A socio-technical approach to feedback and instructional development for teaching assistants

David Gerritsen

August 2018
CMU-HCII-18-102

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Thesis Committee:
Amy Ogan (co-chair)
John Zimmerman (co-chair)
Ken Koedinger
Marsha Lovett

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through grant R305B090023 to Carnegie Mellon University. This material is also based upon work supported by the National Science Foundation under Grant Numbers 1464204 and 1747997. Any opinions, findings, and conclusions or recommendations expressed in this material are my own and do not represent the views of the Institute or the U.S. Department of Education, nor do they necessarily reflect the views of the National Science Foundation.

Keywords: teaching assistants, pedagogy, professional development, instructional development, educational development, teaching consulting, longitudinal studies, classroom interventions, smart classrooms, technology enhanced learning

To Vivian for the inspiration,
Ruth for the ambition,
and Tycho for the motivation.

Abstract

Teaching assistants (TAs) in the United States play a prominent role in educating undergraduates. Their influence can make the difference between students continuing in their majors or leaving them. However, most TAs use teacher-centered, transmission models of teaching, i.e., lecturing to disengaged students. Part of the reason for this is that most TAs receive little training on how to teach, and almost no grounded feedback about their teaching behaviors.

In this thesis I describe my work investigating the use of technology to increase feedback and training for TAs. My focus is understanding how their knowledge, skills, beliefs, and attitudes should drive the design of algorithms for gathering classroom behavioral data and delivering computer-mediated feedback and consultation. My work evaluates a novel framework for investigating how TAs interact with their data, reflect on what it means, and decide what (if anything) to change in their teaching. I examine how initial beliefs can impact their system interactions, how those beliefs change over time, and the resulting implications for designing data-driven training artifacts.

Acknowledgments

Formal learning has been one of the most prominent features of my life. I like it so much that I ended up researching it, designing it, and turning it into a career. Yet as important as formal education is in stretching my mind, it is informal learning that leads my heart. The soul of my dissertation work was born and raised through each of the unscripted, unexpected interactions I had in the distinguished halls of Carnegie Mellon University, the buzzing corridors of the Human-Computer Interaction Institute, and the inclined floors of the Program in Interdisciplinary Education Research. In thanks to those who made this adventure possible, here is what you all taught me.

Amy never relented in pushing me to do hard things, despite my persistent resistance. I like to think that we're both stronger, but I know I got the better end of that deal

John taught me what really mattered. More important, he taught me to recognize what did not.

Marsha showed me how to think bigger. Ken taught me to keep asking questions.

Trohoc—Anthony, Brandon, Chris, Dan, Jenny, Nikola, and Tati—showed me how to be a friend, enjoy a dinner, and produce ridiculous email threads all within in the midst of the madness. Your families are mine. JoAnna, Curie, Annie, Kabir, Caitlin, Shuang, and Ryan have made us all even better. Being the last one to finish, let me take this opportunity to publish one more clap.

Iris, Jim, and Tyler taught me how to play again, and reminded me that I love building big, pointless things. Let's meet up in 1.13.

In no particular order, Kelly, Samantha, Jeff, Beka, Will, Jason, Eliane, Eiji, Chris H., Kevin, Gabi, Stephen, Martina, Julia, and Nesra all taught me how to be supportive and kind to those who come after you.

Elijah taught me how to live well sharing chicken tikka masala and countless revisions.

Erik Harpstead, Steven Dang, Robert Xiao, and Adrian de Freitas each saved my code at various points, and helped me admire the magic of MySQL, JavaScript, NumPy, and dry-erase sketches of state machines.

Jessica Hammer and the members of the Oh!Lab taught me everything from how to fix my CV to how to play Carcassonne. I'm especially looking at Judith, Judy, Alexandra, Amy C., and Evelyn. You are my human side.

Evelyn Yarzebinski in particular. You have literally been there from day one, showing me how a colleague truly gives moral, practical, and spiritual support in every endeavor.

Queenie showed me how to be patient with someone who constantly forgets what he was supposed to be doing for his PhD requirements.

There are simply too many others in the HCII community for me to list all of their contributions here. Suffice it to say that you all showed me how a department should continue to look inward and iterate.

The broad PIER community taught me nearly everything I know about the learning sciences. Additionally, Sharon Carver showed me that my intuition and life experiences are my best guides.

Emily Keebler instructed me in the art of graceful parenting while under pressure. Audrey Russo taught me to bring my baby boy to the dinners, because he's the one folks really want to see. And David Klahr convinced me to finally, one and for all, just calm down.

The Eberly Center and my Graduate Teaching Fellows taught me the lion's share of what's in Chapter 2. You all made the intellectual aspects of this research grow beyond what I thought was possible.

Sherice, Amy, and Rebecca showed me how to collaborate. I miss our meetings.

My family—Angela and Steve, Micah and Cori, David and Sarah, Murr and Old Dave, Anna, and the estimable Drs. Nivea Maria Vega Longo Reidler and Ruth Gerritsen-McKane. You and the rest of our kin taught me how to be proud of myself.

Vivian taught me that love is action. I am forever ready.

Tycho. You're new here. But you have already taught me the most profound lessons, like how to carry something precious. I hope that you come to love learning as much as your silly father does. I will do my best to make it fun.

Table of Contents

Chapter 1: Introduction

People need feedback in order to learn.

This is true in every domain and at every age: learning to stack blocks, spell, drive a car, or navigate a career. As the complexity of a task increases, the feedback becomes more nuanced, and more critical. Blocks fall or stand. Teachers (or word processors) point to incorrect spellings. Cars emit sounds, smells, and shakes. Meeting with peers and mentors, promotions, and even reprimands all shape the development of a career.

Today, technology for delivering feedback and oversight in complex situations looms over every profession, including the modern classroom. Teachers are already familiar with "smart" boards and interactive multimedia. Students are familiar with more diret interventions such as computational learning environments and educational games. Behind the scenes, researchers and engineers in the learning sciences have built massive datasets tracking what students do in online learning spaces. As a result, most research on technology enhanced learning in the classroom has focused on bypassing teachers and delivering feedback directly to students, or else providing teachers with feedback about their students as they work. Although results vary, some students have seen substantial learning gains when they receive direct, personalized feedback. And experienced teachers with a solid foundation of pedagogical content knowledge can use real-time data to make sophisticated decisions about the needs of their students.

To date, though, these technologies have done very little to support teachers in improving at their craft.

In order to teach well, professionally trained adults receive complex feedback. Years of coursework are available to soon-to-be educators where they can master pedagogical principles for improving the learning of their future students. New teachers can improve with feedback from their seniors or peers, through slow processes of human observation and review. Perhaps a consultant attends class, takes note of what happens, reviews the notes, then sits down for an in-depth conversation after the fact. The feedback from these interactions is rich and informative for these budding experts. But in reality, this deep interaction is rare, and particularly in universities, it often does not happen at all.

1.1 Teaching in Higher Education in the United States

On today's American college campus, instructors are typically hired for content expertise rather than teaching skill, particularly in large state schools and research institutions. The impact these instructors have is considerable. They may encounter thousands of students over their time teaching. Yet college instructors currently get very little feedback or training, and few if any incentives to hone their craft at the expense of time spent in research or service.

The situation is a challenge for freshman and sophomore courses, where teaching assistants (hereafter, TAs) bear the brunt of the burden of teaching. TAs are responsible for a large amount of course material, especially in state universities and for introductory courses. As a result, a large portion of teaching responsibility has been offloaded to students only slightly more advanced than those students in their courses. There are risks when leaving untrained novices in charge of guiding freshman and sophomore undergraduates. These early-stage students are already vulnerable to dropping majors when they encounter difficult subject matter. Some students may also be at risk of leaving the university

altogether. TA-led recitations should be a benefit to and support structure within a broader environment of higher learning. But they are often an obstacle for students to overcome.

