activities. In chapters 4 and 5 I will discuss the designs of the two tools. It will be difficult as the designer and author to identify exactly what knowledge led to each decision (or even what decisions are worth remarking upon). I hope that a reader of this chapter will have developed a similar intuition. Regardless, I advocate that designers of education technology do conduct some amount of inquiry into the context and do so before they introduce their system. Further, that multiple courses be explored to develop an intuition for what varies and what remains constant between courses, and better design for operating at scale.

Finally, the rigor of this process and critical reflection at each stage can lead to a reframing of what the larger problem is that one is attempting to solve. Every design seeks to turn a current state into a preferred state. The lens by which we see now and conceive of what is preferred dictate the designs we can imagine. Domains for which the frame is especially prismatic are known as “wicked problems” (Rittel & Webber, 1973). That is, moving the frame or looking at it from different angles changes the scope and character of what is seen. Education writ large is such a wicked problem. Like urban crime, when you begin to operationalize the exact criteria of a preferred state, what is preferred becomes very slippery. For example, if we could make it so everyone had the same level of education would that be preferred? What if everyone could learn what they would in college without attending, is that something that institutions responsible for education would support? This isn’t to say that improving education is intractably sociopolitical. It is to say that the frames of how we understand the problems in education can be moved and still lead us closer to someone’s preferred state.

Through the active exploration of how learning in college could work, I came to see many aspects of the status quo as vestiges of a bygone era. The agrarian calendar when few are farmers, lectures in big halls when students can watch videos, assessments that can be fed into a Scantron, abstractions of student competencies into course grades and diplomas, the market bundling of instruction, assessment, certification and, for early undergraduate years, overnight camp. The business and technology of education are poised for a radical transformation. This fieldwork helped me imagine a preferred and achievable future that has different challenges than the present. One hallmark of the future of learning, where all knowledge is instantly available, will be decreased importance knowing a domain and the much greater importance of knowing how to learn efficiently from the overwhelming abundance of knowledge. That is the larger design frame that guides this work. How can we develop in students the best skills and dispositions for effective learning?

4 Nudge: Supporting Students' Study Time Allocation

4.1 Introduction

In the contextual design study (Chapter 3) I observed that many students did not know how to be better students and instructors did not include study knowledge (declarative, procedural or dispositional) in the curriculum. I identified this as an opportunity 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: Nudge supports student time management by making course tasks explicit and notifying students of them when they are relevant.

This chapter describes the iterative development of the Nudge system and a semester-long study in a large chemistry course to evaluate its efficacy as an operant probe.

4.2 Background Theory

Not all students are studious. Many students cram (see Benjamin & Bird, 2006), although there is abundant evidence that spacing learning leads to better retention (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). Cramming is often due to procrastinating, which from 46% (Solomon & Rothblum, 1984) to 95% (Ellis & Knaus, 1979) of college students do regularly. While more senior students may have college more figured out, they also procrastinate more (Semb, Glick, & Spencer, 1979). Procrastination is largely due to fear of failure (50%) and to averseness of the task (18%; Solomon & Rothblum, 1984) and has been associated with a variety of difficulties, including test anxiety, missed deadlines for assignments, poor semester grades, depressed affect, low self-esteem, and social anxiety (e.g., Beswick, Rothblum, & Mann, 1988; Ferrari, 1991; Ferrari et al, 1995; Lay, 1986, 1987; Lay & Burns, 1991; Solomon & Rothblum, 1984). Little surprise then that most students see their procrastination as a problem they would like to eliminate (Solomon & Rothblum, 1984).

Time management is difficult for students, but an important factor in their success. In a longitudinal study of cumulative GPA, a regression with time management skill and SAT scores showed time management to be a better predictor of GPA four years later (Britton & Tesser, 1991). Time management is made difficult by the human susceptibility to “planning fallacy”, the tendency for people and organizations to underestimate how long they will need to complete a task, even when they consider their previous under-estimates (Kahneman &

Tversky, 1979). One technique for abating the planning fallacy is to decompose the task, and this technique is more effective for tasks of greater complexity (Kruger & Evans, 2003).

Students don't choose their study behaviors based solely on the largest direct benefit to their learning (Thiede & Dunlosky, 1999). For example, students often use self-testing not as a learning activity but to diagnose their learning. Ironically, in a study comparing flash card practice with reading, students generally believed that more instruction (reading) would produce more learning, but chose flash card practice most frequently (Kornell, 2009). In another study, students reported that they went to lectures before reading their textbooks, despite thinking that reading the textbook and then going to class was more effective, probably because they also rated reading the textbook first as more difficult than going to the lecture first (B. G. Lee, 2006).

Despite these dissociations between beliefs and behaviors, students can be taught to be self-regulated learners. A classroom-based intervention study by (B. J. Zimmerman, Moylan, & Hudesman, 2011) showed struggling math learners how to self-reflect (i.e., self-assess and adapt to academic quiz outcomes) more effectively. Students receiving self-reflection training outperformed students in the control group on instructor-developed examinations and were better calibrated in their task-specific self-efficacy beliefs before solving problems and in their self-evaluative judgments after solving problems. The self-reflection training also increased students' pass rate on a national gateway examination in mathematics by 25% in comparison to that of control students (B. J. Zimmerman et al., 2011).

4.3 Core Features

Nudge began as an intention to develop a scalable software application to address the need perceived by both students and instructors to support students' time management. Following the fieldwork, I had established several design requirements for the application,

1. Require no upfront action by the student in order to benefit
2. Require no changes to the instructor's curriculum or schedule
3. Require little or no time from the instructor to offer in her course

Through a wide review of the relevant theoretic literature, I settled on several features for the system (Table 4-1). The first feature is to transform the course syllabus to organize course activities by date. The rationale for this was that explicit and salient dates more likely to be met, based in the findings that external deadlines boost task performance more than self-determined deadlines (Ariely & Wertenbroch, 2002) and students generally do whatever's due soonest (Kornell & Bjork, 2007). The second feature is to break down course study activities into smaller actions, such as turning an exam date into a series of tasks like "review lecture notes" and "take a practice test" each due well before the

exam itself. The rationale for this was that decomposition of tasks improves time allocation and decreases aversiveness, based in the findings that smaller tasks abate the planning fallacy (Forsyth & Burt, 2008; Kruger & Evans, 2003), students procrastinate largely due to fear of failure (Solomon & Rothblum, 1984) and that in shared task lists, vague information preferred (Blandford, 2001). The third feature is to help students maintain and track their assignments, study time and progress through the course. The rationale was that recording task status increases awareness and inclination, based in the finding that self-monitoring of study behaviors improves learning (Richards, 1975). Table 4-1 lists two additional features to motivate students through rewards. These were derived from the theoretical literature but never implemented.

4.4 Iteration

With these core features defined, I then developed Nudge through a series of successive iterations, driven by field observations and theory. Nudge evolved from scenario sketches (see Chapter 3) to core features (above) to paper prototypes to graphic mockups (shown in Figure 4-1) to a production prototype (shown in Figure 4-2) for use by students in a real classroom setting.

In this production version (Figure 4-2) students log into the system to see a dashboard of all the tasks for the whole semester. They are laid out in a table with columns indicating the milestone for which they should do the task (e.g. Exam 1 or Lecture 8), its importance (e.g. Required, Advised, or only If Needed), the expected time the task will take, a description of the actual task with a link to resources needed to carry it out (like the homework web page), an indication of their currently reported status (e.g. S for started) and when will be or was due. Students can filter any of these columns, as in a spreadsheet but with more relevant categories. For example, the filter on the Due column has options for *Ever*, *Soon*, and *Past Due*. The status column has a filter for what's left *To Do*.

A progress dashboard is accessible by a link at the top of the work list (Figure 4-3). Here students can see quantitatively how much work they've done and how much is left. They can compare the counts of their status reports, for example how many they've finished versus skipped. They can also see what proportion of tasks they've completed with each importance. For example, 4 / 14 required tasks due so far. Finally they can review their report on each task with the time spent and any notes to self.

The production prototype was programmed with the Ruby 1.9 programming language, the Rails 3.0 (and later 3.1) web development framework, HTML5 document object model and SCSS for CSS3 document styling. The system was run on a Heroku platform-as-a-service dyno instance.

Nudge was first evaluated in a lecture course of 95 students in the spring of 2011. It was introduced in the 10th week and one quiz grade was replaced with points for how much they reported into the system about what they had done. The effects of Nudge were measured by normalizing scores on each exam, averaging

these z scores for the two pre-Nudge exams and the three post-Nudge exams, and comparing pre- with post-. Students who used Nudge when it was first made available in the 10th week ($N = 9$) saw their exam scores go up 0.36 sigma while students who started using Nudge in the 12th week ($N = 12$) saw their exam scores go down 0.31 sigma. Of course, this is not a true experiment, and that is why the formal evaluation below manipulates students' Nudge experience through randomly assigned experimental conditions.

While none of the features was evaluated in a controlled way, observations of their use revealed some factors that informed the next design iteration of Nudge.

- Students did not like logging into the web site. The experimental semester-long evaluation iteration of Nudge optimized for email-based interaction (Figure 4-4 Screenshot of an email sent in All-messages condition). Each email was a web form with radio buttons to indicate task status. At the bottom of the email users clicked a Record button to submit the data to the server and view the progress dashboard screen. Students asked that it be integrated with Blackboard but after months of effort investigating the technical and logistic requirements of the university, integration was abandoned as technically and logistically infeasible.
- The dashboard to track course progress was used regularly by few and most never interacted with it (Figure 4-3 Screenshot of course progress screen). This was left in the next iteration but not improved upon.
- Student's self-reports of time on task were very noisy. Systems that prompt for time on task, should be careful to motivate and support students in accurate reporting. In the next iteration this feature was left but not emphasized.
- Some notes in the "notes to instructor" field were feedback on the difficulty of assignments, but most were empty statements to mechanically maximize participation points.
- The instructor never spontaneously looked at any of the reports and few students expected them to. E.g. "The system seems ambiguous in terms of feedback. I don't think that my instructor will look at any comments, so I don't write any for them specifically." Student feedback that is never read may harm trust in the system and instructor. Such features should provide indications of whether feedback has been read. They may also need to push reports to the instructor. However this feature wasn't changed from the pilot.

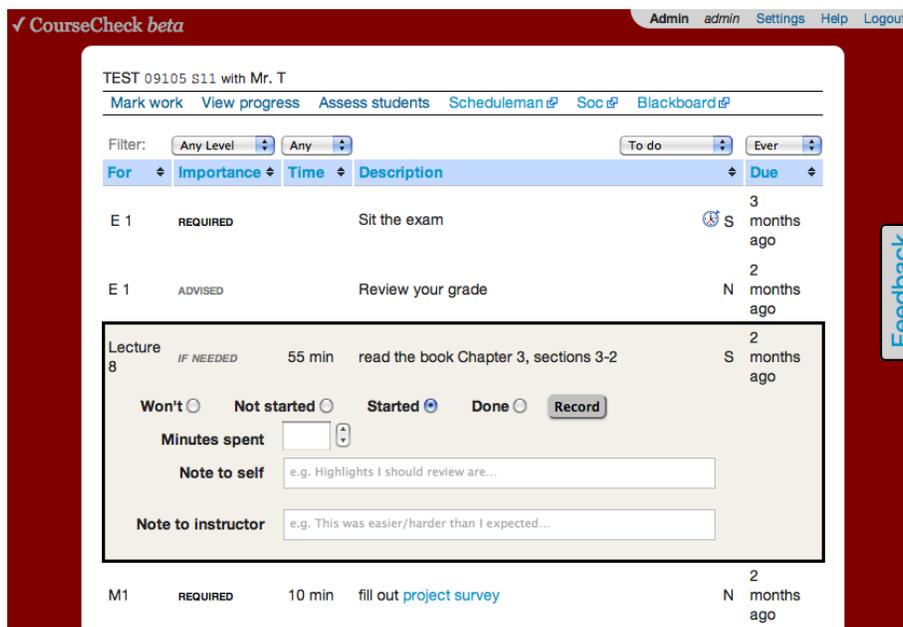
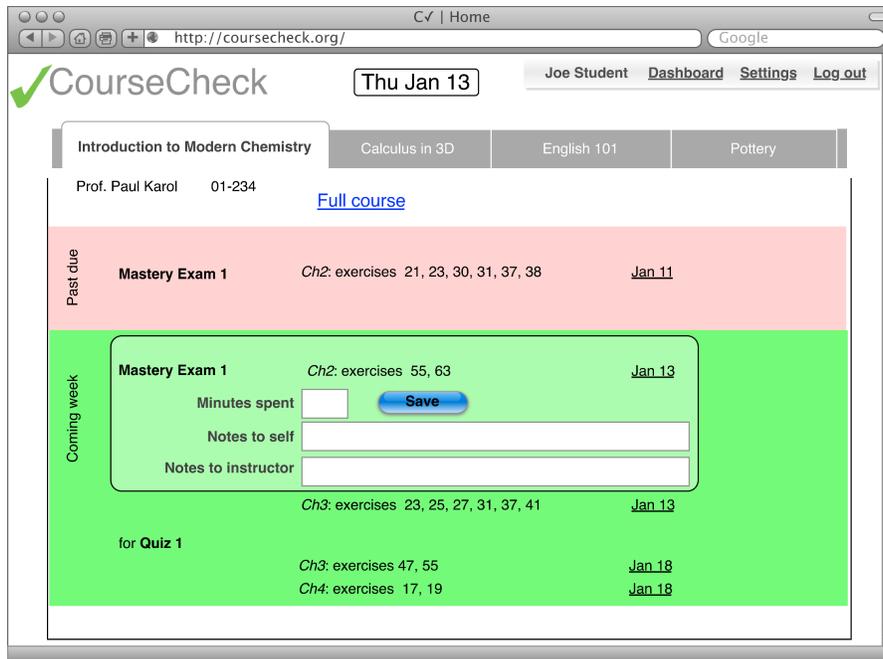


Figure 4-1: Nudge mockup

Figure 4-3 Screenshot of course progress screen

| | | | |
|-----------------|----------------|------------------|----------------|
| Done | Started | Not started | Skip |
| 3 | 1 | 2 | 3 |
| Required (done) | Advised (done) | If needed (done) | Offered (done) |
| 4 / 14 | 5 / 43 | 0 / 3 | 0 / 0 |

| Milestone | Task | Minutes spent | Note to self |
|-----------------|--|---------------|--------------|
| HW1 | Ch2: 21, 23, 30, 31, 37, 38 | | |
| Participation | Take the concepts quiz in first lecture | 20 | |
| HW2 | Ch12: 18, 49, 55, 57, 75, 77, S1 | | |
| HW7 | Do and submit HW7 | | |
| HW7 | Ch14: 45, 67, S8 | | |
| HW8 | Ch5: 27, 31, 67, 83, 89 | | |
| Participation | Fill out mid-semester questionnaire required for study (if enrolled) | | |
| Mastery exam IV | Practice Mastery problems | | |
| Participation | Fill out 3rd exam questionnaire about CourseCheck | | |

Figure 4-2: Screenshot of task list in pilot and final evaluation

CourseCheckup

Hi, [redacted]. Here are some things to do for your courses. Please mark your progress on them. Press *Record* at the bottom to enter them into CourseCheck.

Introduction to Modern Chemistry

| For | Importance | Time | Description | Due |
|---|------------|------|---|---------------|
| | Skipped | | Practice Exam IV problems | 2 days ago |
| HW9 | REQUIRED | | Do and submit HW9 | 2 days ago |
| <p>Skip <input type="radio"/> Not started <input checked="" type="radio"/> Started <input type="radio"/> Done <input type="radio"/></p> <p>Minutes spent <input type="text" value="0"/></p> <p>Note to self <input type="text"/></p> <p>Note to instructor <input type="text"/></p> | | | | |
| | Skipped | | Ch19: 45, 47, 49, 77 | 2 days ago |
| HW10 | REQUIRED | | Do and submit HW10 | 7 minutes ago |

Figure 4-3: Screenshot of an email sent in All-messages condition

Table 4-1: Nudge feature matrix

| Feature | Claim | Warrant | Status |
|---|--|---|-------------|
| Course task assigned to dates and organized centrally | Explicit and salient dates more likely to be met | External deadlines boost task performance more than self-determined deadlines (Ariely & Wertenbroch, 2002) Students generally do whatever's due soonest (Kornell & Bjork, 2007) | Implemented |
| Break-down of study activity into smaller actions | Decomposition of tasks improves time allocation and decreases aversiveness | Smaller tasks abate the <i>planning fallacy</i> (Forsyth & Burt, 2008; Kruger & Evans, 2003) Students procrastinate largely due to fear of failure (Solomon & Rothblum, 1984) In shared task lists, vague information preferred (Blandford, 2001) | Implemented |

| Feature | Claim | Warrant | Status |
|---|---|--|---------------------|
| Maintaining and tracking assignments, study time and progress | Recording task status increases awareness and inclination | Self-monitoring of study behaviors improves learning (Richards, 1975) | Implemented |
| Reinforcement of effort demonstrated | Ss will spend more effort when effort itself is rewarded | Rewards on student effort can enhance achievement-directed effort (Brophy, 1987) Task-orienting strategies facilitate performance of Ss who de-emphasize role of effort (Stipek & Kowalski, 1989) | Not yet implemented |
| Surprise challenges and intermittent accolades | Game-like features increase fun | Intermittent rewards more motivating (Alberto & Troutman, 2008) | Not yet implemented |

4.5 Experimental Design

With the experience of the pseudo-experimental pilot, parts of the Nudge system were improved (as discussed above) and an in vivo randomized controlled experiment was designed to evaluate Nudge as an operant probe. This formal study tested whether Nudge fit the context, achieved its desired effects, and could provide data to inform models of how its affects were achieved.

