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 Examplify 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 Examplify

6.4.1 Motivation

Examplify 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

Examplify 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. Examplify 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 Examplify 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 Examplify and its evaluation is that metacognitive tutors can be effective without evaluating student work. In Examplify, 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

Examplify 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 Examplify 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 Examplify research pursue: optimization and explanation. Like the supersonic jet research described in Chapter 2, as an operant probe Examplify 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 Examplify 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 Examplify 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 Examplify 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 Examplify 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 Examplify 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, Aleven, 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 Examplify. 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).

6.5.6.1 Riskier designs

Both Nudge and Examplify 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.

Examplify 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.

6.5.6.2 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.

6.5.6.3 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 Examplify 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. Examplify 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.

6.5.6.4 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.

Examplify 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 Examplify can be monitored and improved over time. In designing Examplify, 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, Examplify 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.

6.5.6.5 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.

6.5.6.6 Involve stakeholders in judging the quality of designs

Nudge and Examplify 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 Examplify 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 Examplify 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 Examplify, 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. Examplify (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 Examplify 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.