The blame should not fall on the TAs themselves. TAs face several important obstacles when they teach. As a population, they have very low rates of prior teaching experience. Many are teaching for the first time in their lives. One of the goals of including current students as instructors is to prepare them for future teaching by gaining experience in front of the classroom. However, entering the teaching pool, TAs receive extremely little pedagogical training, and best practices from learning science are rarely employed to improve how undergraduates learn or how college classes are taught. TAs are often not trained to encourage active learning; to notice what is happening in their classes; to engage their students in the material; or to avoid the natural "information transmission" style of teaching that is natural from domain experts. Research shows that this "teacher-centered" method of lecturing does not work for many students. It emphasizes memorization over comprehension, and shallow recall over deep encoding.

This approach to education prioritizes the least meaningful stages of learning, and rewards students who can afford to laboriously teach themselves the material. Students who have obligations that extend beyond the curricular requirements of their majors, such as family or work, may find it nearly impossible to advance. This effectively filters out students with unique challenges, possibly undermining the diversity needs of the academic and professional worlds.

As an educational system, there exist unnecessary barriers for students in need. Institutions rely on TAs but fail to train and support those TAs. My research addresses the problem of inexperienced TAs who exclusively use lecture as their teaching method, and the ways we can move forward to help them grow as educators.

1.2 Contributions of this thesis

1.2.1 Contributions to learning science

My work combines threads from human-computer interaction and the learning sciences to produce new insights into how TAs in higher education might improve. Each sample of TAs in this document exhibits familiar teaching strategies that emphasize top-down, information-transmission modes of instruction. A sufficiently experienced student stands in front of a class and gives students as much raw information as possible in 50-minute increments. They hope it is enough for everyone to finish the homework or pass the test.

I use data about these teaching patterns to elicit reflections from the TAs. I ask them to assess their teaching performance and to set goals for improving their practice. I expose them to "discursive teaching techniques" to help them reach those goals. These techniques are specific steps for engaging students in meaningful spoken interactions with each other and with the instructor. Throughout this process I discover:

- A viable cycle of instructional design/training for TAs
- Archetypes of TA reactions that lead to more (and less) instructor development
- A framework for how to align TA instructional patterns with technology-enhanced learning environments

- How the beliefs and attitudes of the TA may interact with the success of their instructional development

Research shows that when college students speak more during class—by thinking through deep questions and participating in class discussion—they learn more than when they listen to a lecture and solely take notes. This is true even in STEM fields (science, technology, engineering, and mathematics), where traditional perspectives emphasize the memorization of facts and formulas over the higher-order use of synthesis and analysis.

My research provides an innovative approach to increasing interactivity in the novice-led university classroom. I focus on TAs as learners. I explore how the new world of sensor-based data collection in the classroom might deliver learning opportunities to these new instructors. I learn how the population of TAs might react to these opportunities.

1.2.2 Contributions to classroom technology

I address a gap in the current literature by investigating technology in the hands of non-expert instructors, for the purpose of giving them feedback about their teaching. This interaction requires an all-new type of professional development for instructors, which I will refer to throughout this document as SmartPD.

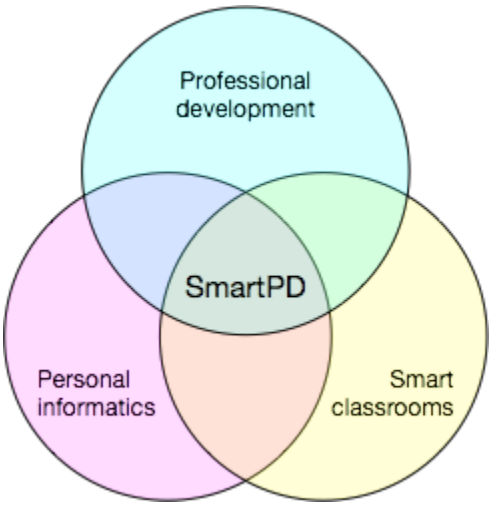


Figure 1.1: The intersection of disciplines and technologies needed to produce SmartPD.

SmartPD is a combination of sensors, models, algorithms, and educational resources delivered over time. It combines features of Personal Informatics, technology-enhanced ("smart") classrooms, and professional development (PD) for teaching. PD of this type means many things. For this work I focus primarily on teacher feedback through consultations with educational experts. I choose this approach because of its potential to emphasize direct evidence about teaching behaviors.

I posit that in-class sensors offer unique opportunities to improve teaching. Monitoring and recording real-world actions allows technologists to build a database of "teaching informatics;" like a cross between learning analytics and fitness tracking. The potential to use these data has not been critically analyzed by designers of educational technology. Much of the thinking in this space has been focused on personalized learning and informatics for students alone. I lay the groundwork for using personal informatics in pursuit of grounded, evidence-based feedback to teachers as they learn about teaching.

1.2.3 One technology in particular: *ClassInsight*

PD opportunities are necessary in order for teachers to reflect on their experiences, set goals, and change their praxis over time. SmartPD is a new genre of tools to create those opportunities with the personal informatics data from sensors in a smart classroom, help determine the skills of a teacher based on the granular detail of what happens in their classroom, identify their appropriate learning pathway, and possibly provide hints or recommendations in the form of intelligent tutoring systems for professional learning. *ClassInsight* is the first instantiation of SmartPD.

In this early stage, I use human generated data to simulate plausible sensor technologies. This allows me to explore the design requirements for a SmartPD artifact. These design requirements will help inform the continued development of *ClassInsight* as it integrates with actual sensors and communicates with instructors. I use a design-based research (DBR) approach to generate research questions, design and test prototype systems, and explore how participants learn and change over time, in part due to the tools we build.

1.2.4 Contributions on the people who teach

My questions are about the cognitive, temporal, and dispositional aspects of TA professional learning. It is in the intersection of these questions that we break new ground. This work is not about advancing the sensors used in smart classrooms. It is about supporting the *people who teach*. This leads to at least three critical outcomes of my thesis, all of which can advance theory, but which are deeply practical in nature.

One important output of this work is the Plan/Act/Reflect framework, a theoretical contribution and conceptualization of teachers' experience as they hone their skills, the inputs and outputs through the phases of learning how to teach. I use this framework to define the necessary features of a technology-enabled intervention, and to determine when and how to implement those features and insert them into the rhythm of TA teaching time.

A second important output of the work is a set of variables and constructs that relate TA behavior with the system to their personas in the classroom over time. These variables—*productive self-doubt*, *shallow dismissal*, and *confirmatory assessment*—emerged through repeated implementations of SmartPD across multiple studies, embedded in research artifacts and elicited through qualitative analysis. Taking a mixed-methods approach, we lay the groundwork for how future quantitative work can enrich our understanding of these variables.

Uniquely, the design of the studies in this thesis collects fine-grained and deep data on our TAs and their experiences throughout periods lasting many weeks, well beyond the scope of most learning science interventions. This includes classroom data, annotated and timestamped down to the second, collected through the emulation of smart classrooms. As an emulation of passively observing classroom sensors, this data were collected even when TAs did not follow through with adherence to study designs. Our data also include exit interviews conducted in parallel with traditional, discretely measurable assessments like surveys.

This unusual data collection leads to a third output of this work. Using the new variables of TA behavior, we can begin to more deeply understand adherence to professional development regimens, the people who stick with those regimens, and those who don't. Our findings begin to cast doubt on whether non-participants or dropouts have been modeled appropriately in past research. The picture

that emerges is of a rich feedback effect, and a nuanced relationship between an educator's personal characteristics, technology interventions, and social incentives. It reframes the phenomenon of non-adherence as something more complex than automatic failure. This ecosystem of factors will add nuance to future studies at all granularities: of *ClassInsight*, of SmartPD as a framework, and broadly of studies of educator professional development throughout the learning sciences literature.