4.5.1 Context

The study took place in one section ($n = 136$) of a large introductory chemistry course at a competitive private university. The instructor used Blackboard and a personal web site to provide students with a calendar of lectures and assessments, and regular announcements. The data collection and system intervention took place over the whole semester (Figure 4-5).

4.5.2 Task list

The course syllabus was recomposed into 60 tasks (14 required, 43 advised and 3 supplemental), which are all listed in Appendix E. In this evaluation the conversion was rather formulaic so it didn't require any domain or metacognitive knowledge. First I entered each assessment into a spreadsheet with its date, marking those tasks as *required*. Then I entered several ways to prepare for the assessment and marked them as *advised*. For homeworks these were simply the problems recommended by the instructor from his homework assignment listing. For exams, these were to take practice exams and review notes. The most time consuming part was to encode the web links to the resources to use for studying

(e.g. linking to the actual practice problems online). In the table of tasks, each bracketed string was such a link. The expected time for completion for each task was very difficult to estimate and omitted from most tasks.

Entering the tasks into the system takes negligible time. As the software developer, I was able to enter the tasks into the system from this spreadsheet programmatically in minutes, taking less than a half hour total. The current authoring tool could require a novice user up to an hour for the task decomposition and entry, but the authoring interface is a rudimentary prototype. With optimization to speed entry and scaffold the elaboration of the syllabus, a novice could produce a better task list in less time. Because the entry doesn't require an expert understanding of the course, they could also outsource it. In a casual evaluation, Mechanical Turk workers typed in information from PDF syllabi for \$1 and wrote by email to ask for more work.

4.5.3 Conditions

I evaluated the effects of Nudge by randomly assigning students to *all nudges* and *no nudges* conditions. Students in the *all nudges* condition were sent every task before it was due, grouped in emails sent at least weekly (e.g. Figure 4-4). Each email prompted students to reply with their status of completing each task (skipped, not started, started or completed.) Students in the *no nudges* condition were not sent any reminders before tasks were due. To collect data on their work for the course, after each exam they received one email with all the tasks and were prompted to indicate their completion status for each. All students also filled out questionnaires before, during and after the semester's instruction.

4.5.4 Hypotheses

Nudge was expected to help students' study time allocation and grades.

H-allocation

Students sent all Nudge messages exhibit better time use than students sent no Nudge messages.

The theory of operation of Nudge is that it helps students study more effectively by scaffolding, and ultimately causing, more effective allocation of time to study activities. For example, it should help students to review lecture notes immediately after the lecture to verify and repair their understanding.

H-grades

Students sent all Nudge messages perform better on assessments than students sent no Nudge messages.

Through better allocation of study time, students will learn more and will ultimately perform better on course assessments.

H-disposition

Students with poor study time use benefit more from Nudge messages.

The ethnographic observations were that some, but not all, students have difficulty with knowing how to study well. Nudge is designed for these students with poor study dispositions and is not expected to help as much students with good study habits.

4.5.5 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 that was cumulative and could replace a student's lowest non-cumulative exam grade.

4.5.6 Explanatory measures

Each student's personal attributes affect how she uses Nudge, which in turn affect how the tool affects her and her learning. Toward understanding how the tool works differently for different students, I logged user activities and collected several large questionnaires over the term.

Behavioral measures include students' interactions with Nudge, the data they reported through Nudge, and questionnaires about their time and study behaviors.

Study time allocation was operationalized as the Time/Environment scale ($\alpha = .71$) of the Motivated Strategies for Learning Questionnaire (Pintrich, 2002; Pintrich, Smith, Garcia, & McKeachle, 2001). The scale has eight items and some were adapted to target math and science classes. E.g. "I make good use of my study time for math and science courses."

To see how students' goals in the course mediated their use and performance, the questionnaires also included several standard measures of goals. Of note in this analysis are the measures of the 2x2 achievement goal framework (Elliot & McGregor, 2001). This framework distinguishes students' conception of competence by two dimensions: personally mastering a domain versus demonstrating performance (definition dimension) and whether they are oriented to approaching success or avoiding failure (valence dimension).

4.5.7 Attrition and Missing Observations

11 students signed up for the study, but never did any coursework and were omitted from all analysis. 7 (13%) were in the *all nudges* condition and 4 (8%) were in the *no nudges* condition. The difference is not significant.

Of students who started the course, 2 (2.2%) dropped before the end (one from each condition). They are included in analyses for which their data are available.

4.5.8 Timeline

To help interpret the following results, Figure 4-5 Timeline of Nudge study shows a timeline of the course, assessments, and questionnaires and when changes were made to Nudge.

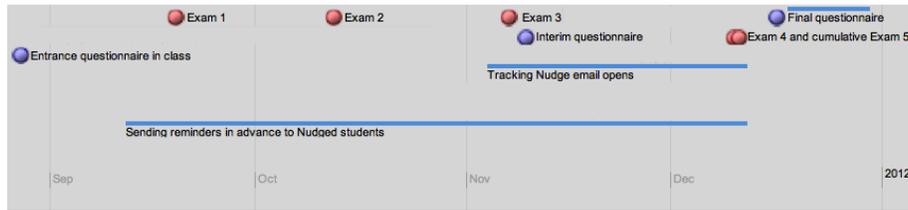


Figure 4-4: Timeline of Nudge study

4.6 Results

4.6.1 Descriptive statistics

Table 4-2 presents mean learning outcomes and pre/post time management scores by condition.

Table 4-2: Incoming attributes and outcomes

| Group | Study time habits score (pre MSLQ: T/E) | Exam scores (Exams 1-5 mean) | Study time habits score (post MSLQ: T/E) | Passed course |
|--------------------------------|---|---|---|----------------|
| No nudges (<i>n</i> = 48) | 5.2 (<i>n</i> = 42, <i>sd</i> = 0.7) | 70.9 (<i>n</i> = 43, <i>sd</i> = 11.9) | 4.8 (<i>sd</i> = 0.8) | 92% (44/48) |
| All nudges (<i>n</i> = 45) | 5.3 (<i>n</i> = 40, <i>sd</i> = 0.8) | 69.5 (<i>n</i> = 41, <i>sd</i> = 12.3) | 4.9 (<i>n</i> = 31, <i>sd</i> = 1.0) | 93% (42/45) |

Subjective rating A questionnaire was given after the 3rd exam asking the usefulness of several features of the course. 75% of respondents who received all nudges (*n* = 28) rated “Email reminders about course work” as “Good” or “Great”. Five (17%) didn’t perceive it as useful and two chose “didn’t know about it”, even though the survey was sent to the same addresses as the Nudge messages. Among students not in the study, who could choose whether to receive

Nudge messages or not, 20% (14/70) opted out and a few opted to reduce their rate but continue to receive weekly messages.

About 40% of students responding to the final questionnaire agreed with the statement, “The reminder emails helped me in the class” (13/32). About 46% of Nudged students responding to the final questionnaire agreed with the statement, “I wish I could have email reminders for all my classes” (15/32), even though some of these students disagreed that it helped them in this class. 44% (14/32) agreed with the statement, “Without the reminders I would have forgotten to do something.” Again some of these students disagreed with the previous statements.

In a measure of overall course satisfaction, students rated their agreement with “I achieved my goals for the course.” The main predictor, not surprisingly, was their grade. Accounting for average exam grade ($p < .0001$), Nudged students agreed more ($F(1, 58) = 5.0, p = .029$). To see if this was due more to expectation or outcomes, a second covariate was tested, their responses on the midterm questionnaire indicating the final grade they expected to receive ($p = .002$). Nudged students were sure to agree ($p = .018, 95\% \text{ CI } [0.09, 0.97]$) and not Nudged likely to disagree ($95\% \text{ CI } [-0.67, 0.22]$).

Nudge usage

Table 4-3: Nudge reception

| Group | Evidence of opening email (email image tracker in 4th quarter) | Evidence of opening email repeatedly (same message) | Number of messages opened (among evidence of opening) | Proportion opened more than two messages (among evidence of opening) |
|------------|--|---|---|--|
| No nudges | 38% (18/48) | 25% (12/48) | 1.6 (out of 2, $n = 18, sd = 0.5$) | - |
| All nudges | 80% (36/45) | 22% (10/45) | 3.6 (out of 12, $n = 36, sd = 3.3$) | 36% (13/36) |

Table 4-3 presents usage measures of Nudge by condition. Nudged students opened about 80% of the messages sent to them once the email image tracker was in place. This is a floor estimate because some students may not have had told their email client to load the images which the tracker required. Only 38% of no-nudge students appear to have opened an email, but they also had only two occasions versus 12 for the nudged students (after the tracker). Over 22%

of both groups opened some email messages repeatedly. The mean number of messages opened by nudged students was 3.6 out of 12, with a mode and median of 2.

Table 4-4: Nudge replies

| Group | Replied to task polls ever | Median proportion of completes in <i>non-required</i> tasks | Median proportion of completes in <i>required</i> tasks | Agreement with “What I enter is accurate”. (7pt Likert) |
|------------|----------------------------|---|---|---|
| No nudges | 83% (40/48) | .48 ($n = 40, sd = 0.27$) | .87 ($n = 40, sd = 0.25$) | 5.5 ($n = 26, sd = 1.4$) |
| All nudges | 87% (39/45) | .20 ($n = 39, sd = 0.22$) | .75 ($n = 39, sd = 0.26$) | 5.8 ($n = 29, sd = 1.3$) |

The no-nudge students received Nudge tasks after the exams and were asked to reply with their task status then. The rate of reply was roughly equal between the groups, but the no-nudge students reported higher rates of completion of the tasks, especially on non-required tasks. Because the two conditions were measured differently, and 15% of participants never replied to any task poll, the response data is not used in the following analyses. However it’s worth noting that there was no difference in students’ agreement with “The data I reported were accurate.”

4.6.2 H-allocation

Students sent all Nudge messages exhibit better time use than students sent no Nudge messages.

For this hypothesis, the originally intended operationalization of study time allocation was students’ reports of completion of advised tasks. However this measure is confounded with the different time and context of the task status polling between conditions. Nudged students were polled periodically in small batches before due dates while non-Nudged students were polled in large batches after all the due dates related to an exam. Table 4-4 shows that nudged students not only reported far fewer completions of non-required tasks, they also reported fewer completions of *required* tasks (such as turning in a graded homework assignment). I take this to be an effect of when and how they were polled, making comparisons between conditions on these measures uninformative.

Table 4-5: Reported hours spent on different activities

| Week type | Group | Attending lecture | Attending recitation | Reading the book | Reviewing notes | Studying and solving problems |
|-----------|----------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-------------------------------|
| Regular | No nudges ($n = 34$) | 2.5 ($SD = 0.7$) | 1.6 ($SD = 0.6$) | 0.9 ($SD = 1.0$) | 0.8 ($SD = 1.0$) | 2.2 ($SD = 2.0$) |
| | All nudges ($n = 33$) | 2.3 ($SD = 0.7$) | 1.9 ($SD = 0.7$) | 0.8 ($SD = 1.0$) | 1.0 ($SD = 1.0$) | 2.3 ($SD = 1.5$) |
| Exam | No nudges ($n = 34$) | 2.4 ($SD = 0.7$) | 1.7 ($SD = 0.7$) | 1.2 ($SD = 1.1$) | 1.3 ($SD = 1.0$) | 4.2 ($SD = 1.8$) |
| | All nudges ($n = 33$) | 2.4 ($SD = 0.6$) | 2.0 ($SD = 0.8$) | 1.1 ($SD = 1.2$) | 1.7 ($SD = 1.2$) | 4.6 ($SD = 2.0$) |

The best available measures of student time use are their reports in a questionnaire at the end of the course asking the hours per week they spent on different activities (see Table 4-5). The response levels were ordinal and recoded to continuous (“Less than 1 hour”=0.5, “1-2 hours”=1.5, “2-3 hours”=2.5, “>3 hours”=3.5). Nudged students reported spending more hours in a regular week attending their recitation sections ($F(1, 64) = 5.5, p = .023$). During an exam week the difference was only marginally significant ($p = .08$). They also spent marginally more time reviewing notes during exam weeks ($F(1, 68) = 3.0, p = .086$) and regular weeks ($p = .074$ one-tailed), however this may be an artifact of multiple comparisons. No significant differences were observed on time spent attending lecture, reading the book or studying and solving problems.

4.6.3 H-grades

Students sent all Nudge messages perform better on assessments than students sent no Nudge messages.

There were no main effects of nudge messaging observed on exam performance.

4.6.4 H-disposition

Students with poor study time use benefit more from Nudge messages.

In a model of the interaction of nudging with students incoming Time/Environment dispositions score, better time management led to better exam scores ($F(1, 76.9) = 6.4, p = .014$) but Nudge interacts to help students with poor

management ($F(1, 76.9) = 4.6, p = .036$). That is, Nudge may compensate for poor time management dispositions. Digging deeper, I account for math aptitude ($F(1, 63.7) = 20.4, p < .0001$), the number of email messages opened within each nudge condition ($F(2, 62.4) = 3.0, p = .059$) and its interaction with the Time/Environment score ($F(2, 62.6) = 5.1, p = .009$). In this model, the worse a student's time use the greater the benefit of opening each Nudge message. For students with the highest Time/Environment scores, the number of messages they opened had no relationship with their exam scores. For students with the lowest Time/Environment scores, a predicted exam score of 58% would be 61% if they opened one message and by opening six messages they could match the predicted score of the best time managing students who opened none (76% on exam with 720 math SAT). Opening all twelve messages predicts a 90% score in this model. Because the number of messages opened is subject driven and not experimentally manipulated, it cannot be claimed to be the cause of the higher scores. It could instead be evidence of a third cause, which is the student's motivation to succeed in the class. However, the fact that for students with good time skills their exam scores exhibit hardly any relation with the number of messages they opened suggests that the Nudge system created an opportunity whereby motivated students with poor time management could overcome this deficit.

4.6.5 Student perceptions

Different students perceived the system differently, but in several pretty consistent themes. In the final questionnaire, students were asked, "If the email reminders were a person, what kind of person would it be?" The responses mostly described one's relationship to the person. "My mother" was echoed by in several responses. "They would be sort of like the mother that's always around and making sure you're on top of your school work." From another,

If the email reminders were a person they would be very annoying and nosy, but good-hearted. They would be something of a mother, always checking up on you and wondering how much you have done, and even though you may at times get irritated, you would never give up a friendship with this individual.

Several echoed the idea of a friend, e.g. "A close friend who is pushing me" and "My best friend. The only person that would tell me to get my work done because I tend to forget about assignments sometimes. This person would be on top of their work as well and displayed great academic success."

A more negative theme was a well-intentioned but annoying and dense nag. E.g. "the email reminders would be a person that did not really care about what they were saying. they need to be more upbeat or motivating and something thst [sic] someone would actually listen too." Many said simply, "persistent", and some explained that eventually they ignore such formulaic persistence: "That person who runs by your house every day to the point when eventually you stop

noticing him” and “They'd be the kind of person you say 'Hi' to because you feel obligated when passing them on campus, but in reality, you do not associate with that person in any way.” Some students expressed that despite being annoying, you appreciate this person:

They would be a person that slightly annoyed me, but more because I wouldn't want him/her pointing out my flaws or what I've missed/forgotten. In the long run, I'd appreciate that person a lot, as he/she helped me and kept me on track.

4.6.6 Feature Validation

The features of Nudge were based in a few theoretically derived design claims. How does the evidence from this evaluation support or cast doubt on these claims? Because all the features were tested together none has rigorous evidence either way, but some observations from student use may help inform them.

Explicit and salient dates more likely to be met This was implemented through organizing course tasks by date and regularly emailing students with what tasks were coming up (e.g. a homework or exam). The results are consistent with an interpretation in which this claim is true. However some caveats to the rationale are that while students generally do whatever's due soonest (Kornell & Bjork, 2007), many of the “due” dates in the evaluated course were not conventional due dates when a student must do something or lose points. For example, the “take a practice exam” task due a few days before the exam can be ignored with no direct consequences. It's unclear how much students bought into these dates as dates by which they should do the task. The fact that the dates were external (Ariely & Wertenbroch, 2002) may have motivated some students, but it may have led other students to distrust the dates. Some students expressed some of the “due” tasks had no value to them: “there are some assignments that are not required that I don't feel are necessary for my understanding of the material.” To increase adoption, it is worth considering allowing students to customize the due dates. To maintain the externally imposed nature, the dates and tasks could be part of a high level setting of, for example, high/medium/low effort, or more concretely a range of hours per week they will allocate to the course. With access to their course grades it could help them meet a target grade by adapting to their course performance.

The salience of the dates is another issue to better explore. While no students expressly commented on the choice of dates, many were bothered by the frequency of the reminders. “I do not like the everyday emails, because they are sometimes excessive.” One student suggested, “Instead of daily reminders maybe biweekly reminders.” However, one student actually requested a higher frequency: “I would like if the check ups were more often, or if it could send reminders a few hours before homework is do.” It seems the reminders are annoying until they save you: *What I like most about [Nudge] is that it lets you know when you have an assignment due. Usually, I am very good at remembering due dates, but one*

week, I forgot to write down my homework for the week and was therefore under the impression that nothing was due that week. It wasn't until I got an email reminder from [Nudge] that I had assignments due that I realized my mistake.

One way to address this feedback would be an escalating nag factor as the deadline approaches, customized to the user's preference.

Decomposition of tasks improves time allocation and decreases aver-siveness This was implemented by breaking the course tasks, such as taking an exam, into a series of smaller actions. The fact that the effects of the messages interacted with students' time management skills lends support to this design claim and the earlier findings that smaller tasks abate the planning fallacy (Forsyth & Burt, 2008; Kruger & Evans, 2003). The study produced no evidence related to the rationale that vague task information is preferred (Blandford, 2001) but there is data to affirm the idea that students procrastinate largely due to fear of failure (Solomon & Rothblum, 1984).

The questionnaires at the beginning and end of the semester included measures in the 2x2 achievement goal framework (Elliot & McGregor, 2001). The Performance Avoidance goal, for example, is measured by agreement with statements like, "My fear of performing poorly in this class is often what motivates me." Students receiving all the Nudge messages ended up with a stronger orientation to the performance avoidance goal ($F(1, 91) = 4.56, p = .035$), accounting for their earlier rating ($p < .0001$). This goal orientation is a consequence of, but distinct from, a fear of failure (Bartels & Magun-Jackson, 2009; Elliot & McGregor, 2001). It is possible that the effects of the Nudge messages work through increasing students' fear of failing in the class while decreasing their fear of failing in any particular study task. Is this good though? Performance avoidance goals have been found to be negatively correlated with learning outcomes and cognitive self-regulatory activities ("Goals and Goal Orientations," 2008). As a mechanism, performance avoidance goals have been found to be a positive predictor of surface learning (Liem, Lau, & Nie, 2008). This points to an alternative explanation for the interaction of Nudge with time management skills. For students with poor skills, increasing their attention to surface learning activities may have produced a net gain in attention to the course. For more studious students, this greater attention to the course tasks (surface features of learning) may have been at the expense of deeper cognitive engagement, supplanting their own strategies for regulating their learning with those of the course syllabus. Can fear of failure be channeled into deeper learning? Further yet, can time management be supported while inducing a more productive achievement orientation (e.g., performance approach goals, or even mastery approach goals)? Both of these questions highlight important design spaces to explore.

Recording task status increases awareness and inclination This claim was implemented by prompting students to record their task status. In the current design, students not required to record as part of the study did not

record. Because the task status records are so noisy, there is no quantitative measure of specifically whether recording tasks increased their awareness and inclination per the rationale that self-monitoring of study behaviors improves learning (Richards, 1975). However, qualitative data make clear that some students valued the progress monitoring enabled by task recording. “I like the way it keeps track of your progress” and “Sometimes it is nice to look back and see how many of the assignments I have completed.” Specifically regarding motivation, “It makes me realize how much I should spend doing my work” and “I like that it makes me feel accomplished since I get everything done on time.” For one student, the “not started” status option was their favorite feature of the system. “I like the fact that it gives you an opportunity to answer the questions very honestly with the ‘not started’ option.”

4.7 Limitations and Opportunities

4.7.1 Operation on desired outcomes

The usage rates were low. 20% of nudged students never opened the emails. Some of these may have just not opened them during the 4th quarter when the email tracker was in operation or had their email clients set to not load images. However among people who definitely did open, 64% only opened one or two of the twelve sent. Both of these limitations suggest that some students do not take the treatment. Clearly the system should be more tailored to each student’s dispositions and course performance. How exactly is an open question. If it’s completely optional, then poor students may not see the need for it. If it’s mandatory, it may hinder some students. In future work, I would explore the potential for motivating participation, beginning with the last two features in Table 4-1 that have theoretical support but have not yet been implemented.

There was no overall effect of the Nudge messages on student learning and for students with good time management they may have even been counterproductive. The messages appear to help students with poor time management (i.e. those in need of help) but this experiment didn’t have enough power to provide evidence for a main effect in this subpopulation. Future work should study nudge messaging where a larger proportion of students have poor time management. Further, Nudge messages increased students’ performance-avoidance goal orientation, which is generally predictive of worse learning and self-regulation. This may explain the negative impact on good time managing students. Future work should provide better messages, matched to the needs and proximal abilities of different learners.

A key way to improve the impact of the Nudge messages is to improve the messages themselves. The set of tasks defined in this course were limited and did not specify all the good study activities to do well in the course. In the formally evaluated version, there were no reminders for instruction, only practice. There is a body of literature on best study practices that could be operationalized into ideal tasks and messages. What this study demonstrates is that the Nudge

system works in vivo and is easy to deploy. The Nudge operant probe provides a new mechanism for research to test these study ideals in real-world settings and discover the distribution and boundaries of their effects with different curricula and students.

While Nudge is domain general, the messages had to be authored for the course in the study. Adding new curricula to nudge could be facilitated by having a set of templates to elaborate task structures around different typical course events. For example, each exam could have associated with it: reread notes (6 days before), study worked examples (8 and 4 days before), take practice exam (3 and 1 days before), and attend review session (2 days before). Were researchers to test their theories of optimal practices, these could become standard task expansions of traditional course events. Adding courses could be as simple as uploading a syllabus, scanning for key dates (homework due, quiz or exam given) and automatically expanding them into task sets to generate a task set for the whole course. These would certainly be improvements, but not necessary for adoption. The instructor in the study, was asked after seeing the results whether he would take the time to type in the dates to use it again and replied, “Yes, very much. I would say emphatically.”

4.7.2 Probe data for modeling

One aspect of the probing utility of Nudge that didn’t work as hoped was the set of student responses about what tasks they had done. There was a confound for comparing response between conditions, but that could be addressed easily in future work by polling the same times and ways. More problematically, it’s not clear how accurate the task reports are. 41% of students “strongly agreed” to “What I enter is accurate” but 28% didn’t agree, and that’s among the 58% who took the time to respond to that questionnaire. Conservatively, only a quarter of students in the study strongly agreed to having entered accurate task reports. For these data to be useful in modeling student study behaviors naturally, the design of acquiring them needs to be greatly improved. An important factor is students’ incentives for entering accurate data (or any data at all). Game-like motivations could help. Reinforcement of effort demonstrated and intermittent rewards are two planned features, supported by theory, that have yet to be implemented in Nudge.

The particular tasks authored define the data that can be collected from students. While there are some tasks or behaviors that may be more effective for student learning, there may be others that are more valuable for student modeling and intervention diagnostics. If the system could elicit accurate reporting from students it would open another area of inquiry. Key factors for modeling students could be added to the task expansions for the purposes of different probing studies.

4.8 Conclusion

Nudge was designed to improve learning outcomes in university lecture courses using observations from the field and theories from existing literature. In a large introductory chemistry course, Nudge helped students with poor time management dispositions to learn and perform better on course exams. The benefit to such a student was greater the more of the Nudge emails they opened.

The process of designing Nudge helps light the way for the design of similar systems. Explicit and salient dates may be more likely to be met but they can be too explicit and too salient to the point of being ignored. Instead they should be due dates and messaging policies that students buy into through a choice they make of what effort to allocate for the course. There was no direct evidence that decomposition of tasks improved time allocation but it may have decreases aversiveness. Students who received all Nudge messages went up in their performance-avoidance goal orientation, an indication of being motivated to avoid performing poorly. This motivation orientation is not best for excellence, but it is an increase in motivation that can help students perform better. There was also support for the design claim that recording task status increased students' awareness and inclination to perform course work. Each of these merits further exploration.

Nudge highlights the opportunity to support students' time management skills to improve their learning. The formal evaluation and analysis of its design principles shine a light on new opportunities for research and real world impact through operant probes for applied learning science.

5 Exemplify: 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: Exemplify 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 Exemplify 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 Exemplify 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 (Alevén & 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 Exemplify.

5.3 Core Features

Exemplify 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 Exemplify 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).

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.

Exemplify 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.

Exemplify 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).

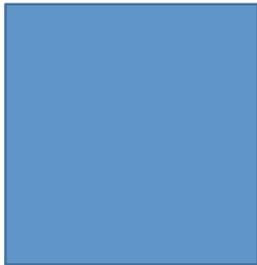
Like Nudge, Exemplify was developed through a series of iterations. The prototypes have many visible states fluctuated widely from iteration to iteration, due to the cognitive complexity of the task. Figure 5-3 shows a screenshot of the PowerPoint prototype, which used animation to reveal the parts of the model solution. Figure 5-4 shows the sequence of screens in the implemented version of Exemplify. 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: Exemplify 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 |

| Feature | Claim | Warrant | Status |
|--|---|--|--|
| Source the explanations from other students | Student explanations may be more effective | Non-experts often make better explanations of work than experts do (Aleahmad, Alevan, & Kraut, 2009) Expert knowledge creates blind spots in instruction (Nathan, Koedinger, & Alibali, 2001) | Implemented |
| 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.

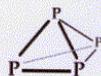


Does the expert's path match up with yours?

- Yes, I'm on the same path.
- No, I took a different path that also works
- No, I was on the wrong path

Figure 5-1: PowerPoint prototype of Exemplify

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.

Why is your work correct?

Check my work

Get help

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)

Ⓟ

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?

Yes and similar.

Yes but different.

Partly right.

Not at all.

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 ?

$:P \equiv P:$ 8

(just like nitrogen, N_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.)

Why did the expert do this?

paste earlier justification private

How good is your explanation?

Poor Good Very good

Do you understand the uncovered work?

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

Continue

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 ?

$:P \equiv P:$ 8

(just like nitrogen, N_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.)

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?

Show another Revise mine

Do you understand the uncovered work?

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

Continue

Figure 5-4 Screenshots of implemented Exemplify

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, Alevan, Renkl, & Schwonke, 2009; Salden, Koedinger, Renkl, Alevan, & McLaren, 2010).

There is no clear consensus on the optimal design for a system with the goals of Exemplify. 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

Exemplify 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).

Screen shots from an example usage of the Solving variant

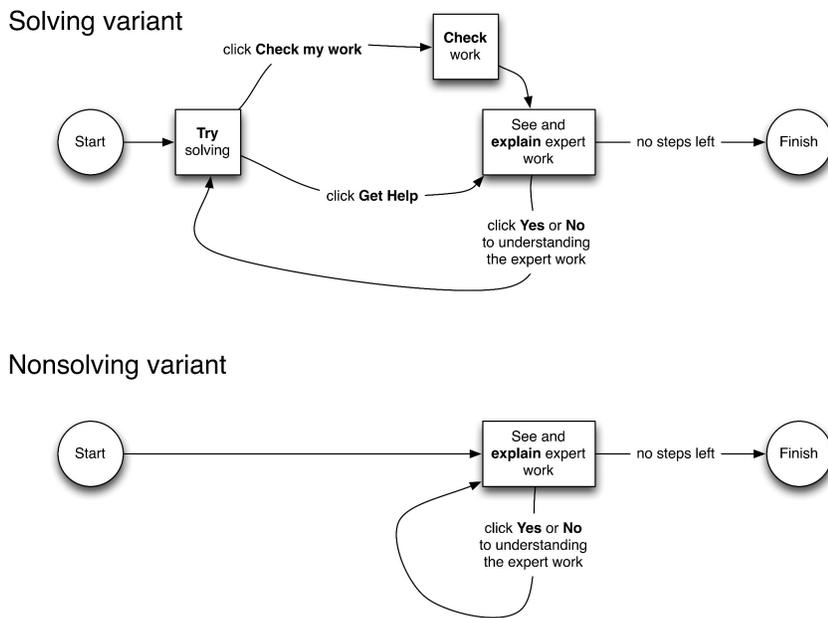


Figure 5-2: States and transitions of Exemplify worked example interaction

5.3.5 Implementation

Exemplify 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 Exemplify. 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.

Problem browser from Exam 3 until end of term

Figure 5-8 Problem browser

from start of semester until Exam 3

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 Exemplify. To add them into Exemplify 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 Exemplify (with solving) to a nonsolving control variant of Exemplify and to a business-as-usual (BAU) control section. Students self-selected into the Exemplify 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 Exemplify section, students had access to the default version of Exemplify with the solving prompt. They accessed the Exemplify 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 Exemplify or an alternate, nonsolving variant of Exemplify. 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 Exemplify 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.

H-immediate

Students with Exemplify 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.

Exemplify 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 Exemplify 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, Exemplify may be more motivating than the BAU static questions and also scaffold better study strategies.

H-delayed

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

Exemplify 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 Exemplify, 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 Exemplify 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 Exemplify section (11%) and 6 in the BAU control section (9%). Within the Exemplify 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 Exemplify study shows a timeline of the course, assessments, questionnaires and when changes were made to Exemplify. 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 Exemplify 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.

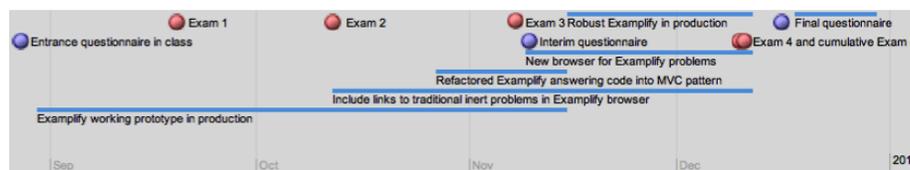


Figure 5-3: Timeline of Exemplify study

5.5 Results

5.5.1 Descriptive statistics

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 Exemplify 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 ($SD = 1.1$, $n = 50$) |
| Exemplify section | 65% (60/93) | 723 ($n = 78$) | 4.5 ($SD = 1.6$, $n = 82$) |
| - <i>Never opened</i> | 33% (1/3) | 730 ($n = 3$) | 5.4 ($SD = 1.9$, $n = 3$) |
| - <i>Nonsolving</i> | 61% (28/46) | 717 ($n = 37$) | 4.3 ($SD = 1.5$, $n = 41$) |
| - <i>Solving</i> | 70% (31/44) | 726 ($n = 38$) | 4.6 ($SD = 1.8$, $n = 38$) |

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

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

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.

Exemplify usage Ninety-seven percent of students in the study opened the Exemplify tool (no difference by Exemplify 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 <i>before</i> exams | Example opens on early topics <i>after</i> 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 ($n = 55$, $SD = 17.8$) |
| Exemplify section | 70.1 ($n = 86$, $SD = 12.3$) | ... | ... | 74.3 ($n = 91$, $SD = 14.1$) | 68.9 ($n = 88$, $SD = 19.9$) | -5.4 ($n = 87$, $SD = 18.0$) |
| - <i>Never opened</i> | 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 ($n = 3$, $SD = 7.3$) |
| - <i>Non-solving</i> | 68.2 ($n = 44$, $SD = 12.5$) | 0.26 ($n = 46$, $SD = 0.77$) | 0.37 ($n = 46$, $SD = 0.77$) | 72.0 ($n = 45$, $SD = 14.8$) | 62.3 ($n = 43$, $SD = 20.0$) | -10.5 ($n = 43$, $SD = 19.5$) |
| - <i>Solving</i> | 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 Exemplify 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 Exemplify 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 ($F(1, 147) = 5.2$, $p = .024$, $d = 0.35$) 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 ($F(1, 87.3) = 5.4$, $p = .023$, $d = 0.36$).

Students with access to Exemplify with solving scored marginally higher than BAU control section students on immediate assessments ($F(1, 145) = 3.1$, $p = .082$, $d = 0.26$) 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 Exemplify 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 Exemplify 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 Exemplify 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).

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

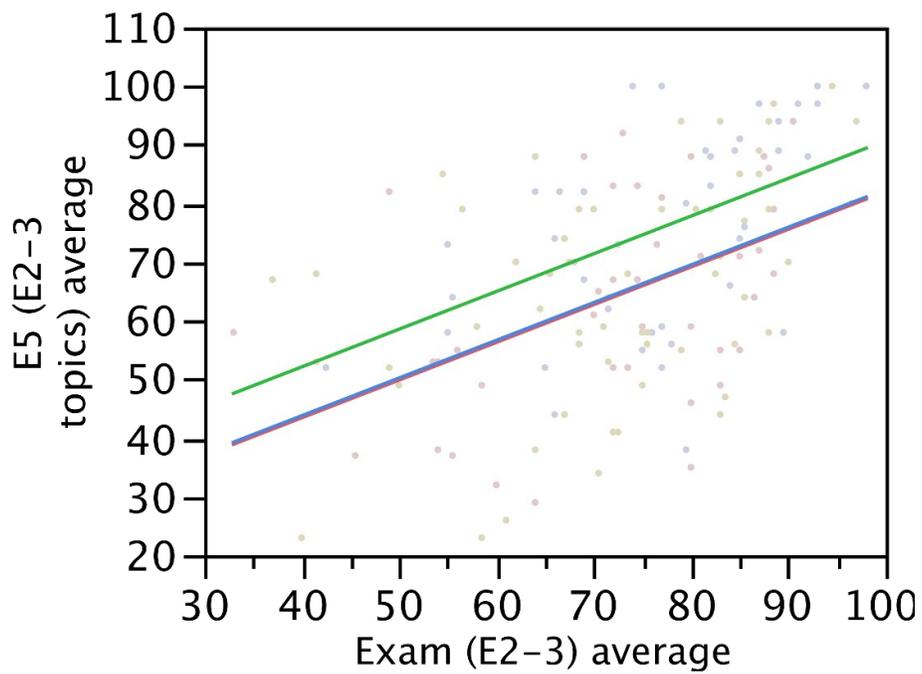


Figure 5-4: Prediction of delayed exam scores by type of study examples available with earlier scores on same topics as covariate. Students with access to solving variant of Exemplify performed significantly better on delayed assessments, taking into account their earlier (superior) performance.

the analysis to students in the Exemplify section who ever opened Exemplify, 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 Exemplify variant ($p = .003$), whether they opened the Exemplify 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 than 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 Exemplify? 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 "Exemplify 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 Exemplify 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 Exemplify 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 Exemplify 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 Exemplify 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

Exemplify 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 Exemplify 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 Exemplify.

Students with the solving variant opened more examples, pointing to a motivational effect. They used Exemplify 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 Exemplify 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 Exemplify with solving leads to better retention than how students otherwise study.