1.3 Structure of this document

This thesis describes the multi-year development of the idea of SmartPD. It presents the outcomes of implementation of not only *ClassInsight*, but also the prototype experiments that led to it. The work is divided into two major parts.

In the first part—chapters 2, 3, and 4—I work to inform this research with a base of knowledge about what works in professional learning, and to identify imminent opportunities of sensor-enabled smart classrooms. Chapter 2 walks through the literature that inspired this research. I review related work in PD, college instruction, PI, and smart classrooms. I make the argument that combining these fields produces a novel set of research questions. I also describe my implementation of educational design-based research methods. In chapter 3 I investigate the problem space and begin designing possible solution strategies. This chapter focuses on what is known about TAs in American institutions who teach STEM courses, then describes a field study where I gather contextual data about events in these classes. I use those data to reveal paper prototypes of PI inspired visualizations to the participants and then follow their reactions and teaching behaviors to explore possible impacts these visualizations may have. In chapter 4 I build and deploy an initial prototype system for training TAs through an online system. It is an iterative design and implementation of a specific instance of what will become SmartPD. Through a combination of DBR and user-centered methods, this experiment uncovers themes which will produce the Plan/Act/Reflect-Training/Sensing framework.

The second part of this dissertation—chapters 5, 6, and 7—builds on these findings to develop a preliminary framework for PD, and enters the classroom with a technical intervention. In chapter 5 I define that framework along with the definition of SmartPD. I explain the iterative design cycles that uncover useful methods for approaching these novice instructors; I explain how the findings of the studies thus far produced that framework, and how it informs our understanding of emerging genres of socio-technical systems in education. I also show how the beliefs of the user may provide an important constraint on future design. In chapter 6 I describe the first test of a possible application of SmartPD. This design advances the previous training system by combining direct instruction with individual, personal reports of in-class data. The findings produce evidence that some individual differences may predict how users respond to this support and how those responses should be interpreted, leading to the introduction of new variables of TA behavior. In chapter 7, a final study implements a new app and set of algorithms that automate many of the previously human-guided steps of TA training. This is the first step in transitioning from traditional PD for instructors into SmartPD. I describe the challenges and successes of this approach, and how this data further informs our variables of TA behavior. With this richer data, I develop a more complex, nuanced set of interpretations of individual differences in TAs and their interaction with the new app.

Chapter 8 synthesizes these many studies, explaining and summarizing the major contributions of the work. This includes design considerations for researchers and practitioners that are interested in using classroom sensors to produce feedback and training for teachers, whether they be TAs in higher education, or educators in any of the many other domains of learning.

Chapter 2: Context of the research

2.1 What we know about college classrooms

American professors have a lot to do. There are endless tasks associated with research, service, and teaching. As it stand, teaching in many universities is held in lower status than more visible (and lucrative) acts of research (Fairweather, 1993; Mauksch, 1986). Because of this, departments often hire domain experts rather than instructional superstars when it is time to fill faculty and TA positions. Rarely is there a culture of curricular collaboration or encouragement to gain teaching expertise (P. J. Baker & Zey-Ferrell, 1984; M. D. Cox, 2004). Most faculty development initiative are completely voluntary (Gullatt & Weaver, 1997), and campuses struggle to adopt a "learner-centered" culture of teaching (B. E. Cox, McIntosh, Reason, & Terenzini, 2011).

Although a majority of faculty believe their teaching could be better, they lack the time, incentives, or motivation to participate in formal training or evaluations (Berman & Skeff, 1988). Most of the feedback college instructors receive comes from student ratings at the end of a course. These have little impact on teaching, however, as professors are skeptical of their validity (Marsh, 1984; Marsh & Roche, 1997; Richardson, 2005). Even for faculty that perceive student feedback as useful input, they have little opportunity to address concerns that are only raised after a semester is complete (Penny & Coe, 2004). The tangible impact of student evaluations is typically their influence on promotion decisions (Haskell & Theall, 1997).

The focus in higher education is often on tenure and promotion rather than teaching. This may explain why few college instructors have developed "active learning" or "student-centered" practices (B. E. Cox et al., 2011; Stains et al., 2018). These approaches, which I explore in greater detail later in this chapter, are skills for teaching with proven advantages for learning, even in college STEM classes (Freeman et al., 2014). Eliciting student participation in class is shown to produce deeper learning than simply requiring them to listen and take notes (Chi & Wylie, 2014; Nunn, 1996). Nevertheless, professors and TAs continue to do most of the talking (Nunn, 1996).

The majority of in-class interactions are top-down lectures with very little student involvement (Stains et al., 2018). During lectures, students encounter mostly shallow questions that elicit only brief answers (Nunn, 1996). Their opportunities to answer are fleeting, as instructors do not usually wait long enough after asking a question before they move on (Larson & Lovelace, 2013). The use of lecture as a primary instructional strategy makes sense in large halls with more than 45 students. Student participation in these classes can be difficult to manage. There may only be a few opportunities for students to speak in such an environment (Fritschner, 2000; Rocca, 2010). But many large classes have breakout sections of recitations, discussion groups, or labs taught by graduate students and advanced undergraduates (Friedman, 2017). These small classes present an opportunity to introduce meaningful interactions between students and their instructors.

2.1.1 Teaching assistants

TAs are responsible for a large number of classes, seminars, discussion groups, and labs. At last count there were over 130,000 TAs employed in the U.S. (Bureau of Labor Statistics U.S. Department of Labor, 2016). In 2015, TAs were the primary instructor for as many as 26% of undergraduate courses in some institutions (Friedman, 2017). This figure does not count the even larger number of labs, discussion groups, and recitations that TAs lead.

Despite teaching a large number of undergraduates, TAs receive very little pedagogical training (Hardré & Burris, 2012). For example, in a large survey of American mathematics departments, only two-thirds provided professional development for graduate TAs (J. Ellis, Deshler, & Speer, 2016), and of those, only half of eligible TAs volunteered. Even for those who receive training, there is no standardization of procedures, and at present the field has no strong theories about the professional development for TAs (Stes, Min-Leliveld, Gijbels, & Van Petegem, 2010). Most of the training that does exist tends to address classroom management and domain content rather than pedagogy, leaving many TAs who desire to work in academia feeling unprepared to become productive professors (Austin, 2002).

As a majority of TAs receive unstandardized training—if they receive any at all—this population tends to exhibit poor teaching performance (O'Neal, Wright, Cook, Perorazio, & Purkiss, 2007; Wilen & Clegg, 1986). They bring to the classroom inaccurate impressions of what constitutes good teaching, mimicking what they have seen as lifelong students, but without any insight into why teachers do what they do (Borg, 2004). This leads to myriad problems, such as a tendency to equate lecturing with teaching (Brownell & Tanner, 2012), and an underestimation of the abilities and motivations of undergraduates (Luft, Kurdziel, Roehrig, & Turner, 2004). Adding to these concerns, institutional emphasis on research can give TAs the sense that teaching is unimportant (Brownell & Tanner, 2012), as well as raise challenges for attempts to institute broad changes in TA training..

To start thinking about what could be done differently, we turn to an overview of strategies that work for effective teaching, and the rich history of professional development for teachers, along many pathways of intervention.

2.2 What we know about good teaching

This thesis involves "professional development for teachers," for which I use the acronym PD. A broad, inclusive definition of this term describes efforts to improve the pedagogical practices of instructors at any level and in any occupational domain. By that definition, PD is a sprawling and highly active field of research and practice. Without clarification, this thesis could be interpreted broadly and decontextualized from the domains that I study. In this chapter, I hope to make my domain of practice clear, to facilitate a specific and accurate portrayal of the findings that I will lay out in the chapters to come.