The study did not have measures of student activity outside of Exemplify, 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 Exemplify 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. Exemplify

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 Exemplify 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 Exemplify 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 Exemplify 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 Exemplify 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 Exemplify 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 Exemplify, 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 Exemplify 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 Exemplify 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.

Exemplify 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.

Exemplify is a simple technology that can provide big gains to learning. In a full semester evaluation in a real-world college course, Exemplify 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 Exemplify are drawn from cognitive psychology and are simple to implement and iterate upon.

6 Summary and Conclusions

6.1 Introduction

This work began with two main lines of inquiry: exploration and reflection on design processes for learning sciences research that operationalize theoretical results and are easily adopted *in vivo*, and case studies in applying those processes to the design and rigorous evaluation of systems to support students' study activities in college lecture courses. In reflecting on the processes and outcomes in these cases and others, I developed this work as an instance of a broader concept of Scientific Research through Interaction Design, an emerging approach to research facilitated by recent developments in computing. In this last chapter, I will first summarize how this approach was pursued in this work. Then I will examine the two cases of Exemplify and Nudge to support the thesis statement,

The Scientific Research through Interaction Design approach can enact preferred states in a manner that explains outcomes, informs

the conditions for applying scientific theory, and generates new experimental hypotheses.

Finally I will reflect more generally upon the design processes I used and invented in service of these goals.

6.2 Process Overview

The phases of the process roughly fit the mold of the Integrative Learning Design Framework: Informed Exploration, Enactment, Evaluation for Local Impact, and Evaluation for Broad Impact (Bannan-Ritland, 2003). Like the ILD framework, this work was also driven by the question, “How should we systematically create, test, and disseminate teaching and learning interventions that will have maximum impact on practice and will contribute significantly to theory?” In this work, the Informed Exploration was preceded by Planning of methods. I adopted HCI user experience design methods and a frame of Research through Design (J. Zimmerman et al., 2007). Further, through considering the affordances of current technology and the power of the available methods, I set out to design a particular kind of intervention that could impact practice and maintain a live connection to theory: the operant probe. By setting the operant probe as the designed artifact, I could cleanly separate the concerns of design and science to provide productive interfaces between them.

The goal of developing an operant probe shaped the Informed Exploration phase of the work. Because an operant probe is intended to operate in natural use and contribute to science, this required a map of opportunities for which designs would both be accepted and facilitate the rigorous manipulation and instrumentation of scientifically interesting variables. While I used traditional HCI methodologies like Contextual Design (Beyer & Holtzblatt, 1997) and newer methods of user experience sketching such as Needs Validation (Davidoff et al., 2007), I devised a new technique, Scientific Impact Evaluation, to evaluate the users’ needs by the ability of solutions to those needs to contribute to science.

In the Enactment phase, I used traditional HCI prototyping techniques. I again augmented these with a theoretically driven Empirical Feature Rationale map for core features of the system. With this map of lab-based principles driving the design decisions, the qualitative and quantitative aspects of the Evaluation phase can help inform the mechanisms of any outcomes, conditions for applying these empirical principles in the studied context, and new experimental hypotheses around these principles.

Finally in the Evaluation phase, I tested the systems in authentic classroom settings. While this work did not have a separate Broader Impact phase, the systems were evaluated for factors contributing to their Acceptance and Scalability. In the design of operant probes, the potential for broader impact is considered from the very beginning. The operant probes were also evaluated by their Effectiveness to improve outcomes in the context and the Insight they

provided into the mechanisms of those outcomes and future applications of the principles.

6.3 Nudge

6.3.1 Motivation

Nudge was driven primarily by the observation in the Informed Exploration phase that students needed help with time management. The technique of Needs Validation demonstrated that both students and teachers felt this need. It also scored well in the Scientific Impact Evaluation, connecting to key principles for organizing instruction and studying. A more thorough literature review added evidence that time management is difficult for students, but an important factor in their success. In a longitudinal study of cumulative GPA, a regression with time management skill and SAT scores showed time management to be a better predictor of GPA four years later (Britton & Tesser, 1991). Time management is made difficult by the human susceptibility to “planning fallacy”, the tendency for people and organizations to underestimate how long they will need to complete a task, even when they consider their previous under-estimates (Kahneman & Tversky, 1979). One technique for abating the planning fallacy is to decompose the task, and this technique is more effective for tasks of greater complexity (Kruger & Evans, 2003).

The user interviews and analysis pointed to several design principles:

- Computer support for students to use their limited time most effectively
- Require no upfront action by the student in order to benefit
- Require no changes to the instructor’s curriculum or schedule
- Require little or no time from the instructor to offer in her course

6.3.2 Solution

Nudge was designed to help students by breaking the course syllabus down into actionable tasks and supporting students in monitoring their statuses at carrying out those tasks. It was implemented as a web-based application that sent email messages when tasks were coming due. In each email was an embedded form whereby students could click to update their task statuses: Skipped, Not Started, Started or Completed. They would then see their progress through the course tasks. Any tasks they had done were stricken from future emails.

6.3.3 Effectiveness

In a randomized controlled trial over a semester of an introductory chemistry class, Nudge messages led students to spend more time in their recitation sections and helped students with poor time management to earn better grades. However, there were also some potentially negative outcomes.

Among students who reported excellent time management skills, those receiving all reminder messages performed worse on exams than those sent no reminders. One possible explanation is that the Nudge messages were effective in causing students to study in the manner modeled by the set of tasks. Because the tasks modeled a middle of the road student, the better students would be less studious than they would have been otherwise. Easy solutions to this would be to email only students who need the support, or to email messages that model more studious behaviors to students who can reach those levels. Another observed negative effect is that students receiving all messages ended the semester with higher Performance Avoidance goal orientation than students receiving no messages. The performance-avoidance orientation is basically fear of failure, which can have negative effects on learning.

6.3.4 Acceptance

It appears that students would eagerly adopt such a technology if offered more broadly. Three quarters of the respondents to an end-of-term questionnaire rated “Email reminders about course work” as “Good” or “Great”, including those who had high time management skills. Students not in the study could choose how often to receive Nudge messages and 80% did not choose to stop them. This suggests that even the students whom would have performed as well without the Nudge messages perceived them as valuable.

6.3.5 Insight

The benefits of Nudge message did not require opening the messages, but they were greater for students who opened more of them. This was not merely selecting a correlation with being a better student; the opening more messages made more of a difference for poor time managing students than those who already had managed time well. The fact that opening messages had no relation with exam scores among students with good time management helps inform the limitations of applying the principles behind Nudge.

6.3.6 Scalability

Nudge required no instructor time or changes to the course. It simply required that someone type the syllabus dates into a tool. This doesn't require any domain expertise and could easily be outsourced, but instructors may be willing to do it. When the instructor in the study was asked if he'd take the time to do it himself given the results, he replied, “Yes, very much. I would say emphatically.” The costs of operating Nudge are minimal. A simple web server can handle hundreds of courses and sending 10,000 emails costs \$1 today.

6.3.7 Future Work

As an exploratory design research project, Nudge poses more questions than it answers. One overall question is whether students should have time management

scaffolded for them, when it such an important skill to develop. The ultimate goal of Nudge is not to supplant the need for time management skills but to model them for students and support them until they have developed the skills. The contextual inquiry data support this position. Many students expressed a desire to be better students and ignorance of how to do it. Instructors valued supporting students' development of these skills but did not have the needed expertise or time to spare in their curriculum. Nudge provides this scaffolding for students with minimal instructor time to set the tasks.

Future work could explore whether Nudge-style systems shape students' enduring behaviors (positively or negatively) or just help in the course with Nudge. A benefit of the Nudge system is that it enables these sorts of long-term evaluations with relatively little costs to the researcher. In a design over multiple semesters, Nudge could be provided to students in one of two introductory classes and outcomes measured in a subsequent required course. In a design over one semester, performance in other classes could be measured as an outcome of Nudge in one of students' courses. With these, we could determine whether Nudge leads students to manage their own time better or worse (or neither) and whether the outcomes are predicted by student attributes. In the future Nudge could deliver, messages appropriate to each kind of student.

Towards the goal of broader impact, I of course would like to see evaluations of Nudge in more environments. The effects of Nudge may be more pronounced in school environments where more students struggle with time management. In particular, I would like to study Nudge in community colleges where more students must balance studies with work and family. I would also be interested in adapting Nudge to a K12 environment. K12 teachers walk structure students' study time very much already, but a system like Nudge could separate these time management skills from instruction and gradually fade for capable students to encourage internalizing the skills.

Before any future study, I would like to develop a more theoretically validated model of what activities students should perform for different standard class events like lectures, quizzes and exams. The tasks in this evaluation were a shallow attempt at distributing practice, but more fine-grained scaffolds may increase the value that students gain and perceive in the system. Further, it would increase the value of Nudge as a probe; to determine how students currently allocate their study time and how much adopting theoretically optimal study habits would affect their learning.

6.4 Exemplify

6.4.1 Motivation

Exemplify was motivated by the observation in the Informed Exploration phase the students needed help to study more effectively. Students expressed the need for active engagement to hold their attention and how the study techniques they use are ad-hoc. Students want more immediate and regular feedback on their

understanding, but quality feedback costs lots of instructor time that they do not have to spare. One solution to this is large banks of multiple-choice questions, but students and instructors agree these are shallow and do not assess deep understanding. In addition, they require time to create. Intelligent tutoring systems can get at deeper knowledge constructs, but require an inordinate amount of expert time to create.

Designing a system to support student feedback required comparing competing theoretically driven design factors. For example, worked examples without solving have been shown to be more efficient for learning than direct problem solving. As students approach mastery, the effect reverses and problem solving is more effective. Which would help students more in a real-world course setting? Further, which would students use more? Problem solving may be less motivating because it requires more work. In addition, explaining solutions to oneself is beneficial in both cases. Could a software interaction elicit this behavior from the students?

The user interviews and analysis pointed to several design principles (along with the generic latter three principles of Nudge):

- Scaffold effective study techniques for students that work even for students who don't know them
- Be interactive enough that students are engaged
- Help students to accurately assess what they know and don't know
- Be self-paced so that students can go quickly over what they are already confident in
- Map well to course assessments so that students know when they are prepared

6.4.2 Solution

Exemplify was designed to provide immediate and high quality feedback to students through an interactive problem solving activity. The key insight is that many instructors already produce answer keys to their exams. Exemplify lowers the costs of authoring interactive exercises by repurposing the troves of answer keys in instructors' filing cabinets and hard drives. Each page of a key is made interactive through simply drawing boxes over answer steps to mark what should be revealed in what sequence. Two variants of the system were developed, one without any prompt to solve or compare one's own work to the expert solution.

6.4.3 Effectiveness

In a randomized controlled trial of the variants, students with the solving version used the system more and performed better on learning measures. They performed especially better, about a grade letter, on delayed measures of learning.

In a non-randomized controlled comparison with business-as-usual, the solving version had similarly sized benefits over the non-interactive answer keys.

6.4.4 Acceptance

In a questionnaire given three quarters into the term, 40% rated the interactive tool as Good or Great, but 25% didn't yet know about it. To increase student awareness, the tool was linked to more prominently from Blackboard and the navigation was improved. In this last quarter of the term, students with the solving variant of Exemplify opened an average of 5 examples and the top half of users opened an average of 8. The nonsolving variant was used less over the term, averaging 2 in the last quarter of the term.

The instructor was skeptical of the effectiveness results because he'd been pitched many other technological systems that claimed to improve student scores. He said he'd like to see it work again and was eager to include it in the next semester course.

6.4.5 Insight

The exact mechanisms of the benefits are difficult to determine given the experimental design. The theory of the design of the solving variant points to the testing effect, but the self-explanation prompts and easily available solutions were also in play. It's an open question whether the system provided *better* testing than the business-as-usual non-interactive testing or simply motivated students to test themselves *more*. It's also not clear whether the nonsolving variant had poor effects because it didn't work as well, or students simply didn't like using it and thus didn't reap its benefits. Another factor is the expertise reversal effect, by which the nonsolving variant may have been helpful early on and the nonsolving variant after having studied. However, the solving variant can act as the nonsolving variant whenever the student wishes by bottoming out through the Get Help button.

Some data suggest that the benefits on the delayed measures of learning are due to more robust learning before the earlier measures of the same topics. For the solving variant, the difference in scores is more correlated with the number of examples opened before the early exam than after. For the nonsolving variant, there is almost no correlation with opens before the early exam.

The most important scientific insight of Exemplify and its evaluation is that metacognitive tutors can be effective without evaluating student work. In Exemplify, the student is responsible for evaluating his or her own work against the expert's work. This drastically simplifies the system and reduces costs of authoring and implementation. It is conceivable that it also leads to greater metacognitive development by requiring users to evaluate their work in order to advance. If so, such an interaction is unlikely to work for unmotivated learners but it does point strongly to a line of research to pursue.

6.4.6 Scalability

Exemplify scales easily because it re-uses existing content. It requires no changes to the curriculum and only one minute per page to annotate. Instructors can easily avoid spending this time by giving the work to their teaching assistants. The instructor in this study said, “In a situation like that, I would find help from the TAs for their labor. Should be easily within their talents.”

A bigger scalability issue is that the re-use of existing content depends on there being existing content. Not all instructors have troves of answer keys. Some of those who do may not wish to share them so that they can re-use exam questions. In practice, this may not much limit the adoption of Exemplify because the types of questions it is suited to are those that have multiple steps, where students must show their work, and thus are easier to produce.

6.4.7 Future Work

There are two main directions I would like to see Exemplify research pursue: optimization and explanation. Like the supersonic jet research described in Chapter 2, as an operant probe Exemplify can bifurcate for these two goals.

To optimize the outcomes, I would first like to validate the system in more courses and evaluate its broader impact. This will first require converting the answer keys from more courses into Exemplify activities, which will help inform the cost estimates of scaling up. I would also like to explore how well problems from one course can help students in other courses by testing them on similar but differently oriented problems. For these new settings, I would like to grow out from this chemistry course into new student populations (e.g. community college and high school) and domains. I am especially interested in whether the problem solving with examples extends past procedural domains and into ill-defined domains like history, business or design.

A priority for the next iteration is to improve the elicitation of self-explanation. Are they self-explaining and just not typing it in? How can the system better motivate sharing of explanations? In reflecting on the results, I have some ideas for a better navigation structure that prompts participation in a way that they may see more value in. For example, when they do not understand then prompt them to ask a question which someone will answer. When they do understand, they can browse and answer these questions. This would also help the instructor to see what students are struggling with at a conceptual level.

To help explain the outcomes, I would like to continue studying Exemplify experimentally. First, I would improve the logging system to better model what mechanisms of the activity are improving student learning. I am especially interested in capturing how they study outside the system, and would explore ways to poll students at a fine grain that remain ecologically valid. I would abandon the worked examples without solving because the Exemplify interaction degrades into a worked example when students click to Get Help. My next

randomized manipulation would be to compare students who have access to only Exemplify interactive problems with those who also have access to the classic noninteractive versions. Does the necessity of working at a computer hinder their studying? Do they spend less time but reach the same outcomes? Does Exemplify work well for everyone or only the students who elect to use it? I would also carry out this experiment at multiple sites to have a better representation of students and instructional settings.

6.5 Scientific Research through Interaction Design

6.5.1 Motivation

One of the top challenges of the learning sciences is in improving education as it is practiced. Educators on the front lines perceive little value in the outputs of education research. Traditional experimental research methods, in isolating variables, often lose fidelity to learning as it actually occurs. Leaders call for more “usable knowledge” (Lagemann, 2002).

The “design-based research” movement in education research attempted to place research in the learning context to improve its ecological validity. This has been at the expense of other forms of validity that science requires. Design-based research as commonly practiced has significant challenges in reproducing studies, controlling variables, and managing vast data that may be relevant.

The challenges are implicit in the tension between the design and empirical communities, in their methods, goals and reward structures. What’s needed is a better way to link research and design (Schoenfeld, 2009), and move research more rapidly into practice.

6.5.2 Solution

Scientific Research through Interaction Design offers a new way to interface science and design to produce systems that have positive real world impact. The methods and values of Interaction Design are maintained without compromising them to a “science of design”. Instead, scientists are treated as stakeholders in the familiar design processes, such that the preferred state for which they are designing is both to improve world and to place scientific instrumentation within natural contexts. I offer a name for this type of artifact, an operant probe.

6.5.3 Operant probe

Operant probes are a research apparatus that can advance learning sciences by linking the design and traditional research communities. My work is not the first operant probe, but I believe reifying this concept and producing more instances will improve both research and practice. I have offered a definition: an *in vivo* research apparatus that operationalizes theoretical constructs and collects data by which to both evaluate its effects and model the mechanisms.

In vivo experimentation is growing in education (Koedinger, Alevan, Roll, & Baker, 2009) and even iterated design of in vivo experimental interventions (E. Walker, 2010). Operant probes are not a new type of experimentation but a new emphasis on the research apparatus as a designed artifact. While in vivo experimentation helps create “usable knowledge”, the operant probe is a means for researchers to create *usable artifacts*. These systems can operate in real world settings and put the products of research directly into practice.

6.5.4 Opportunity mapping

Researchers often begin the design process with an opportunity in mind. They use HCI techniques for user-centered design like iterative prototyping, but they don’t question their framing of the problem. The success of an operant probe design depends on its adoption. In this work, I used a broader user experience design approach to discover opportunities for systems that users would likely accept, that would likely work, and that could contribute to science. I argue that designs using this method are more likely to be adopted in real world settings in ways that are sustainable, ecologically valid, and productive for research.

6.5.5 Scientific impact evaluation

An important contribution to the opportunity mapping process is the Scientific Impact Evaluation technique. I used this to prioritize among the needs that users felt for the ones for which designed solutions would 1) be predicted by lab-based principles to work, 2) inform future applications of those principles, and 3) fit the expertise of the research team in order to succeed scientifically.

This technique of evaluating scientific impact in the design process stands in contrast to normal design practice. Through this process, I was able to filter out systems that would be easy to design but not contribute to research. For example, the asynchronous question-asked backchannel in lectures. The need to ask questions in lecture without risking embarrassment was strongly felt, but the scientific opportunity for me as a researcher was not strong. This would be an excellent system for someone to implement commercially, but probably not as an operant probe.

This technique of filtering scientific principles by user acceptance also stands in contrast to the traditional pipeline of lab to practice. Many learning principles, such as spaced practice, are scientifically robust and have the potential to improve real world education, but are hindered by user acceptance. For example, I was eager to implement a system to take spaced practice to a new scope. Students and faculty expressed frustration that students forget so much of what they learned when they walk out of the final. In interviews I described OlderCheck, a system that quizzes people months after they’ve finished a course, to help them retain that knowledge. This would have been interesting scientifically, but students and teachers rejected it completely. It didn’t fit at all into how courses operate today. The Scientific Research through Interaction Design approach can

be seen as a way to focus the scientific inquiry towards knowledge that could fit more easily into real world use.

6.5.6 Evaluation of the design process

As part of the exploration of these methods, I used them to develop Nudge and Exemplify. Reflecting on the design and results of those two systems, how effective was my Scientific Research through Interaction Design approach? I discuss each of Walker's criteria for productive design research (D. Walker, 2006).

Riskier designs Both Nudge and Exemplify involved considerable risk. For one, they are new types of systems, not iterations upon or features added to existing systems. There is no prior art to automated task polling in education. Nudge did draw on designs of general productivity task management systems, but the hard-coded set of tasks may not have turned out well. (Indeed, it is not clear that that part did.) It was somewhat surprising that students did fill out the tasks and that 80% of students who had a choice kept receiving the emails. I would not have invested the time to build the system if not for the promising results from the earlier pilot and the qualitative data from the opportunity finding process.

Exemplify bears some resemblance to intelligent tutoring systems, but takes away an essential element: intelligence. Would students still learn when they could deceive the computer? Would they use it voluntarily? They did learn and did not attempt to deceive the system, likely for just the reason that use was voluntary. This has opened up a new class of tutoring support systems.

Cycles of studies The risk of these less conventional designs was minimized through the inexpensive iterative process that focused energy on ideas most likely to be accepted. To do this required failing fast on less productive ideas. The opportunity finding process helped me as the design researcher to quickly discover that some of my most precious ideas were not acceptable to the users for whom I was designing. For example, students in the interviews expressed frustration with having to learn things that were not connected to their career goals. I sketched a system to support customized curricula and social supports for sub-groups of the class with similar career goals. In interviews, students were uninterested and faculty explained, "most of the students have no idea what they want to be". Another, a system to support retention of material past the end of the course, was found to be untenable in the current university structure.

Study the resource requirements of designs Part of the opportunity finding activity is to consider the perceived benefits *and costs* to each stakeholder. Many of the other systems that were ruled out would require more effort on the part of teachers and students. Nudge and Exemplify require very little time from the instructor and fit into their existing activities. For example, Nudge tasks can be set up while making the syllabus. Exemplify exercises can be input while

making the exam answer key. Because these authoring activities require so little expertise beyond the standard materials, they can be outsourced to students or online workers for pay or recognition. When such a scale warrants the up front costs, the authorship can be lower to zero marginal cost by algorithms that interpret the instructor's raw materials.

Compare practices The market orientation of the opportunity finding method treats existing practices as competition in the market. The new designs have to be not just better than existing options, but so much better, they warrant adoption. (Or so much cheaper.)

Nudge for many students was not so much better. In interviews, organized students explained they already have their time and task management methods such as a paper calendar or dorm room whiteboard. For students without good existing practices, Nudge helped. This is likely in part because Nudge did not require any effort on their part to configure. Should teachers take on the burden (albeit minor) of supporting students' time management? It depends on their goals and incentives.

Exemplify was better than existing options. In the study, the solving version was compared both to the nonsolving variant and to the business-as-usual bank of noninteractive exercises. The solving variant was so much better that students' rate of use went up over the semester. Further, as software Exemplify can be monitored and improved over time. In designing Exemplify, it was also positioned against simple online testing systems with automated scoring and with sophisticated intelligent tutors. While there is no direct evidence comparing them, Exemplify activities are compatible with work that can't automatically scored and they are significantly cheaper to author than quality automatic scoring questions or intelligent tutors. Whether they cause better learning is an open question.

Consider sustainability and robustness The two systems have been shown to work in a classroom with negligible experimenter participation. They have not yet been shown to work in any other classroom or hostile deployment. However, because they are designed as operant probes, they are easy to replicate, iterate and monitor in new settings. Monitoring can detect early when the system usage is somehow going off the rails. Further, the qualitative research in the opportunity finding give confidence that the systems were designed with a decent understanding of the realities of the college course environment. Moreover, failures provide opportunities to explore and expand the applied knowledge of how to operationalize the basic theories.

Involve stakeholders in judging the quality of designs Nudge and Exemplify were each assessed by questionnaires with stakeholders and scored well. I believe the operant probe orientation of the work led to these systems that are easier for stakeholders to wrap their heads around to evaluate. They

are not hypothetical or contingent upon other changes. They work as is in the classrooms of today. Further, because they were designed to be domain general they are easy for stakeholders to imaginatively assess their transfer into other courses with other structures and curricula.

6.5.7 Future Work

As I described in the Nudge and Exemplify sections above, I would like to try them both in new settings such as community colleges and observe contrasts in use and perceptions. I also would like to further explore the costs of content production and specific interaction features.

For the broader design process research, I would like to apply these concepts again to new contexts. In this study I worked within the practical constraints of completing a dissertation, restricting the design space *a priori* to systems that could be informed, designed, implemented, and studied experimentally *in vivo* over a full semester all primarily by a single graduate student. How do these methods work in a team? Over several years or several months? I would like to see whether they reliably lead to productive operant probes. Moreover, I hope others will experiment with these concepts to assess whether they add value to their own design work.

However, the future work I am interested in is validation of these systems as boundary artifacts. Strong evidence would be an independent party taking up Nudge or Exemplify and either running with it, to hack away and make it as fast as possible, or walking with it to model how exactly it is working. Even more inspiring would be for the results of either of those inquiries to feed back across the boundary.

6.6 Final Thoughts

In this dissertation, I have described my work in innovating design concepts and processes for education research that better puts theory into practice. I have also described the two fruits of this labor, Nudge and Exemplify, which have been shown through *in vivo* randomized controlled trials to have benefits to learners. Nudge especially helped students with poor time management to perform better on exams. Exemplify (with solving) helped students across the board. Students who merely had access performed better than students who did not. The benefits were most pronounced on delayed measures in which students with Exemplify performed a letter grade better.

These systems operationalize theory and put it into practice. Where does this fit in the future of education research? Are designers and technologists in the learning sciences tent or will operant probes serve to delineate its boundaries? My hope is that as an applied science in a terribly complex system, developing products and shepherding them to adoption will be a valued research contribution.

7 References

- Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*(7), 861–876. doi:10.1002/acp.1391
- Alberto, P. A., & Troutman, A. C. (2008). *Applied Behavior Analysis for Teachers* (8th ed.). Prentice Hall.
- Aleahmad, T., Alevin, V., & Kraut, R. (2009). Creating a Corpus of Targeted Learning Resources with a Web-Based Open Authoring Tool. *IEEE Transactions on Learning Technologies, 2*(1), 3–9. doi:10.1109/TLT.2009.8
- Alevin, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive science, 26*(2), 147–179. doi:10.1207/s15516709cog2602_1
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences, 4*(2), 167–207. doi:10.1207/s15327809jls0402_2
- Ariely, D., & Wertenbroch, K. (2002). Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological Science*.
- Assessing our own competence: Heuristics and illusions. (1999). Assessing our own competence: Heuristics and illusions. In *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application, Attention application and performance*. The MIT Press.
- Assessment for learning: putting it into practice*. (2003). *Assessment for learning: putting it into practice*. Open University Press.
- 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. doi:10.1037/0022-0663.95.4.774
- Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research, 61*(2), 213–238. doi:10.3102/00346543061002213
- Bannan-Ritland, B. (2003). The Role of Design in Research: The Integrative Learning Design Framework. *Educational Researcher, 32*(1), 21–24. doi:10.3102/0013189X032001021
- Bartels, J. M., & Magun-Jackson, S. (2009). Approach–avoidance motivation and metacognitive self-regulation: The role of need for achievement and fear of failure. *Learning and Individual Differences, 19*(4), 459–463. doi:10.1016/j.lindif.2009.03.008

- Beckett, M., Borman, G., & Capizzano, J. (2009). Structuring Out-of-School Time to Improve Academic Achievement. IES Practice Guide. NCEE 2009-012. *What Works Clearinghouse*.
- Bell, P., Hoadley, C. M., & Linn, M. C. (2004). Design-based research in education. In *Internet environments for science education* (pp. 73–84). Mahwah, NJ: Lawrence Erlbaum Associates.
- Beyer, H., & Holtzblatt, K. (1997). *Contextual Design: Defining Customer-Centered Systems (Interactive Technologies)* (1st ed.). Morgan Kaufmann.
- Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in Self-Explanation and Self-Regulation Strategies: Investigating the Effects of Knowledge Acquisition Activities on Problem Solving. *Cognition and Instruction, 13*(2).
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In *Metacognition: Knowing about knowing*. (pp. 185–205). The MIT Press.
- Blandford, A. (2001). Group and individual time management tools: what you get is not what you need. *Personal and Ubiquitous Computing*. doi:10.1007/PL00000020
- Britton, B. K., & Tesser, A. (1991). Effects of time-management practices on college grades. *Journal of Educational Psychology, 83*(3), 405–410. doi:10.1037/0022-0663.83.3.405
- Brophy, J. (1987). Synthesis of Research on Strategies for Motivating Students to Learn. *Educational Leadership, 45*(2), 40–48.
- Brown, A. L. (1992). Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *The Journal Of the Learning Sciences, 2*, 141–178.
- Buie, E., Dray, S., Instone, K., Jain, J., Lindgaard, G., & Lund, A. (2010). How to bring HCI research and practice closer together. *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, 3181–3184.
- Burkhardt, H., & Schoenfeld, A. H. (2003). Improving Educational Research: Toward a More Useful, More Influential, and Better-Funded Enterprise. *Educational Researcher, 32*(9), 3–14. doi:10.3102/0013189X032009003
- Catrambone, R., & Yuasa, M. (2006). Acquisition of procedures: The effects of example elaborations and active learning exercises. (R. Catrambone & M. Yuasa, Eds.) *Learning and Instruction, 16*(2), 139–153. doi:10.1016/j.learninstruc.2006.02.002
- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 254–380.

- Chi, M. T. H. (1994). Eliciting self-explanations improves understanding. *Cognitive science*, 18(3), 439–477. doi:10.1016/0364-0213(94)90016-7
- Chi, M. T. H. (2000). Self-explaining: the dual process of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology*, (pp. 161–238). Mahwah, NJ: academic.research.microsoft.com.
- 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(2), 145–182. doi:10.1016/0364-0213(89)90002-5
- Clymer, J., & Wiliam, D. (2006, December). Improving the Way We Grade Science. *Educational Leadership*, 64(4), 36–42.
- Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design Experiments in Educational Research. *Educational Researcher*, 32(1), 9–13. doi:10.3102/0013189X032001009
- Collins, A. (1992). *Toward a design science of education* (No. 1). (E. Scanlon & T. O'Shea, Eds.)New directions in educational technology. Springer-Verlag.
- Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design Research: Theoretical and Methodological Issues. *Journal of the Learning Sciences*, 13(1), 15–42. doi:10.1207/s15327809jls1301_2
- Committee on Developments in the Science of Learning. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition* (2nd ed. p. 374). National Academies Press.
- Cross, N. (1999). Design research: A disciplined conversation. *Design Issues*, 15(2), 5–10.
- Data Points: More College Students Around the World. (2009, September 18). Data Points: More College Students Around the World. *The Chronicle of Higher Education*. Retrieved March 28, 2012, from <http://chronicle.com/article/Chart-More-College-Students/48516/>
- Davidoff, S., Lee, M., Dey, A. K., & Zimmerman, J. (2007). Rapidly Exploring Application Design Through Speed Dating. In J. Krumm, G. D. Abowd, A. Seneviratne, & T. Strang (Eds.), *UbiComp 2007: Ubiquitous Computing* (Vol. 4717, pp. 429–446). Springer Berlin / Heidelberg.
- Dede, C. (2004). If Design-Based Research is the Answer, What is the Question? A Commentary on Collins, Joseph, and Bielaczyc; diSessa and Cobb; and Fishman, Marx, Blumenthal, Krajcik, and Soloway in the JLS Special Issue on Design-Based Research. *Journal of the Learning Sciences*, 13(1), 105–114. doi:10.1207/s15327809jls1301_5
- Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, 32(1), 5–8.

- Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the Bold (and the Italicized): Effects of disfluency on educational outcomes. *Cognition*, *118*(1), 111–115. doi:10.1016/j.cognition.2010.09.012
- diSessa, A., & Cobb, P. (2004). Ontological Innovation and the Role of Theory in Design Experiments. *Journal of the Learning Sciences*, *13*(1), 77–103. doi:10.1207/s15327809jls1301_4
- Druin, A. (2002). The role of children in the design of new technology. *Behaviour and Information Technology*, *21*, 1–25. doi:10.1080/01449290110108659
- Ebbinghaus, H. (1913). *Memory: a contribution to experimental psychology*. Teachers College, Columbia University.
- Elliot, A. J., & McGregor, H. A. (2001). A 2 X 2 Achievement Goal Framework. *Journal of Personality and Social Psychology*, *80*(3), 501–519.
- Ellis, A., & Knaus, W. J. (1979). *Overcoming procrastination*. How to think and act rationally in spite of life's inevitable hassles. New American Library.
- Falchikov, N., & Boud, D. (1989). Student Self-Assessment in Higher Education: A Meta-Analysis. *Review of Educational Research*, *59*(4), 395–430. doi:10.3102/00346543059004395
- Falchikov, N., & Goldfinch, J. (2000). Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks. *Review of Educational Research*, *70*(3), 287–322. doi:10.3102/00346543070003287
- Fallman, D. (2004). Design-oriented Research versus Research-oriented Design (pp. 1–3). Presented at the Conference on Human Factors in Computing Systems.
- Fishman, B. J., Marx, R. W., Blumenfeld, P., Krajcik, J., & Soloway, E. (2004). Creating a Framework for Research on Systemic Technology Innovations. *Journal of the Learning Sciences*, *13*(1), 43–76. doi:10.1207/s15327809jls1301_3
- Forsyth, D., & Burt, C. (2008). Allocating time to future tasks: The effect of task segmentation on planning fallacy bias. *Memory and Cognition*, *36*(4), 791–798. doi:10.3758/MC.36.4.791
- Frayling, C. (1993). *Research in Art and Design* (No. RCS-RP--1). *opengrey.eu* (Vol. 1, pp. 1–5). London : Royal College of Art.
- Gerjets, P., & Scheiter, K. (2006). Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations? *Learning and Instruction*.
- Gerjets, P., Scheiter, K., & Catrambone, R. (2004). Designing Instructional Examples to Reduce Intrinsic Cognitive Load: Molar versus Modular Presentation of Solution Procedures. *Instructional Science*, *32*(1), 33–58. doi:10.1023/B:TRUC.0000021809.10236.71

- Goals and Goal Orientations. (2008). Goals and Goal Orientations. In *Motivation in Education: Theory, Research, and Applications*. Pearson/Merrill Prentice Hall.
- Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using Student Achievement Data to Support Instructional Decision Making* (pp. 1–76). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Hausmann, R. G. M., & VanLehn, K. (2007). Explaining Self-Explaining: A Contrast between Content and Generation. In *Proceeding of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work* (pp. 417–424). Amsterdam, The Netherlands: IOS Press.
- Hausmann, R. G. M., Van De Sande, B., & VanLehn, K. (2008). Are Self-explaining and Coached Problem Solving More Effective When Done by Pairs of Students Than Alone? *Arxiv preprint arXiv:0805.4223*.
- Howley, I. K. (2010). *Designing for the Whole Student*. Retrieved from http://www.andrew.cmu.edu/user/ihowley/website/papers/ihowley_DesigningForTheWholeStudent.pdf
- Kahneman, D., & Tversky, A. (1979). Intuitive prediction: Biases and corrective procedures. *TIMS Studies in Management Science*, *12*, 313–327.
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychology*, *38*(1), 23–31.
- Karpicke, J. D. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language*, *62*(3), 227–339. doi:10.1016/j.jml.2009.11.010
- Karpicke, J. D., & Blunt, J. R. (2011). Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. *Science*, *331*(6018), 772–775. doi:10.1126/science.1199327
- Klein, J., & Rice, C. (Eds.). (2012). *U.S. Education Reform and National Security* (No. 68). *cfr.org*. Council on Foreign Relations Press.
- Koedinger, K. R., Alevan, V., Roll, I., & Baker, R. (2009). In vivo experiments on whether supporting metacognition in intelligent tutoring systems yields robust learning. *Handbook of metacognition in education*, 897–964.
- Koedinger, K. R., Corbett, A. T., & Perfetti, C. (in press). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*.
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370.
- Kornell, N. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*(5), 493–501. doi:10.1080/09658210902832915

- Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224.
- Kruger, J., & Evans, M. (2003). If you don't want to be late, enumerate: Unpacking reduces the planning fallacy. *Journal of Experimental Social Psychology*, *40*(5), 586–598. doi:10.1016/j.jesp.2003.11.001
- Lagemann, E. C. (2002). *An Elusive Science: The Troubling History of Education Research*. University of Chicago Press.
- Lee, B. G. (2006). *Lecture first or text first? Optimizing undergraduate instruction* (Vol. 66). University of California, Los Angeles.
- Liem, A. D., Lau, S., & Nie, Y. (2008). The role of self-efficacy, task value, and achievement goals in predicting learning strategies, task disengagement, peer relationship, and achievement outcome. *Contemporary Educational Psychology*, *33*(4), 486–512. doi:10.1016/j.cedpsych.2007.08.001
- Linn, M. C., Davis, E., & Eylon, B. (2004). The Scaffolded Knowledge Integration Framework for Instruction. In *Internet environments for science education*. Lawrence Erlbaum Associates.
- Lovett, M. (1992). Learning by problem solving versus by examples: The benefits of generating and receiving information. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 956–961). Hillsdale, NJ: Erlbaum.
- McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*(2), 200–206. doi:10.3758/BF03194052
- McLaren, B., Lim, S.-J., & Koedinger, K. R. (2008). When and how often should worked examples be given to students? New results and a summary of the current state of research. Presented at the 31st Annual Conference of the Cognitive Science Society, Austin, TX.
- Metcalf, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review*, *14*(2), 225–229.
- Mor, Y., & Winters, N. (2007). Design approaches in technology-enhanced learning. *Interactive Learning Environments*, *15*(1). doi:10.1080/10494820601044236
- Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education (IJAIED)*, *10*, 98–129.
- Nathan, M. J., Koedinger, K. R., & Alibali, M. W. (2001). Expert blind spot: When content knowledge eclipses pedagogical content knowledge. In *Proceedings of the Annual Meeting of the American Educational Research Association*. Seattle.
- No Child Left Behind Act., gpo.gov (2002). 115.