My use of the term PD includes formal efforts to improve the teaching practices of employed, active teachers. These efforts include a broad range of activities, such as professional seminars, workshops, expert or peer classroom observation, and consultation (Chism, Holley, & Harris, 2012; Stes et al., 2010). These activities can address different goals in the attempt to improve student learning. The curricula might aim to change teachers' beliefs, knowledge, or attitudes about how student learn, or they might try to invoke new teaching behaviors directly without addressing what teachers personally think and feel (Clark & Hollingsworth, 2002).

There are other terms that touch on the same ideas. "Instructional development," or "instructional design," describes how teachers analyze and design their learning environment (Gustafson & Branch, 1997; Hardré, 2005). At a low level, this might describe the process of building class materials. At a high level, it might describe an analysis of outcomes following a teacher's pedagogical decisions. "Educational development" can mean the broad category of programs and activities designed to improve teaching practices (Chism et al., 2012), but can also describe student progress (Zimmerman, 1995). "Faculty development" is a term that is specific to instructors only. It might describe efforts to

improve teaching, but it can also describe other academic activities supporting overall professional ambitions, such as learning to write grants (Gullatt & Weaver, 1997; Sorcinelli, 2007).

In this work I draw from the broad field of PD for teachers as I try to understand and apply some of the extensive knowledge that exists about how people learn to teach. I apply these ideas to the development of teaching assistants in higher education. For the purposes of this thesis, I almost exclusively focus on formal learning practices that instructors experience. Chatting with colleagues, reading elective literature, and other informal learning spaces are legitimate forms of development, but outside the scope of this document.

2.2.1 Student-Centered Teaching

Modern PD highlights "active learning" and "student-centered" practices as core tenets (B. E. Cox et al., 2011; Hill, Kim, & Lagueux, 2008). These approaches emphasize involving students in their own learning process. Teachers today are widely encouraged to use learner-centered techniques in order to better engage their students. Students speaking up in an active learning environment is one aspect of such an approach to teaching (Michael, 2006). This philosophy conceives of learning as a process of conceptual change (e.g., Guzzetti, Snyder, Glass, & Gamas, 1993; Ho, Watkins, & Kelly, 2001) rather than a simple memorization of facts (Land & Hannafin, 1996). Rote learning can be necessary at times. But is not sufficient for deep cognitive development and building connections between ideas.

2.2.2 Active Learning

Active learning offers many advantages. It works well in empirical and scientific domains, moving beyond rote memorization and instead engaging students in discussion, argumentation, and deeper understanding (Land & Hannafin, 1996; Osborne, Simon, Christodoulou, Howell-Richardson, & Richardson, 2013). STEM courses in higher education have shown substantial improvements through the use of active practices over the past 20 years. College students in STEM who experience some level of active learning enjoy an average increase of .47 standard deviations on their final grades and are 1.5 times less likely to fail (Freeman et al., 2014). Active learning includes a wide variety of practices. For example, one subtle adaptation to instruction includes taking short breaks during lecture where students spend 2 minutes discussing the material (McConnell, 1996). A more dramatic example of active learning involves using a "flipped" classroom, where students watch video lectures and read texts on their own time and use class for active discussions and questioning (Lowell Bishop & Verleger, 2013).

Formalizing the practice slightly, the ICAP Framework (Chi & Wylie, 2014) describes four different levels of "activeness" that learners express. In declining order of cognitive engagement, these levels are "interactive," "constructive," "active," and "passive." At a high level, these terms are defined as follows:

- *Interactive:* Dialogic engagement with peers that involves argumentation and comprehensive questioning.
- *Constructive:* Self-explanation and verbalized reflection in an effort to synthesize information.
- *Active:* Repetition, rehearsal, and manipulation of information without any effort to engage deeper comprehension.
- *Passive:* Listening, reading, or watching without doing anything else.

This framework is a useful tool for categorizing student activities. It correlates to both cognitive levels of engagement as well as learning outcomes. It is relatively granular, however, in its use of terms. The majority of authors who describe "active learning," whether from practical, theoretical, or experimental perspectives, are referring to a combination of what Chi & Wylie call interactive and constructive practices (Freeman et al., 2014). The difference—which separates out different cognitive processes—is not relevant in every classroom or every study.

This is the case in my work, as well. The research questions in this thesis do not intend to discriminate between different cognitive processes. Instead, I describe my research in assisting instructors to learn about and experiment with instructional activities that they may find helpful for increasing student participation during class. The outcomes of these instructional activities may challenge TAs' assumptions about how to teach, and if successful, ought to improve their students' engagement and learning. For this work, my use of the term will align with both interactive and constructive activities, following the more common and inclusive definition of "active learning" that is widely used in the literature.

2.2.3 Discursive Teaching and Deep Questioning

Within the domains of active learning and student-centered principles is a set of practices which I call "discursive teaching." These are tactics that instructors use to foster an effective classroom environment by generating student talk. There are many ways to engage students in dialogue. At a high level, students should feel comfortable speaking in class because they know that they are respected as people and that their contributions will be heard (Ambrose, Bridges, DiPietro, Lovett, & Norman, 2010; Bain, 2004; Lemov, 2010). In my work I focus on building up the practical skills that novice college instructors need to use in order to build such an environment (cf. Fritschner, 2000; McConnell, 1996; McShannon et al., 2006; Nunn, 1996; Rocca, 2010; Weaver & Qi, 2005). These include:

- *Asking deep questions during class*
- *Waiting at least three seconds after asking a question and after students respond before speaking again*
- *Calling directly on students to answer questions after sufficient wait time has passed*
- *Praising students for their contributions*
- *Using students' names at any time during class*
- *Asking follow-up questions, e.g., how students discovered an answer*

Perhaps the tactic which is most challenging for novices to implement is the use of deep questions. These are questions which challenge students to build on basic knowledge by engaging in deep explanation (Pashler et al., 2007). They encourage interpretation, analysis, and elaboration. They can help students construct new knowledge as they analyze and synthesize information, build on their assumptions, and generate new ideas (Bolen, 2009; Chen, Clarke, & Resnick, 2014; K. Ellis, 1993).

The difficulty involved in using these questions may be due to how infrequently they are typically used in American education environments. Observational studies show that as many as 80% of questions professors ask target the lowest cognitive levels (Nunn, 1996). This questioning behavior encourages rote memorization and undermines the exploration of underlying concepts (Kreber, 2005). Teaching this way continues a trend of shallow questioning that students encounter all throughout K-12 schooling (Galloway & Mickelson, 1973; Oliveira, 2010; Sahin, 2007). If students only hear shallow questions, they may have trouble asking anything else when they have their turn to teach.

Another tactic that can enhance student learning and build on deep questioning is to create an entire sequence of questions rather than simply delivering a positive evaluation and moving on (Hellermann, 2002). Follow-up questions all students to show that their understanding is replicable, and it keeps them moving forward into new domains of understanding (Lemov, 2010). Concrete examples include asking students how they came to know something, why something is the way it is, asking for evidence.

The evidence that deep questions produce the intended result emerges through active student participation (Oliveira, 2010), i.e., students attempting to explain what they know or explore what they are learning. Deep questions can help students elaborate on what they are still trying to understand (Bolen, 2009; K. Ellis, 1993; Galloway & Mickelson, 1973; Oliveira, 2010). As students respond to these prompts and participate in class discussion, they exercise critical thinking and improve fact retention (Weaver & Qi, 2005). Even quiet students can benefit when other students answer questions, because talkative students are more likely to ask follow-up questions that are relevant to everyone (Howard, Short, & Clark, 1996).