O'Toole, G. (2011, July 28). My Customers Would Have Asked For a Faster Horse

. *quoteinvestigator.com*. Retrieved March 30, 2012, from <http://quoteinvestigator.com/2011/07/28/ford-faster-horse/>

Paas, F. G. W. C. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*.

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K. R., McDaniel, M. A., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning*. IES practice guide.

Phillips, D. C. (2006). Assessing the quality of design research proposals: Some philosophical perspectives. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 144–155). Routledge.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into Practice*.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachle, W. J. (2001). *A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. (No. 91-B-004). *eric.ed.gov* (University of Michigan.). University of Michigan.

Renkl, A. (1997). Learning from Worked-Out Examples: A Study on Individual Differences. *Cognitive science*, *21*(1), 1–29. doi:10.1207/s15516709cog2101_1

Renkl, A. (2002). Worked-out examples: instructional explanations support learning by self-explanations. *Learning and Instruction*, *12*(5), 529–556. doi:10.1016/S0959-4752(01)00030-5

Renkl, A., Atkinson, R. K., & Maier, U. H. (2000). From studying examples to solving problems: Fading worked-out solution steps helps learning. . . . *of the 22nd Annual Conference of the*

Richards, C. S. (1975). Behavior modification of studying through study skills advice and self-control procedures. *Journal of Counseling Psychology*, *22*(5), 431–436.

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, *4*(2), 155–169. doi:10.1007/BF01405730

Roediger, H. L., III, & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*.

Roediger, H. L., III, & Karpicke, J. D. (2006b). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, *17*(3), 249–255. doi:10.1111/j.1467-9280.2006.01693.x

- Roschelle, J., & Penuel, W. R. (2006). Co-design of innovations with teachers: definition and dynamics. In Proceedings of the 7th international conference on Learning sciences (pp. 606–612). Presented at the ICLS '07, Bloomington, Indiana: International Society of the Learning Sciences.
- Salden, R. J. C. M., Alevan, V., Renkl, A., & Schwonke, R. (2009). Worked Examples and Tutored Problem Solving: Redundant or Synergistic Forms of Support?
- Salden, R. J. C. M., Koedinger, K. R., Renkl, A., Alevan, V., & McLaren, B. (2010). Accounting for Beneficial Effects of Worked Examples in Tutored Problem Solving. *Educational Psychology Review*, 22(4), 379–392.
- Scheiter, K., Gerjets, P., & Vollmann, B. (2006). A methodological alternative to media comparison studies: Linking information utilization strategies and instructional approach in hypermedia learning.
- Schoenfeld, A. H. (2006). What Doesn't Work: The Challenge and Failure of the What Works Clearinghouse to Conduct Meaningful Reviews of Studies of Mathematics Curricula. *Educational Researcher*, 35(2), 13–21. doi:10.3102/0013189X035002013
- Schoenfeld, A. H. (2009). Bridging the Cultures of Educational Research and Design. *Journal of the International Society for Design and Development in Education*, 1(2).
- Schworm, S., & Renkl, A. (2002). Learning by solved example problems: Instructional explanations reduce self-explanation activity. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 816–821). Erlbaum.
- Semb, G., Glick, D. M., & Spencer, R. (1979). Student Withdrawals and Delayed Work Patterns In Self-Paced Psychology Courses. *Teaching of Psychology*, 6(1), 23–25. doi:10.1207/s15328023top0601_8
- Shepard, L. A., & (null). (2000). The Role of Assessment in a Learning Culture. *Educational Researcher*, 29(7), 4–14. doi:10.3102/0013189X029007004
- Simon, H. A. (1969). *The Sciences of the Artificial*. Cambridge: MIT Press.
- Sketching User Experiences: Getting the Design Right and the Right Design*. (2007). *Sketching User Experiences: Getting the Design Right and the Right Design*. Morgan Kaufmann Publishers Inc.
- Slotta, J. D., & Aleahmad, T. (2009). Toward a technology community in the learning sciences (pp. 12–14). Presented at the 9th international conference on Computer Supported Cooperative Learning, International Society of the Learning Sciences.
- Snyder, T. D., & Dillow, S. A. (2011). Postsecondary Education. In *Digest of Education Statistics, 2010*. National Center for Education Statistics.

- Solomon, L. J., & Rothblum, E. D. (1984). Academic procrastination: Frequency and cognitive-behavioral correlates. *Journal of Counseling Psychology, 31*(4), 503–509.
- Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30*(9), 641–656. doi:10.1037/h0063404
- Stipek, D. J., & Kowalski, P. S. (1989). Learned helplessness in task-orienting versus performance-orienting testing conditions. *Journal of Educational Psychology, 81*(3), 384–391. doi:10.1037/0022-0663.81.3.384
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science, 12*(2), 257–285. doi:10.1016/0364-0213(88)90023-7
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. Explorations in the Learning Sciences, Instructional Systems and Performance Technologies (Vol. 1, p. 274). Springer.
- Sweller, J., van Merriënboer, J., & Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review, 10*(3), 251–296.
- The Role of School Improvement in Economic Development. (2007). The Role of School Improvement in Economic Development.
- Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 1024–1037. doi:10.1037/0278-7393.25.4.1024
- United States National Commission on Excellence in Education. (1983). *A Nation At Risk: The Imperative For Educational Reform*. Superintendent of Documents, U.S. Gov. Print. Off.
- van den Akker, J., Gravemeijer, K., & Nieveen, S. M. N. (2006a). Introduction to educational design research. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research*. Routledge.
- van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (Eds.). (2006b). *Educational design research*. Routledge.
- van Gog, T., Paas, F. G. W. C., & van Merriënboer, J. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. *Learning and Instruction, 16*(2), 154–164. doi:10.1016/j.learninstruc.2006.02.003
- van Merriënboer, J. J. G. (1997). *Training complex cognitive skills: a Four-Component Instructional Design model for technical training*. Educational Technology Publications.
- VanLehn, K., Jones, R., & Chi, M. T. H. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences, 2*(1), 1–59.

- Viadero, D. (2009, April 1). “No Effects” Studies Raising Eyebrows. *Education Week*, 28(27), 1,14–15.
- Walker, D. (2006). Toward productive design studies. In J. van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research* (pp. 9–18). Routledge.
- Walker, E. (2010). *Automated Adaptive Support for Peer Tutoring*. (K. R. Koedinger & N. Rummel, Eds.) (pp. 1–193).
- Wang, F., & Hannafin, M. J. (2005). Design-Based Research and Technology-Enhanced Learning Environments. *Educational Technology Research and Development*, 53(4), 5–23.
- Whitehurst, G. J. (2003, April 22). The Institute of Education Sciences: New Wine, New Bottles, a Presentation by IES Director Grover (Russ) Whitehurst. *ies.ed.gov*.
- Wittwer, J., & Renkl, A. (2010). How Effective are Instructional Explanations in Example-Based Learning? A Meta-Analytic Review. *Educational Psychology Review*, 22(4), 393–409. doi:10.1007/s10648-010-9136-5
- Wylie, R., Koedinger, K. R., & Mitamura, T. (2009). Is Self-Explanation Always Better? The Effects of Adding Self-Explanation Prompts to an English Grammar Tutor. In Proceedings of the 31st Annual Conference of the Cognitive Science Society (pp. 1300–1305).
- Zimmerman, B. J., Moylan, A., & Hudesman, J. (2011). Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling*, 53(1), 141–160.
- Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). Research through design as a method for interaction design research in HCI. Proceedings of the SIGCHI conference on Human factors in computing systems, 493–502. doi:10.1145/1240624.1240704

8 Appendix A: Output of ideation

| # | Idea |
|---|--|
| 1 | Class study partner pairing system. |
| 2 | Shared note-taking wiki style that all notes coexist but some are prominent, social voting |
| 3 | TA review session voting system (submit questions, everyone votes and popular ones first) |
| 4 | Share the question analytics with students. |
| 5 | Funnel most disputed questions to students. |
| 6 | Intelligent system to prioritize materials during study time. |

| # | Idea |
|----|--|
| 7 | Stats on a question during practice to show how hard it is. (IRT curve) |
| 8 | Seat reassignment system to pair clicker discussion partners. (learning community) |
| 9 | Study time companion, find most appropriate way to study within a certain time window. (e.g. bus ride) |
| 10 | Book edition referent translator. (page numbers between editions) |
| 11 | Public anonymous note-taking on the learning goals to read the same material in different ways. |
| 12 | Study partner match-up system based on performance data. |
| 13 | Easy attendance system. (sort of CAPTCHA) |
| 14 | Quiz system that gathers up learning components you need more help on for review before the exam. |
| 15 | Games/puzzles in lecture to keep everyone engaged. |
| 16 | Voluntary delayed post test system for data mining. |
| 17 | Real time feedback to the instructor whether people understand what you're saying (e.g. slides) or doing (e.g. activities) |
| 18 | Mark your confusion at a part of the lecture for someone to help you. (audio recording) |
| 19 | Contributions that don't require being right or wrong. (e.g. cog psych scenarios, provocative questions) |
| 20 | Wiki study packets for exams. |
| 21 | Make the grade reflect real learning, not motivation. |
| 22 | Self-testing system during exam prep, coupled with wiki instructional materials and worked examples. |
| 23 | Analytics on how much students are working and how. (Anonymous logging and reporting.) |
| 24 | Personal informatics on how you're spending attention. Cognitive/goal costs of compulsive computing. |
| 25 | Writing tutor that teaches the domain (to enable and carry out better assessments) |
| 26 | Assess learning through authoring scenario-based applied questions |
| 27 | Studying informatics to be coached by instructor, peers, computer, etc. |
| 28 | Teacher-accessible question and test informatics/validation tool (product-y) |
| 29 | Integrated exam grading system (Questionmark?) |
| 30 | Rapid system showing students their retention with active learning vs. passive |
| 31 | Study behaviors tutor, tied to real data from learning activities and outcomes. |
| 32 | Make students care, show they care. |
| 33 | Big ideas database to find concepts that cut across findings |

| # | Idea |
|----|--|
| 34 | Automated attendance system in lecture. (geoweb? exchange info with person sitting next to you to validate you're there) |
| 35 | Real-time monitor of student engagement in the class. Let them indicate with devices, or sense with camera in the front (counting eyeballs with IR) |
| 36 | More active processing of lecture (than note-taking) |
| 37 | Peer tutoring support highly instrumented for accountability / class credit |
| 38 | Phone-based quiz answering to avoid paperwork time. (M/C and short answer) |
| 39 | Actionable analytics for teachers on their formative assessments. |
| 40 | Shared course planning, management tools for a team of teachers. A la Lesson Study. |
| 41 | Explain the lecture at the end of the lecture. By calling into a system? Peer review of audio later. Compare with rubric and give written feedback. |
| 42 | Data mining class participation to intervene in different ways, deliver personalized learning. |
| 43 | Graphical tools to easily connect current topic with earlier topics (in lectures?) Activeclass.org |
| 44 | Micro-experiment tracking system for educators. Quick pre/post assessments around a small treatment. |
| 45 | Versioning system for teaching materials with in-class annotations on each version. |
| 46 | Compare learning across labs/recitations to monitor teaching of TAs, help them and improve the course. |
| 47 | Portfolio system to develop the macro concepts of the course. Track recurrence of ideas. |
| 48 | Phone app to display your understanding stats throughout the semester. (Gamelike) |
| 49 | Personalized questions during lecture. Pick your difficulty level. (Phone affords individual display > individual response) |
| 50 | Backchannel with TA moderation to interrupt, queue or ignore on incoming question. |
| 51 | Recording questions keyed to time code and point in the slides. Embed student experience in materials (for future self, and others) |
| 52 | Ideate on fragmenting the attention streams in large lectures so more students can participate / be immediately accountable. |
| 53 | Anonymous polling system to know what students are really doing. (did you read the chapter? have you prepared for class? did you cheat on the homework?) |
| 54 | Confidential assessments, surveys (no name encourages honesty, but still get feedback) |

| # | Idea |
|----|---|
| 55 | Pre-req ramp-up to get students in sync (learn immediately where you are in class, what you need to get by) |
| 56 | Assessment system that is low/no anxiety. |
| 57 | Crowd-source the content of a learning game (authoring/use class/library/bus/home) |
| 58 | Process sharing system. Instructor proposes ways to do things, students reflect and improve on the process. |
| 59 | Teaching repository annotated with student feedback |
| 60 | Daily evals on teaching and instructions |
| 61 | Course question system: email an address, makes a ticket, instructors can make public (AAQ) |
| 62 | Facilitating spontaneous small-scale discussions |
| 63 | Just in time tutoring through micro-labor markets |
| 64 | Algorithm to distribute students throughout the hall to for break-out with different groups |

9 Appendix B: Early design areas for lecture courses

Support mastery learning. Students not seeing results, focused on percentages instead of usable knowledge. Students want faster feedback on what they know and don't know. Detailed feedback can help them master what they want to master for themselves vs. the grade.

Personalizing lecture. Students have different life goals wrt/ the course content. Adapt to different backgrounds and interests.

Supporting explication of learning goals. Related to mastery learning. Student want to focus on exactly what they are expected to learn. Instructors could use more accountability and discourse around harder to assess goals.

Developing time and attention discipline. Students feel bad about these and it leads to low retention. Technology makes them distracted more easily than ever. Compensate for this.

Integrate content into their lives. Students unmotivated by content they see as irrelevant to their lives. Link content into their daily experiences with mobile technology. Incentivize reading through questioning. (like the OLI Stats tutor) Spaced practice.

Instrumenting reëncoding for self and formative assessment. Students spend a lot of time rewriting notes and aren't sure of what they understand or not. Mediate this informatically to inform both them and instructors of what they know and don't.

Community of learners. Lack of motivation, poor study habit. Drive participation with communal activities.

Encouragement through non-grade feedback and achievements. Freshmen have fragile egos. Course grade is a “course” evaluation, based on somewhat arbitrary relative weights. Recognize smaller achievements and personal goals.

Social support for achievements. Students live in social media. Public goal setting with accountability.

Taking advantage of the large size. What can you do in large lectures that you can’t in smaller ones? Harness diversity, anonymity, division of labor.

Motivating interest, nourishing curiosity. Students are unmotivated and incurious. Motivate them as life long learners.

Add fun to the work. Most students do the work because they have to. Freshman courses are intro or breadth and have many people who aren’t passionate about the material. To increase their learning, add game-like engagement. e.g. achievements that aren’t graded or opportunities for creativity.

Increase conceptual learning. Teachers are inclined to use shallow resources by their availability in textbooks. Students value conceptual learning more than facts but it takes more effort which they want to minimize. Teachers give shallow questions to meet student effort, lowering their motivation. Bad cycle. E.g. mechanisms for augmenting shallow resources with deeper reasoning.

Feedback to teachers. Teachers stare out into a “sea of faces”. Hard to know distribution of interest, arousal, learning in the students. Rapid, rich data for personal and pedagogical formative evaluation.