2.2.4 A simple example: *wait time*

One example of a simple yet effective strategy that teachers can lean on to build a discursive teaching environment in their classroom. From the 1970s to the 1990s, a large number of studies examined the effect of waiting for students to answer teachers' questions in K-12 classrooms (McDaniel, 1985; Rowe, 1972, 1980; Swift & Gooding, 1983; Tobin, 1987). The research spanned many grades, courses, and environments. There were descriptive and experimental studies.

The findings were robust and declared that most teachers waited less than 1 second after asking students a question before they either repeated the question, rephrased it, or simply answered it themselves. Some of the benefits that either correlated with or resulted from increasing wait time (depending on the study) included less teacher talk, more student talk, more student-student interaction, more deep questions from the teachers, and more thoughtful and longer responses from the students. Research in higher education has shown that besides asking mostly low-level questions, the amount of time that professors wait is generally below three seconds, and that most questions go unanswered (Larson & Lovelace, 2013).

2.3 What we know about teaching teachers

Given all of the above, we know that best practices exist in teaching with demonstrated effects. Despite decades of research showing the benefits of active, student-centered teaching practices, college instructors are slow to shift away from teacher-centered practices. "Information transmission," i.e., top-down, expert-driven lecture, continues to dominate among most of higher education STEM. Higher education instructors believe their job is to deliver facts. Time constraints and lack of career incentives can explain why teachers in higher education do not seek out help. This section operationalizes the previous description of PD for teachers and investigates the methods by which instructors improve. The goal is to clarify what is known about how people learn to teach so that I can apply it to the research which follows.

2.3.1 Feedback on teaching

Like mastering any new skill, learning to teacher in higher education requires feedback (Gormally, Evans, & Brickman, 2014). Feedback is critical for changing instructors' knowledge about teaching practices and attitudes about student learning. Feedback for faculty and TAs is more effective when it contains accurate data and irrefutable evidence, concrete information, specific data, and when it is focused on specific topics, such as lecturing/discussion or wait time (Brinko, 1993). Successful consultation in learning to teach focuses on describing specific behaviors rather than on evaluating the individual. Instructors should receive feedback frequently enough to support ongoing self-assessment, and soon enough that the practitioner can recall the relevant practice opportunity (Ilgen, Fisher, & Taylor, 1979). Feedback is more effective when it is part of a learning process, not a one-time quick fix. An empirical study of semester-long PD with professors showed that in-class observations and weekly meetings to review strategies can reliably lead to belief and behavior change (McShannon et al., 2006).

2.3.2 Teacher reflection

An interpretation of Chism's model of teaching growth (Nyquist & Wulff, 1996) explains that graduate students learn to overcome teaching challenges in the college classroom by noticing what happens during class, experimenting with new ideas, gathering data about what the new actions have produced, and reflecting deeply on those data. This model draws a line from feedback (data) to reflection that helps explain a possible mechanism for improved teaching. Instructors in higher education have also shown the ability to produce feedback for themselves through self-reflection (Gormally et al., 2014) that can lead to modified teaching behavior.

There are different types of reflection and different reflective processes. Schön famously described *reflection-in-action* as the skill mastery that allows an expert to improvise within a changing set of familiar constraints (Schön, 1983). This is decision-making that relies on deeply encoded schemata that are difficult for experts to explain. *Reflection-on-action*, by contrast, is a qualitatively different experience that incorporates declarative rather than tacit processes. The subject reviews their actions, generates insights, and perhaps constructs a new plan. Boud describes "reflection after events" as a method to enhance learning through three phases: returning to experience, attending to feelings, and reevaluating experience (Boud, 2001). These phases roughly translate to recalling what happened, managing resulting feelings that could inhibit or promote learning, and determining the value of the lived experience in light of overall goals.

Effective reflection can produce change in terms of the underlying values instructors hold regarding learning (Hubball, Collins, & Pratt, 2005), and even help transition instructors from teacher-centered attitudes into student-centered beliefs (Ho et al., 2001).

Research with pre-service teachers has shown that through reflection people can examine their lived experience as former students in a new light of being an instructor (Boud, 2001; Boyd, Gorham, Justice, & Anderson, 2013). This process of re-evaluation is an important step in gaining an identity as a teacher.

Reflection supports metacognition about how one teaches, which in turn supports teacher self-regulation (Hartman, 2001; Prytula, 2012). It assists instructors in learning to notice helpful and unhelpful behaviors that might normally bypass awareness (Sherin & Elizabeth, 2005). This is a critical step in developing pedagogical content knowledge, i.e., knowing not just of the content

students need to learn, but how students learn given their specific contexts, purposes, and overall curriculum (Ball, 2008).

Ways that researchers have studied teachers' reflection include showing them videos of their teaching under the guidance of an expert (Sherin & Elizabeth, 2005), asking them to write in journals (Hatton & Smith, 1995), and prompting them to share anecdotes with colleagues within a formal training context (Henderson, Beach, & Finkelstein, 2011). These methods have shown that teachers can develop beliefs about themselves as capable teachers, and their students as capable learners, through repeated practice opportunities, feedback, and the experience of successes. This is a (simplified) example of teachers developing self-efficacy.

2.3.3 Teacher self-efficacy

Self-efficacy for teaching is an important component of how practitioners learn to teach. For example, it can influence both the willingness to try new teaching tactics as well as awareness of ongoing awareness teaching behaviors (Bandura, 1997; Prieto & Altmaier, 1994; Tschannen-Moran & Hoy, 2001). As a general theory, self-efficacy derives partly from a homeostatic model of self-regulation and motivation that promotes a cycle of action and reflection (Bandura, 1997). Albert Bandura described how the process of acting, reflecting, adapting, and noticing successes builds the individual's self-efficacy for their given occupation. He described how self-efficacy for an activity comes over time as one not only increases content knowledge within a domain, but also sees evidence of success in expressing that knowledge through action. In other words, self-efficacy occurs as teachers see themselves acquire skills that produce student results.

The literature describes multiple pathways in acquiring skills that build toward self-efficacy. Some instructors, for example, need a change of perspective on how their students learn before they try to enact new (student-centered) teaching methods (Henderson et al., 2011; McShannon et al., 2006; Postareff, Lindblom-Ylänne, & Nevgi, 2007). Other instructors (often novices) enact new strategies as they receive supportive feedback from colleagues and advisors, regardless of whether they are originally convinced that those strategies are best for their own students (Tschannen-Moran & Hoy, 2007).

Without these external supports, novices can have difficulty perceiving their own deficits (Settlage, Southerland, Smith, & Ceglie, 2009). Novices are often unable to produce the metacognitive skills necessary to recognize their shortcoming, and as a result they can overestimate their skills (Kruger & Dunning, 1999). Without effective intervention, these perspectives can be difficult to change. Instructors in higher education, particularly within the natural sciences, commonly orient toward teacher-centeredness (Lindblom-Ylanne, Trigwell, Nevgi, & Ashwin, 2006; Postareff et al., 2007). Novices with this orientation can require as many as 30 training sessions before they show evidence of adopting student-centeredness attitudes (Postareff et al., 2007).

2.3.4 Consultations in teacher education

The type of training that instructors in higher education receive matters (Chism et al., 2012). Although there are many methods of academic instructional development, the current gold standard relies heavily on consultation following a live teaching observation, either from an expert or a knowledgeable peer (Bell & Mladenovic, 2008; McShannon et al., 2006; Stes et al., 2010). Consultations can have a powerful impact on multiple measures of teaching excellence, from faculty change to increases in student ratings (Finelli et al., 2008). These observations generally follow an

empirical protocol that helps the experienced observer construct personally meaningful feedback for each instructor (Akiha et al., 2018). This feedback typically comes through a face-to-face meeting following the teaching event where the expert guides the instructor through a thoughtful reflection of their data.