10 Appendix C: Scenarios developed for Needs Validation

| | Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS prin- ci- ples | LS con- tribu- tion po- ten- tial |
|---|-------------------------|--|---|-------------------------------|---|----------------------------|---|
| 1 | progress | To see more immediate fruits of study efforts. Faster feedback on what they know and don't know. | Do you want to know how you've progressed after a study session? | Motivate students to study. | Do your students underestimate how much studying helps them? | OIS 6a, OIS 6b | Motivation. |
| | | | | | | | <p><i>Peter has an exam on Tuesday and sits down in the library Sunday afternoon to study.</i></p> <p><i>Before cracking the book, he takes a short online quiz on his laptop. The system gives him feedback on what he knows and doesn't know.</i></p> <p><i>While reading through his book and notes, he pays more attention to what the system said needs help in.</i></p> <p><i>A few hours later he opens his laptop again and takes another quiz. It shows him he did better on all but one part. He is satisfied with his progress and leaves the library.</i></p> |
| 2 | retention | Hard to retain information. | Do you find it hard to remember what you learned earlier in the semester? | Higher retention of material. | Do you find that your students forget what they've "learned" earlier? | OIS 6a, OIS 1 | |

| Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS prin- ci- ples | LS con- tribu- tion po- ten- tial |
|-------------------------|---|--------------------------------------|---|--|--|---|
| | <p><i>Peter crammed and did poorly on the first exam. He feels like he knew it at some point but forgot. To help he goes online to enroll in the optional RetentionBuddy system Prof. Treakle has set up for the class.</i></p> <p><i>The next morning checking his email at the dinner table, he sees a message with questions about some stuff covered last week and two weeks ago in class. He sort of remembers the stuff from last week and is frustrated he doesn't remember the stuff from two weeks ago, but he thinks through it and comes up with the answer. He replies to the email with his answers.</i></p> <p><i>A week later he gets an email with questions from one, two and three weeks go. He is glad that he can answer the stuff from three weeks ago that he had trouble with last time.</i></p> | | | | | |
| 3 | relevance | Connecting through shared interests. | Do you ever question what the value is of what you're learning? | Connect learning with shared inter-ests. | Are your students able to connect with people who share their interests? | SOST 4, OIS 4 Motivation. |

| Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS prin- ci- ples | LS con- tribu- tion po- ten- tial |
|-------------------------|--|--|---|--|--|---|
| | <i>Claire is in her dorm room working on the first homework assignment in her Intro to Psychology class. She has to describe her career plans and imagine how an understanding of psychology could help her.</i> | | | | | |
| | <i>She writes what she wants to be a nurse traveling around the world. Her friend Lisa wants to be a teacher.</i> | | | | | |
| | <i>The next week in lecture, there is a slide listing subgroups in the class based on people's answers to the survey. Prof. Forbes instructs the groups to find each other in the lecture and meet up. Throughout the semester they will have customized assignments related to their goals.</i> | | | | | |
| | <i>Over the weekend she talks with her friend Lisa about their homework on Attachment Theory. Claire's group had to relate it to baby incubators in hospitals while Lisa had to relate it to kindergarten teachers.</i> | | | | | |
| 4 | allocation | Spend study atten- tion on what you don't know. | Do you have any trouble prioritiz- ing what you should be reviewing in your notes? Ever studied the wrong stuff? | Help stu- dents focus on the impor- tant parts. | Do you wish you had data on what students think they know? | OIS 6b |

| Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS prin- ci- ples | LS con- tribu- tion po- ten- tial |
|-------------------------|--|---------------------------------|---|--|--|---|
| | <p><i>Claire got her Intro to Business midterm back and was disappointed that she missed some key areas. She was pissed because she spent so much time on another area that was barely on the exam.</i></p> <p><i>Before the next quiz she goes to the library and tries the DoYouKnow system on her laptop for help on what she should focus on. First it lists the areas she is expected to know and then she has to rate her confidence in each area.</i></p> <p><i>She answers a series of quiz-like questions that adapt to be harder or easier based on her answers. When she's done, she gets a report on how well the system thinks she knows the different areas. She is surprised that the system recommended she study more on supply/demand and market growth because she thought she understood those.</i></p> <p><i>She starts studying those trouble areas and worries less about the others.</i></p> | | | | | |
| 5 | competition | Competitive motiva- tion. | Do you want to be the best at something in class, even if it's not the best exam grade? | Motivation through your com- peti- tion. | Would your students benefit from some friendly competi- tion? | OIS 6a, 6b Metacognition. |

| Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS prin- ci- ples | LS con- tribu- tion po- ten- tial |
|-------------------------|---|---|--|---|----------------------------|---|
| | <p><i>Claire sits down at her kitchen table and logs into PeerQuiz to do her homework. She has to write and answer questions that require applying psychology theories to particular situations. She writes three questions.</i></p> <p><i>Next she looks over other questions. She clicks to rate them and types comments to critique them. When she finds a question she likes, she types a paragraph to answer it. Before she logs off she sees that one of her questions is now the 3rd highest rated.</i></p> <p><i>The next day she gets a text message with the current ratings of her questions and answers. She sees that her top-rated question has fallen to 5th place. She also sees that her answer is rated highly.</i></p> | | | | | |
| 6 friends | Social awareness. | Do you feel like your education is too separate from the rest of your life? | Integrating class with their social lives. | Do some of your students need more social support for their learning from outside the school? | SOST 4 | |

| Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS principles | LS contribution potential |
|-------------------|--------------|---------------------|--|----------------------|---|--|
| | | | | | | <p><i>At the first lecture in Intro to Physics, everyone has to create a list of people to share their homework and grades with using LearnShare. Peter share his exam grades and attendance with his parents, his problem sets with his friends who are interested and everything with his sister.</i></p> <p><i>After his problem set on mechanics he gets an email from LearnShare. His math whiz friend says it's cool that Peter is learning this stuff. He says that Newton's laws are only an approximation and the relativistic model is even more interesting.</i></p> <p><i>After the first exam he gets an email from his mom, "Good job, son. An A on your first exam!"</i></p> <p><i>A few weeks later he is struggling and gets a D on the quiz. His sister calls him to ask how he's doing.</i></p> |
| 7 | recognition | Achievement | Do you | Acknowledgment | Do you | SOST 4 |
| | | that aren't just %s | wish you got some recognition in class besides just a grade? | by more than grades. | wish it was easier to recognize your students' efforts? | |
| | | | | | | <p><i>Claire is taking Intro Psych to meet her science requirement. At night she goes online to write her two weekly posts to the class forum. She didn't have time to read the textbook so instead she challenges other students to say why the stuff matters. She gets into heated exchanges and posts many more than two messages per week.</i></p> <p><i>Later in class the midterm grades are posted on the projector. She sees the top names and looks down to find hers with 76%. She sees other names for Best Questions in Class, Perfect Attendance, Most Improved, and see her name again under Most Provocative Poster.</i></p> |

| | Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS prin- ci- ples | LS con- tribu- tion po- ten- tial |
|---|-------------------------|--|---|--|--|----------------------------|---|
| 8 | strategy | Learn better study strategies. | How confident are you in your study strategies? | Improving student study strategies. | Would your students benefit from feedback on their study strategies? | OIS 6, OIS 1 | Optimal timing on longer scale. |
| | | <p><i>In the first day Intro Psych, Prof. Treakle tells the class that part of their grade will be for reporting their study activities and the data will be used for in-class analysis on the psychology of learning.</i></p> <p><i>Peter doesn't study very carefully. (He just takes notes in class and skims the textbook.) The lecture starts and he records on his phone that he is taking notes in class.</i></p> <p><i>After class he reads over his notes several times and records this.</i></p> <p><i>Weeks later he receives his first exam back with a C. Along with his grade is a full-page report on the study activities of the students who did best and those who did poorly. Other students who spent as much time but got better grades quizzed themselves while reading, so he decides to do that next time.</i></p> | | | | | |
| 9 | timegoals | Develop good time management. | Do you ever feel overwhelmed or regretful of how you spend your time? | Improving students' time use planning. | Do you think your students should improve their time management? | OIS 6, OIS 1 | Optimal timing on longer scale. |

| Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS prin- ci- ples | LS con- tribu- tion po- ten- tial |
|-------------------------|--|------------------------------------|---------------------------------|--|---|---|
| | <p><i>Claire gets her grade back on the last exam and it's lower than she expected. For this next unit, she decides to study regularly.</i></p> <p><i>At home in the evening Claire uses the TimeGoal system to set goals for what she will accomplish each day to prepare for the next exam. She resolves to read or at least skim the reading before class on Wednesday. She also wants to visit the TA's office hours on Thursday.</i></p> <p><i>Tuesday night at 9pm she receives a message on her phone asking how much she read. She clicks to postpone it. Thursday a message asks whether she went to office hours. She clicks No.</i></p> <p><i>On Friday she doesn't feel confident for Monday's quiz and pulls up a report of whether she's studying more regularly like she wanted. It shows her that she skipped both the reading and the office hours. She has to cancel some weekend plans to cram for Monday's quiz.</i></p> | | | | | |
| 10 | timeawareness | Be confident in how time is spent. | Do you ever lose track of time? | Improving student's awareness of time spent. | Do you wish you had data on the time students spend on different parts of your class? | OIS 6 Optimal timing on longer scale. |

| | | | | | | LS con- tribu- tion po- ten- tial |
|-------------------------|-----------------|--|---|--|--|---|
| Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS prin- ci- ples | |
| | | | | | | <i>Thursday night, Peter realizes he has a big essay due Monday morning and he'll have to work hard on it over the weekend.</i> |
| | | | | | | <i>At 9pm on Friday night Peter is tired of reading and decides it's time to go out. He records on his phone that he's switched from Studying to Partying.</i> |
| | | | | | | <i>Saturday afternoon Peter picks up his essay again and records Studying on his phone. After an hour he records he'll check Facebook for just 10 minutes. He loses track of time until the phone beeps to remind him and he returns to studying.</i> |
| | | | | | | <i>Sunday evening Peter has completed his final draft and submits it early. With satisfaction he records Watching TV.</i> |
| 11 | interest | Motivating interest in the mate- rial. | Do you have any questions or curiosities related to the material that don't get answered by the course? | Students inter- ested in what I'm teach- ing. | Do you wish your students' learning would be driven by their own questions? | OIS 7 Motivation. |

| Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS principles | LS contribution potential |
|-------------------|--------------|--|---|--|---|--|
| | | | | | | <p><i>On the first day of his Intro Psych class, Peter has to fill out a form with the big questions he has about human psychology and his goals for the course.</i></p> <p><i>After the first week Peter is at home on his computer and gets an email reminding him of his initial interests. It asks how the week helped satisfy them and prompts him for new questions he has.</i></p> <p><i>Each week he gets another email with his past questions and answers. He adds new questions and expands earlier answers. He feels like he's learning what he cares about.</i></p> <p><i>At the end of the course he receives a full report of the big questions he had and what he learned to answer those questions. He read it and feels a sense of accomplishment.</i></p> |
| 12 | adaptive | Engagement with adaptive difficulty and topic. | Are you ever frustrated with the lecture material being too fast or slow? Hard or easy? | Matching lectures to students' levels. | Do you worry that your lectures are above or below some students? | OIS 5b Deeper questioning. |
| | | | | | | <p><i>In lecture Prof. Treacle pushes to ask the class questions. She presents many questions on a slide with different difficulty levels side by side. Peter chooses the hardest one and enters his answer on his phone.</i></p> <p><i>He gets back a message that he was right and how many other students attempted and succeeded at that question.</i></p> <p><i>Prof. Treacle can see that Patrick answered the easiest one and calls on him to explain his answer.</i></p> |

| | Scenario | Student | Student | Teacher | Teacher | LS | LS |
|----|-------------|---|------------|---------|------------|-------|-------------|
| | key | need | lead | need | lead | prin- | con- |
| | word | | | | | ci- | tribu- |
| | | | | | | ples | tion |
| | | | | | | | po- |
| | | | | | | | ten- |
| | | | | | | | tial |
| 13 | backchannel | Engaging | Do you | More | Do you | OIS | Motivation. |
| | | quiet | ever have | fre- | wish it | 7, | For- |
| | | stu- | a question | quent | were | SOST | ma- |
| | | dents, | during | feed- | easier for | 5 | tive |
| | | larger | class but | back. | students | | assess- |
| | | back | hold back? | | to ask | | ment. |
| | | channel. | | | questions | | |
| | | Im- | | | during | | |
| | | prove | | | lecture? | | |
| | | teach- | | | | | |
| | | ing. | | | | | |
| | | <i>During her Developmental Psychology lecture, Claire has a question but thinks it's too dumb to raise her hand. Instead she asks the question anonymously on her phone.</i> | | | | | |
| | | <i>Allison the TA is reading each question and ranking them privately.</i> | | | | | |
| | | <i>Every 20 minutes Prof. Treakle shows the top questions on the projector. Claire's is on the list and Prof. Treakle answers it.</i> | | | | | |
| | | <i>Later in the lecture she has another question which is off topic. She sends this one in with her name to be sure she gets an answer. In the next question block Prof. Treakle doesn't choose it to answer.</i> | | | | | |
| | | <i>That night though, she gets an email. Another student liked her question and wrote an answer. Allison the TA marked the answer as acceptable but added a few clarifications.</i> | | | | | |

| | Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS prin- ci- ples | LS con- tribu- tion po- ten- tial |
|----|-------------------------|--|---|------------------------------------|---|----------------------------|---|
| 14 | valued | Feel valued. Know what you know. Break illusion of knowing through reading. | Do you wish your classwork were useful to more than you? | Motivation by seeing value. | Do you wish your students were motivated by something other than a grade? | OIS 1, 5b, 6b, OIS 7 | Motivation. Student questioning. |
| | | <p><i>In the library one evening Peter works on his homework. He has to write questions that test the key ideas from lecture. He logs into CrowdExam and types in questions and their answers. One of them is too similar to another student's question and he changes it.</i></p> <p><i>The next day he receives an email that the TA has rated his questions and two of them are 5 stars out of 5.</i></p> <p><i>While taking the midterm, he gets to a question that is one of his. The professor picked it among the top-rated questions to include on the exam.</i></p> | | | | | |
| 15 | reception | Formative data on how students are receiving the lectures. | Do you wish your professor took feedback on each lecture? | Data on impact/activity of slides. | Do you often wonder how your lecture went? How it could improve? | | |

| Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS prin- ci- ples | LS con- tribu- tion po- ten- tial |
|-------------------------|--|---|--|---------------------|---|---|
| | <i>In her office, Prof. Treacle edits the slides she will use tomorrow in class. For one, she adds a new YouTube video to demonstrate Change Blindness.</i> | | | | | |
| | <i>In class, while she shows each slide students click on their devices with a rating of whether they understand the points. Some students text in with suggestions.</i> | | | | | |
| | <i>At a break she solicits questions. Allison the TA writes these down. Students rate their satisfaction with the answers the professor gives.</i> | | | | | |
| | <i>In her office that afternoon, Prof. Treacle looks at her slide deck with the student feedback and questions beside it. She identifies some slides that could use more work. She also sees that students were more confused by the YouTube video than last semester's were with her earlier animation.</i> | | | | | |
| 16 | longterm | Robust learning that impacts your life. | Do you ever wish you remembered more from your past classes? | Long term learning. | Are you concerned about students forgetting everything after the final? | OIS 1, OIS 5b, SOST 4 Easy measurement of robust learning. Motivation. |

| Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS principles | LS contribution potential |
|-------------------|---|---|-----------------------|--|------------------------------|--|
| | <p><i>In his Developmental Psychology class Peter has an assignment to grade the quiz of a student who finished the class last semester. He logs into OlderCheck and reads their answers. He writes the correct answers and their explanations. He does this for three different students.</i></p> <p><i>During the midterm, he is able to answer some questions by recalling his explanations on OlderCheck.</i></p> <p><i>Several months after his final, he gets an email with a quiz on the same stuff he tutored in OlderCheck. He tries answering the questions but can't remember everything. He looks forward to the explanations from the current students so that he can remember what he learned and not have it be a waste.</i></p> | | | | | |
| 17 | connections Deeper conceptual learning. | Do you feel like the stuff you learn is separate information that's not really related? Does it bother you? | Developing big ideas. | Do you wish your students would take away deeper conceptual understanding? | OIS 7, OIS 4, SOST 4, SOST 3 | Learning impact of affinity diagraming. Related to Concept Map literature. New assessment technique. |

| Scenario key word | Student need | Student lead | Teacher need | Teacher lead | LS prin- ci- ples | LS con- tribu- tion po- ten- tial |
|-------------------------|--|-----------------|-----------------|-----------------|----------------------------|---|
| | <p><i>In lecture after the midterm Prof. Treacle says the students need to see the bigger picture of psychology and introduces the Relations system. Peter types out all the most important ideas and research he can think of from the class.</i></p> | | | | | |
| | <p><i>The next week he meets with his assignment group and they lay out strips of paper on the floor with all their ideas printed on them. For the first phase, they have to organize them by experimental methodology. They debate with each other how the ideas all fit together and when they agree they snap a photo which they upload to Relations.</i></p> | | | | | |
| | <p><i>In the next class Prof. Treacle shows a slide with a giant map of all the concepts that everyone put together. He shows what relationships most students agreed on and what ones were controversial. The class has a big discussion on the differences and Prof. Treacle explains how most scientists would organize them. Peter starts to see how it all fits together.</i></p> | | | | | |

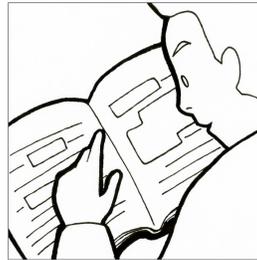
11 Appendix D: Scenario sketches used in Needs Validation

Scenario I



Peter has an exam on Tuesday and sits down in the library Sunday afternoon to study.

Before cracking the book, he takes a short online quiz on his laptop. The system gives him feedback on what he knows and doesn't know.



While reading through his book and notes, he pays more attention to what the system said needs help in.

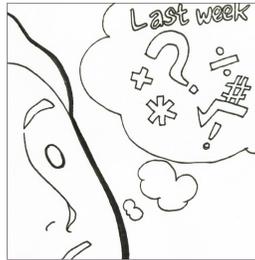


A few hours later he opens his laptop again and takes another quiz. It shows him he did better on all but one part. He is satisfied with his progress and leaves the library.

Scenario 2



Peter crammed and did poorly on the first exam. He feels like he knew it at some point but forgot. To help he goes online to enroll in the optional RetentionBuddy system Prof. Treakle has set up for the class.