In practice, it is difficult to meet the standard of teacher observations and guided reflection. In a classroom, each observation can only target a single session, and each follow-up requires scheduling valuable time. It is impractical for even the most self-motivated instructor to request more than one or two observations per semester. We now turn to the final section of this chapter, where we consider technology in the classroom and the role it might play in changing this status quo.

2.4 The incoming wave of monitoring and sensors in education

2.4.1 Smart classrooms

Technology is rapidly changing the landscape of American education. New tools for augmenting learning spaces arrive faster than teachers can learn to use the old ones. As a practical field, technology-enhanced learning has much to offer. As a research field, it still has much to explain (Goodyear & Retalis, 2010). Instrumented, or "smart," classrooms incorporate interactive technologies that expand the borders of traditional formal learning activities (Bautista & Borges, 2013). For example, tele-presence brings students from different locations into synchronous contact (Shi et al., 2003). Interactive whiteboards and immersive displays offer teachers opportunities to enhance student engagement through spontaneous and rich media (Blau, 2011). Clickers help instructors involve all of the students, especially in large college classes (Caldwell, 2007; Martyn, 2007). Many of today's technical artifacts are visible to students and teachers. However, research is expanding to incorporate ubiquitous, invisible technologies that automatically gather data about the learning space.

Much as mobile technologies have increased access to data about mobile users, the technologies used to enhance classrooms are opening opportunities to gather large datasets about class instruction. Most of the applied research on sensor-based data collection in classrooms has focused on building models of student behavior. These data track student behavior to assist with orchestration of activities (Dillenbourg & Jermann, 2010; Shi et al., 2003), or to provide notification of where students are in their work (Bakker, van den Hoven, & Eggen, 2014). Cutting edge systems are in development that build models of student engagement, or instructor cognition. Microphones gathering audio signals can automatically detect stages of teacher-student interaction based solely on when different actors speak (Anguera et al., 2012; Zhu, Barras, Lamel, & Gauvain, 2008). Automatic speech recognition and machine learning help to understand classroom speech patterns and automatically catalogue teachers' questions (Blanchard et al., 2016; Chen et al., 2014; D'Mello, Picard, & Graesser, 2007).

The field of Learning Analytics reveals some of the productive ways researchers are able to leverage data in order to support teachers and students (Clow, 2012). By building a model of each student's knowledge and skills, systems can give feedback to teachers about their students, or give feedback to students directly (Greller & Drachsler, 2012; Verbert, Duval, Klerkx, Govaerts, & Santos, 2013). Direct feedback to students can help them to practice self-regulated learning and metacognitive development (R. S. J. d. Baker & Siemens, 2014). Dashboards for instructors can provide rich data about the states of the students, helping them decide how to adapt their teaching (Holstein, McLaren, & Aleven, 2017).

Similar to the work produced in the learning analytics community, this thesis investigates the use of models and graphical representations of surface behaviors that are difficult to observe. Additionally, I describe how researcher can use these technologies to support changes in behavior and belief by scaffolding repeated practice and reflection. I also draw on research investigating the use of sensors in classrooms. These instruments offer new possibilities for understanding how people learn to teach in real-world learning spaces.

Instructional spaces that use myriad sensing and feedback technologies, with the intent of improving education, are the education industry's closest parallel to the relatively new area of Personal Informatics (PI), described next. There is some theoretical overlap between traditional PD and PI; there are also unique features to each. These divergent features could potentially complement each other when combined into a new socio-technical training system for college instructors. But before thinking about the combination, it is worth exploring what is known about PI outside of education.

2.4.2 Personal Informatics

PI, also called Quantified Self, is an area of research on socio-technical systems that help people build awareness of their invisible behaviors (Choe, Lee, Lee, Pratt, & Kientz, 2014) and set goals (Consolvo, Klasnja, McDonald, & Landay, 2009). PI first gathers data about what a user does within a specific class of activities, e.g., health and exercise (Consolvo et al., 2008; Kay et al., 2012; Lin, Mamykina, Lindtner, Delajoux, & Strub, 2006), sustainability (Comber, Thieme, & Rafiev, 2013), or finance (Rapp & Cena, 2016). Systems then give the user data visualizations that inspire reflection and goal-setting (Li, Dey, & Forlizzi, 2010). A common persuasive feature of PI systems is the use of achievement badges when users reach specific goals (Fritz, Huang, Murphy, & Zimmermann, 2014). Even users who abandon their data-gathering PI practices can show long-term changes in behavior as a result of their experience (Epstein et al., 2016).

The Stage-Based Model of Informatics is the most commonly cited explanation of how PI works (Li et al., 2010). The five stages are Preparation, Collection, Integration, Reflection, and Action.

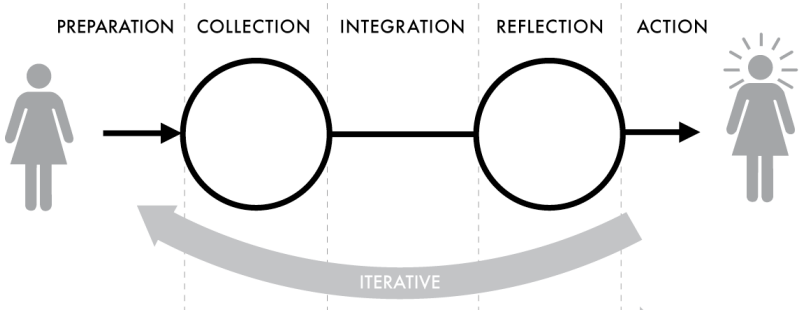


Figure 2.1: The Stage-Based Model of Personal Informatics Systems (Li, Dey, & Forlizzi, 2010).

In the Li et al. model (Li et al., 2010), the user ("Quantified Selfer") manages each of the stages by preparing to gather data, gathering data, cleaning and making sense of the data, and reflecting on the data. Finally the user takes action as a result of gaining new insights on her own behavior. When uninterrupted by barriers, this process continues indefinitely. Many of the original members of this Quantified Self movement were involved in gathering data from many different sources about many different aspects of their lives. Mastery in one domain did not typically bring an end to the overall practice of quantification.

The Quantified Self community remained relatively isolated for many years because the means of gathering and managing data was difficult. Challenges such as tracking, storing, and visualizing data were automated as mobile technology, wearable computers, and cloud-based architectures became ubiquitous. GPS receivers, gyroscopes, and accelerometers became smaller and more accurate. Mobile devices became robust enough to generate, store, or offload large amounts of data. Advances in app design and development created a marketplace for data retrieval. Personal Informatics was able to focus on the most challenging aspect of this research arc: helping people change.

2.4.3 PI and Change

Digging deeper into the mechanics of PI, different researchers have applied different theories for how and why users might change through its use. The Trans-Theoretical Model of Behavior Change (He, Greenberg, & Huang, 2010; Prochaska & Velicer, 1997) and Social Cognitive Theory (Andrew, Borriello, & Fogarty, 2011; Froehlich, 2011; Kamal, Fels, & Blackstock, 2011) have been two popular theories from the social sciences. In both cases, PI researchers have promoted the idea that these theories of change would help users develop increased self-efficacy (Kamal et al., 2011), i.e., a belief in the ability to enact specific behaviors and ultimately achieve specific goals (Bandura, 1997).