The next morning checking his email at the dinner table, he sees a message with questions about some stuff covered last week and two weeks ago in class. He sort of remembers the stuff from last week and is frustrated he doesn't remember the stuff from two weeks ago, but he thinks through it and comes up with the answer. He replies to the email with his answers.



A week later he gets an email with questions from one, two and three weeks ago. He is glad that he can answer the stuff from three weeks ago that he had trouble with last time.

Scenario 3



Claire is in her dorm rooms working on the first homework assignment in her Intro to Psychology class. She has to describe her career plans and imagine how an understanding of psychology could help her.

She writes what she wants to be a nurse traveling around the world. Her friend Lisa wants to be a teacher.



The next week in lecture, there is a slide listing subgroups in the class based on people's answers to the survey. Prof. Forbes instructs the groups to find each other in the lecture and meet up. Throughout the semester they will have customized assignments related to their goals.

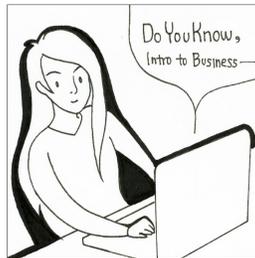


Over the weekend Claire talks with Lisa about their homework on Attachment Theory. Claire's group had to relate it to baby incubators in hospitals while Lisa had to relate it to kindergarten teachers.

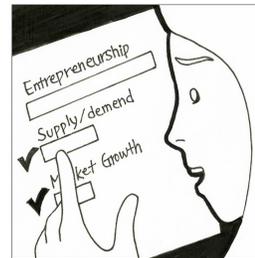
Scenario 4



Claire got her Intro to Business midterm back and was disappointed that she missed some key areas. She was pissed because she spent so much time on another area that was barely on the exam.



Before the next quiz she goes to the library and tries the DoYouKnow system on her laptop for help on what she should focus on. First it lists the areas she is expected to know and then she has to rate her confidence in each area.



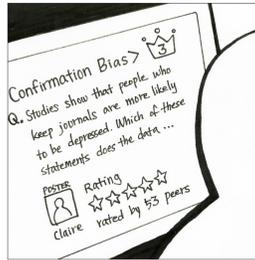
She answers a series of quiz-like questions that get easier or more challenging depending on whether she is getting them right. When she's done, she gets a report on how well the system thinks she knows the different areas. She is surprised that the system recommended she study more on supply & demand and market growth because she thought she understood those.

She starts studying those trouble areas and worries less about the others.

Scenario 5

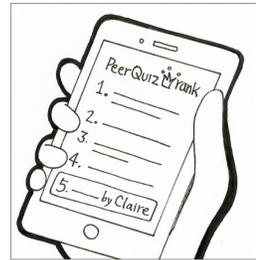


Claire sits down at her kitchen table and logs into PeerQuiz to do her homework. She has to write questions and answers that require applying psychology theories to particular situations. She writes three questions.



Next she looks over questions that others wrote. She clicks to rate them and types comments to critique them. When she finds a question she likes, she types a paragraph to answer it.

Before she logs off she sees that one of her questions is now the 3rd highest rated.



The next day she gets a text message with the current ratings of her questions and answers. She sees that her top-rated question has fallen to 5th place. She also sees that her answer is rated highly.

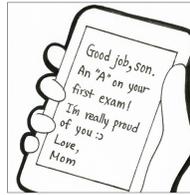
Scenario 6



At the first lecture in Intro to Physics, everyone has to create a list of people to share their homework and grades with using LearnShare. Peter shares his exam grades and attendance with his parents, his problem sets with his friends who are interested and everything with his sister.



After his problem set on mechanics he gets an email from LearnShare. His math whiz friend says it's cool that Peter is learning this stuff. He says that Newton's laws are only an approximation and the relativistic model is even more interesting.



After the first exam he gets an email from his mom, "Good job, son. An A on your first exam!"

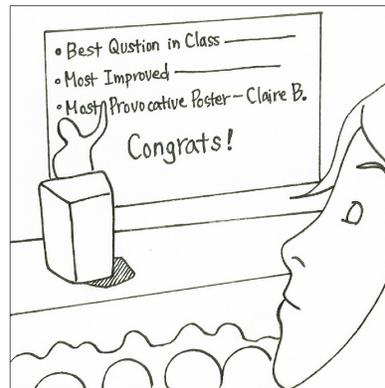


A few weeks later he is struggling and gets a D on the quiz. His sister calls him to ask how he's doing.

Scenario 7



Claire is taking Intro Psych to meet her science requirement. At night she goes online to write her two weekly posts to the class forum. She didn't have time to read the textbook so instead she challenges other students to say why the stuff matters. She gets into heated exchanges and posts many more than two messages per week.

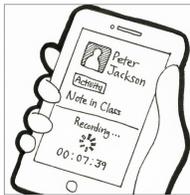


Later in class the midterm grades are posted on the projector. She sees the top names and looks down to find hers with 76%. On other slide she sees other names for Best Questions in Class, Perfect Attendance, Most Improved, and see her name again under Most Provocative Poster.

Scenario 8



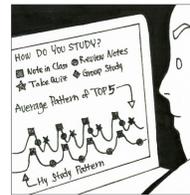
In the first day of Intro Psych, Prof. Treakle tells the class that part of their grade will be for reporting their study activities and the data will be used for in-class analysis on the psychology of learning.



Peter doesn't study very carefully. Normally, he just takes notes in class and skims the textbook. The lecture starts and he records on his phone that he is taking notes in class.



After class he reads over his notes several times and records this.



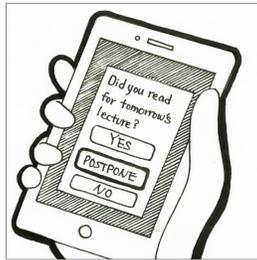
Weeks later he receives his first exam back with a C. Along with his grade is a full-page report on the study activities of the students who did best and those who did poorly. Other students who spent as much time but got better grades quizzed themselves while reading, so he decides to do that next time.

Scenario 9



Claire gets her grade back on the last exam and it's lower than she expected. For this next unit, she decides to study regularly.

At home in the evening Claire uses the TimeGoal system to set goals for what she will accomplish each day to prepare for the next exam. She resolves to read or at least skim the reading before class on Wednesday. She also wants to visit the TA's office hours on Thursday.



Tuesday night at 9pm she receives a message on her phone asking how much she read. She clicks to postpone it. Thursday a message asks whether she went to office hours. She clicks No.



On Friday she doesn't feel confident for Monday's quiz and pulls up a report of whether she's studying more regularly like she wanted. It shows her that she skipped both the reading and the office hours. She has to cancel some weekend plans to cram for Monday's quiz.

Scenario 10



Thursday night, Peter realizes he has a big essay due Monday morning and he'll have to work hard on it over the weekend.



At 9pm on Friday night Peter is tired of reading and decides it's time to go out. He records on his phone that he's switched from Studying to Partying.



Saturday afternoon Peter picks up his essay again and records Studying on his phone. After an hour he records he'll check Facebook for just 10 minutes. He loses track of time until the phone beeps to remind him and he returns to studying.



Sunday evening Peter has completed his final draft and submits it early. With satisfaction he records Watching TV.

Scenario 11



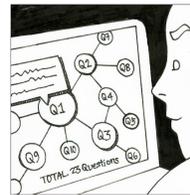
On the first day of his Intro Psych class, Peter has to fill out a form with the big questions he has about human psychology and his goals for the course.



After the first week Peter is at home on his computer and gets an email reminding him of his initial interests. It asks how the week helped satisfy them and prompts him for new questions he has.



Each week he gets another email with his past questions and answers. He adds new questions and expands earlier answers. He feels like he's learning what he cares about.



At the end of the course he receives a full report of the big questions he had and what he learned to answer those questions. He read it and feels a sense of accomplishment.

Scenario 12



In lecture Prof.Treakle pushes to ask the class questions. She presents many questions on a slide with different difficulty levels side by side. Peter chooses the hardest one and enters his answer on his phone.

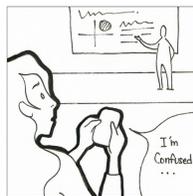


He gets back a message that he was right along with how many other students attempted and succeeded at that question.



Prof.Treakle can see that Patrick answered the easiest one and calls on him to explain his answer.

Scenario 13



During her Developmental Psychology lecture, Claire has a question but thinks it's too dumb to raise her hand. Instead she asks the question anonymously on her phone.

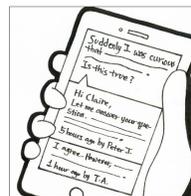


Allison the TA is reading each question and ranking them privately.

Every 20 minutes Prof. Treakle shows the top questions on the projector. Claire's is on the list and Prof. Treakle answers it.



Later in the lecture she has another question which is off topic. She sends this one in with her name to be sure she gets an answer. In the next question block Prof. Treakle doesn't choose it to answer.



That night, she gets an email. Another student liked her question and wrote an answer. Allison the TA marked the answer as acceptable and added a few clarifications.

Scenario 14



In the library one evening Peter works on his homework. He has to write questions that test the key ideas from lecture. He logs into CrowdExam and types in questions and their answers. One of them is too similar to another student's question and he changes it.



The next day he receives an email that the TA has rated his questions and two of them are 5 stars out of 5.



While taking the midterm, he gets to a question that is one of his. The professor picked it, among the top-rated questions to include on the exam.

Scenario 15



In her office, Prof. Treakle edits the slides she will use tomorrow in class. For one, she adds a new YouTube video to demonstrate Change Blindness.



In class, while she shows each slide students click on their devices with a rating of whether they understand the points. Some students text in with suggestions.



At a break she solicits questions. Allison the TA writes these down. Students rate their satisfaction with professor's answers.



In her office that afternoon, Prof. Treakle looks at her slide deck with the student feedback and questions beside it. She identifies some slides that could use more work.

She also sees that students were more confused by the YouTube video than last semester's were with her earlier animation.

Scenario 16



In his Developmental Psychology class Peter has an assignment to grade the quiz of a student who finished the class last semester. He logs into OlderCheck and reads their answers. He writes the correct answers using the class answer key and explain them in his own words. He does this for three different students.

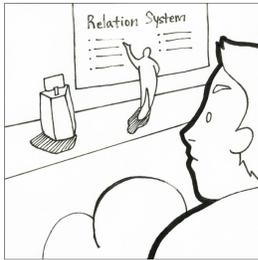


During the midterm, some questions are easier because he remembers his explanations on OlderCheck.

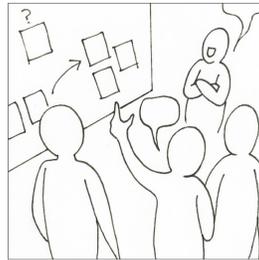


Several months after finishing the course, he gets an email with a quiz on the same stuff he tutored in OlderCheck. He tries answering the questions but can't remember everything. He looks forward to the explanations from the current students so that he can remember what he learned and not have it be a waste.

Scenario 17



In lecture after the midterm Prof.Treakle says the students need to see the bigger picture of psychology and introduces the Relations system. Peter types out all the most important ideas and research he can think of from the class.



The next week he meets with his assignment group and they lay out strips of paper on the floor with all their ideas printed on them.They debate with each other how the ideas all fit together and when they agree they snap a photo which they upload to Relations.



In the next class Prof.Treakle shows a slide with a giant map of all the concepts that everyone put together. He shows what relationships most students agreed on and what ones were controversial.The class has a big discussion on the differences and Prof.Treakle explains how most scientists would organize them. Peter starts to see how it all fits together.

12 Appendix E: All tasks reminded or polled by Nudge

| Task group | Due | Importance | Time estimate | Description | Note |
|------------------|--------|------------|---------------|---|---|
| Participation | 29-Aug | required | 20 min | Take the concepts quiz in first lecture | |
| HW1 | 30-Aug | advised | | *Ch2*: 21, 23, 30, 31, 37, 38 | |
| HW1 | 1-Sep | advised | | *Ch2*: 55, 63; *Ch3*: 23, 25, 27, 31, 37, 41 | |
| Quiz 1 | 4-Sep | advised | | Practice [Quiz 1 problems] | |
| Participation | 5-Sep | required | | Fill out [Week 1 questionnaire] in Blackboard | |
| HW1 | 6-Sep | advised | | *Ch3*: 47, 55; *Ch4*: 17, 19 | |
| Mastery exam I | 7-Sep | advised | | Practice [Mastery problems] | |
| Quiz 2 | 12-Sep | advised | | Practice [Quiz 2 problems] | |
| HW1 | 13-Sep | required | | Do and submit [HW1] | |
| HW1 | 13-Sep | advised | | *Ch12*: 21, 27, 31, 37, 39, 43 | |
| HW2 | 15-Sep | advised | | *Ch12*: 18, 49, 55, 57, 75, 77, [S1] | |
| Exam I | 18-Sep | advised | | Practice [Exam I problems] | It's best to try working through the problems several days in advance to see how you'll do on the exam. |
| HW2 | 20-Sep | advised | | *Ch12*: 11,13, 83, 85, 87, 125, 129 | |
| HW2 | 20-Sep | required | | Do and submit [HW2] | |
| Mastery exam II | 25-Sep | advised | | Practice [Mastery problems] | Practicing in advance will help you learn and focus on what to review. |
| Quiz 3 | 27-Sep | advised | | Practice [Quiz 3 problems] | |
| HW3 | 29-Sep | advised | | *Ch13*: 17, 33, 51, 65 | |
| HW3 | 4-Oct | advised | | *Ch13*: 7, 55, 57, 61, 67, [S2] | |
| HW3 | 6-Oct | advised | | *Ch13*: 37, 39, 41, 72, [S4] | |
| HW3 | 7-Oct | required | | Do and submit [HW3] | |
| HW4 | 11-Oct | required | | Do and submit [HW4] | |
| HW4 | 11-Oct | advised | | *Ch3*: 77, 79 | |
| Exam II | 12-Oct | advised | | Practice [Exam II problems] | It's best to try working through the problems several days in advance to see how you'll do on the exam. [...] |
| HW5 | 13-Oct | advised | | *Ch13*: 81, 89, [S5] | |
| Quiz 6 | 17-Oct | advised | | Practice [Quiz 6 problems] | |
| HW5 | 18-Oct | advised | | *Ch14*: 4, 31, 34a | |
| HW5 | 20-Oct | advised | | *Ch14*: 35, 41, 59 | |
| HW5 | 20-Oct | required | | Do and submit [HW5] | |
| Mastery exam III | 21-Oct | advised | | Practice [Mastery problems] | If you haven't passed the Mastery yet, here's your next chance. |
| Participation | 24-Oct | required | | Fill out [3rd exam questionnaire] about CourseCheck | It takes 5-10 minutes and is the last general questionnaire until the end of the term. |
| Quiz 7 | 26-Oct | advised | | Practice [Quiz 7 problems] | |
| HW6 | 27-Oct | advised | | *Ch14*: 33, 37, 39, 43 | |
| HW6 | 1-Nov | advised | | *Ch14*: 13, 19, 23, [S7] | |
| Quiz 8 | 1-Nov | advised | 20 min | Practice [Quiz 8 problems] | |
| Exam III | 2-Nov | advised | 90 min | Practice [Exam III problems] | |
| HW6 | 3-Nov | advised | | *Ch14*: 25, 73 | |

| Task group | Due | Importance | Time estimate | Description | Note |
|-----------------|--------|------------|---------------|---------------------------------------|---|
| HW6 | 3-Nov | required | | Do and submit [HW6] | |
| Exam III | 5-Nov | advised | 50 min | Timed practice of [past Exam III] | |
| HW7 | 8-Nov | required | | Do and submit [HW7] | |
| HW7 | 10-Nov | advised | | *Ch14*: 45, 67, [S8] | |
| Participation | 10-Nov | required | 4 min | Fill out [mid-semester questionnaire] | If you enrolled in the study, taking a few minutes to fill out this survey is required to maintain participation. |
| HW8 | 15-Nov | advised | | *Ch5*: 27, 31, 67, 83, 89 | |
| HW8 | 17-Nov | advised | | *Ch16*: 16, 17, 18a-e, 21, 23 | |
| Quiz 9 | 17-Nov | advised | 20 min | Practice [Quiz 9 problems] | |
| Mastery exam IV | 18-Nov | advised | | Practice [Mastery problems] | If you don't pass this one, you only have one more shot. |
| Mastery exam V | 25-Nov | advised | | Practice [Mastery problems] | This is your last chance to pass the Mastery Exam if you haven't already! |
| HW8 | 1-Dec | required | | Do and submit [HW8] | |
| HW8 | 1-Dec | advised | | *Ch19*: 7, 29, 33, 37,63, [S9] | |
| Exam IV | 5-Dec | advised | 90 min | Practice [Exam IV problems] | |
| HW9 | 6-Dec | required | | Do and submit [HW9] | |
| HW9 | 6-Dec | advised | | *Ch19*: 45, 47, 49, 77 | |
| HW10 | 8-Dec | advised | | *Ch19*: 41, 51 | |
| HW10 | 8-Dec | required | | Do and submit [HW10] | |
| Exam V | 8-Dec | advised | 5 min | Plan for studying [Exam V topics] | |
| Exam IV | 8-Dec | advised | 5 min | Plan for studying [Exam IV topics] | |
| Exam IV | 9-Dec | advised | 50 min | Timed practice of [past Exam IV] | |
| Exam IV | 9-Dec | advised | | Review notes for Exam IV | |
| Exam V | 9-Dec | if needed | | Practice [Exam V problems] | |
| Exam V | 10-Dec | if needed | 90 min | Timed practice of [past Exam V] | |
| Exam V | 11-Dec | if needed | | Review notes for Exam V | Will you be taking Exam V? |