In practice, PI has struggled to connect self-efficacy to changes in behavior. Despite using it as an explanatory mechanism of change throughout the literature, actual research protocols do not typically operationalize the idea, nor test for it. This may explain the fact that although much of PI has been able to increase users' awareness of actual behavior, it has mostly failed to produce broad changes in behavior across populations (Clawson, Pater, Miller, Mynatt, & Mamykina, 2015; Rapp & Cena, 2014). Despite widespread adoption of tracking technologies, there also exists widespread abandonment (Lazar, Koehler, Tanenbaum, & Nguyen, 2015). Researchers are beginning to recognize that PI has often focused on data tracking as its own goal, rather than as a step toward achieving some other goal, such as behavior change or perhaps even changes in self-efficacy (Lazar et al., 2015; Rapp & Cena, 2016).

On the other hand, more recent research suggests that some of the impacts of PI have simply been slow to emerge. Changes can take place outside of the constraints of a given study, and even persist beyond tracking abandonment (Epstein et al., 2016).

2.4.4 PI and PD

PI offers valuable insights for how to design data capture and visualization. It shows methods for supporting reflection and goal setting. Additional findings from the PI literature, described above, highlights important high-level insights in how to use data and reflection in attempting to change teaching behaviors:

- Feedback alone may not be sufficient
- Users may need to think about their motivations, not just their behaviors
- Longitudinal observations might reveal slow changes

An important consideration in translating PI to teacher education is that PI research has historically had the advantage of guiding people to reflect on behaviors they are likely to understand from the outset. Saving money, walking more, eating healthier, and recycling more often are all common themes in PI research. These topics are unlikely to challenge users' pre-conceived opinions, unless they are complete novices to a given topic (Rapp & Cena, 2016). When it comes to supporting teachers through

data reflection, this new population of users might encounter concepts about learning and teaching that they have never heard before. Alternatively, they may be asked to adopt controversial ideas about which they already have negative opinions. Providing support well beyond data abstractions may prove to be an important design consideration.

Fortunately, decades of work in the learning sciences provide some guidance regarding how people learn. The literature on self-efficacy for learners (and teachers) has had more time to mature than the literature on self-efficacy for data-trackers. At the same time, classroom sensors and PI offer advantages in gathering rich, complex data about teaching and learning behavior in the wild rather than the laboratory.

To date, most of the focus and change driven by smart classrooms has increased awareness of what students are doing with their class time. Increasing teacher awareness of students' states is certainly useful for expert instructors. But this is a limited view that may not be as valuable to novices. However, these same technologies could also gather information about what the instructor is doing. If positioned carefully as a tool for teacher reflection and professional development, there are exciting new possibilities for supporting teachers to gain not only insights into their students, but insights into themselves as instructors. In this thesis, I begin that work.

I advance this research by offering a new application for this technology. Beyond capturing *student* behavior, I capture what *teachers* are doing. I reframe the teacher as the learner. Rather than developing radically new sensing technologies, I limit myself to the capabilities that this work and related work on sensing has shown to be currently viable. This means that the findings can be realized in practice in the near-term, without relying on advances in sensing technologies. This work provides technology developers with a blueprint for helping teachers learn about themselves and their teaching, using classroom instrumentation as a useful source of input that goes far beyond the logistic limits of classroom observations that define today's "gold standard" of PD. In the following section I describe the methods I use to develop this work.

2.5 Research Methods

This dissertation draws on disparate areas of research and asks novel questions about how to design future feedback systems for teachers. It is almost entirely exploratory in nature. As such, educational design-based research (DBR) is an appropriate methodological construct for the investigation.

2.5.1 Design-based research and education

DBR has a rich history in the study of technology-enhance learning environments (Wang & Hannafin, 2005). My work adds to the catalogue with a series of longitudinal investigations into highly variable, real-world learning environments. I chose this approach because of the impact potential it provides. DBR aims to implement and improve theories about education and learning while also impacting practice (Barab, 2014). It has a coherent framework for organizing multiple studies over several years. It indicates the scope and purpose of each study as researchers attempt to build novel solutions to real-world educational problems.

There are various names for this research approach; design experiments, design research, design-based implementation research, educational design research, and design-based research (Barab, 2014; Brown, 1992; Collins, Joseph, & Bielaczyc, 2004; Hoadley, 2002; McKenney & Reeves, 2012; Penuel, Fishman, Haugan Cheng, & Sabelli, 2011; van den Akker, Gravemeijer, McKenney, & Nieveen, 2006;

Wang & Hannafin, 2005). The overall emphasis is that researchers draw from existing research, build practical and innovative solutions, and contribute back to educational theory (McKenney & Reeves, 2012). Solution building involves iterative cycles of design and testing with increasingly sophisticated interventions and larger samples of a research population.

There are multiple generic models of iterative design, testing, and theory building through DBR (Kennedy-Clark, 2013). One that provides very clear descriptions of how to enact each phase of research is Wademan's generic research model (Figure 2.2; Wademan, 2005). The phases for this model are *problem identification*, *preliminary investigation of design principles*, *tentative products and theories*, *prototyping and assessment of preliminary products and theories*, and *problem resolution and advancing theory*. For my work, I have adapted Wademan's model to visualize my particular research process (McKenney, 2001; Wademan, 2005).

These research stages flow through general phases of needs analysis, design, development, and evaluation. Along the way, researchers identify and examine multiple interacting variables to produce "system-level understanding." The goal is to expose educational phenomena in contextually rich environments, and draw forth theoretical implications. Mechanisms emerge by uncovering the challenges and opportunities that surface during classroom research (Barab, 2014). Over years of investigation, DBR researchers attempt to describe what happens, why it happens, and eventually discover if it is systematic (Shavelson, Phillips, Towne, & Feuer, 2003).

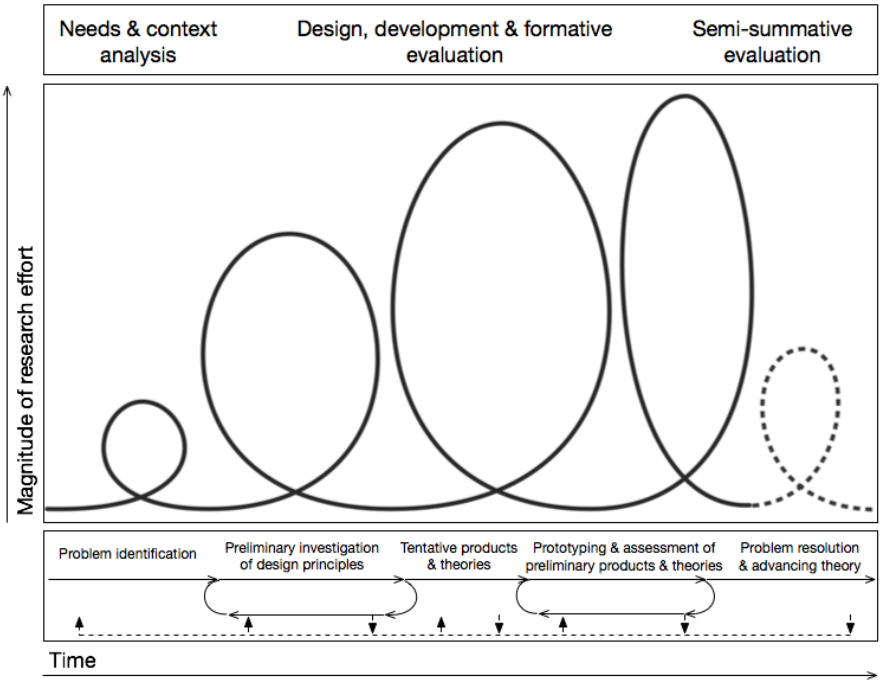


Figure 2.2: An adapted "generic development research method" (McKenney, 2001; Wademan, 2005)

The research in this document follows the general phases of DBR as outlined by the adapted generic development research method. The first study (Chapter 3: Field observations and problem identification) identifies problems in the given context and proposes possible solution strategies. The second study (Chapter 4: Personal question trainer) is a preliminary investigation to uncover important design principles and to test tentative prototypes and theories. Chapter 5 introduces a tentative

practical/theoretical framework for the remainder of the research. The third study (Chapter 6: Including data visualizations) continues prototype development and refines the solution space. The fourth and final study (Chapter 7: Development and formative evaluation of *ClassInsight*) investigates the possible theoretical contributions of the research in terms of system development as well as knowledge about the participants.

To be considered valid, DBR requires evidence-based claims about learning. To be meaningful, it must advance theoretical knowledge of the field. The iterative and nested cycles of design, development, and testing in DBR are amenable to many research methods, whether they are qualitative or controlled and randomized (Easterday, Lewis, & Gerber, 2014). The dotted line at the end of the cycles in Figure 2.2 indicates a partial resolution in that this document does not "resolve" the problem per se, but it does advance theory.

2.5.2 Intended outcomes of this work

Throughout the interventions I design in this thesis, I use research on teacher reflection to guide my designs. Specifically, through this design research, I build a foundation for a series of goals in educational research:

- *Providing teachers with rapid feedback on their actions through newly available data from the thoughtful use of sensors in smart classrooms*
- *Scaffolding grounded reflection-on-action, driven by smart classroom data, by prompting the instructor to elaborate on the events of the day*
- *Encouraging instructors to build self-efficacy, view their own actions as having an impact on their classes, and choose to improve their teaching*
- *Implementing active learning in education research itself, by including the voices of participants throughout the interventions that I, and others, build.*

My research addresses the state of teaching in higher education by designing an intervention that (a) highlights the value of student-centered teaching, and (b) could scale to a wide audience, implement learning through algorithms, and diminish the need for expert oversight. I aim to borrow from the PI literature while expanding that research realm; implementing additional goals and operationalizations than what is typically included in instruments from that literature. Furthermore, I explore the nuance of designing for change when some elements of the instruction may run counter to the user's prior beliefs, such as when instructors hold teacher-centered orientations.

I do not see traditional methods of instructional development as problems to overcome. For the foreseeable future they will be likely to outperform the outcomes of any algorithmic or heuristic system. But there is a strong possibility that an embedded system, situated within the learning context, that provides instructional support to novices would be a meaningful complement to the existing avenues of PD for higher education instructors, with availability that far exceeds anything seen in the status quo. As technologists, we are building educational systems in a time of rapid change. It is imperative that we understand the tools that are being added into our classrooms, and design for those tools with student well-being and student learning at the heart of the work.

Chapter 3: Field observations and problem identification (Study 1)

Researchers have begun investigating what can and should be "sensed" in a smart classroom. Current work investigates both what information can be inferred, and what information might be valuable. It is not yet clear what these technologies might produce, what impact they may have on classrooms, nor how they should be used to improve teaching and learning. For context, note that the primary goal of sensors in any domain is to gather large arrays of data embedded within a lived context. This is similar to the activity monitors found in fitness devices, to other systems used in support of personal informatics and quantified self, and to internet of things (IoT) sensors being designed for a variety of contexts. or the many data collection systems in modern smart homes.

This is very different from the majority of research in controlled studies, where information about users is gathered based on a very limited amount of data about the participants and their actions. In the absence of smart classroom sensors, typical classroom research places human observers, and often cameras and microphones, in the classroom. The work produces qualitative field notes in conjunction with quantitative counts of some specific variables of interest in service of a research question. These counts are occasionally gathered in the field, but more often produced through a methodical review of video, audio, or transcript data. This slow process of data production creates a highly valid and useful dataset that can answer questions that are scientific in nature. However, it is limited by the small number of classes and classrooms that can be realistically observed.

My goals for smart classroom sensors would produce data in a very different way: my research explores questions about a possible future technology. One of the most important features of big data in the wild is that it generates immediate models of what is happening in a given context. A heart-rate monitor must produce immediate samples of data. It would not be helpful if it required someone to manually produce the samples, long after the data was originally collected. Luckily, sensitive and accurate sensors already exist to produce heart rates. In more multimodal environments like classrooms, cameras, depth sensors, gyroscopes, and microphones make up a large portion of sensing technologies. Automatic speech recognition, text segmentation, computer vision, and machine learning are tools that overlay these sensors to produce potentially meaningful relationships between many low-level variables. But these do not produce data that can be measured as easily as a heart rate; instead, they produce a messy, interlocking set of signals all happening in parallel. So the first question to address in this first study on potential uses of smart classrooms is this: how do we gather data and deliver feedback to instructors?

I chose a few targeted questions about student-centered teaching for this study. As noted in Chapter 2, the population of novice TAs in higher education is likely to rely on lecture and struggle to develop meaningful, rich interactions with their students (Borg, 2004; Brownell & Tanner, 2012; Luft et al., 2004; O'Neal et al., 2007). For this reason, technologies that detect speech acts and automatically detect meaningful speech patterns would be likely to produce data that would be useful in training this population. Traditional PD consultants for this population already produce qualitative data about whether or not the TA is engaging their students (Penny & Coe, 2004). Would a system that could support these observations with concrete data about classroom interactions be a useful contribution to the field?

With this research, rather than developing a framework for sensor technology in the abstract, I aim to pair the technological work with open questions in the research literature on professional development. By doing so, my hope is that both sides of the research question—sensors and learning—can be improved and made more robust. This work included reviewing the literature on what kinds of sensors are in development, piloting rough sensors of my own, and considering what types of data might be used in actual classrooms in the near future. Likewise, I wanted to gather data that would highlight areas of concern for the given population in ways that participants would understand.

This study is the first phase of DBR: Problem identification/Needs & context analysis (McKenney, 2001). As a low-cost investigation, the study implements data collection techniques that allow a human observer to immediately produce computer readable data about classroom interactions. The observational protocol mixes quantitative counts of live behaviors with qualitative fields for commenting on those events. The goal is to create "good enough" data capture to simulate a future smart classroom. With these data, it is possible to show TAs visualizations of their teaching patterns and outcomes to see how they respond. Following observations and visualizations, the study moves into a probing interview with each participant to explore what they see in their data and their teaching.

The motivating factors for this study, based on the ideas above and the related work in Chapter 2, generate the following research questions:

1. *Questions based on Professional Development for teachers*
    a. Are TAs aware of what they do that is effective for learning, and what is not?
        i. Do TAs perform actions that support student learning?
        ii. What pedagogical skills do TAs exhibit? Are they relevant for their contexts?
        iii. How do they feel/think about discursive teaching strategies?
        iv. Do TAs believe in their ability to enact appropriate strategies?
        v. Do TAs exhibit student-centeredness or teacher-centeredness?
2. *Questions based on Personal Informatics and Smart Classrooms:*
    a. Of the types of data that current and coming technologies can provide, which would be relevant for this population?
        i. How does this population respond to data delivered following the stage-based model of PI?

3.1 Method

The protocol for this exploratory study focused on attending classes for a sample of TAs in a computer science course and live-tracking a high fidelity, fast record of important discursive acts. These speech acts included the length, frequency, and source of spoken turns, the length of silence (non-speech acts), the presence of TA questions and whether those questions were based on class content or some other topic, and the timing/order of all of these speech acts.

3.1.1 Participants

Five teaching assistants from the Pittsburgh campus of Carnegie Mellon's School of Computer Science participated in the study. Each TA was male, Indian, and in his twenties. They were all graduate students leading recitations (i.e., small class sections typically used as a supplement to the primary lecture) in a mid-level, undergraduate computer systems class. None had taught before, and prior training had been limited to practical teaching matters, such as delivering information or grading assignments.

3.1.2 Observations

I attended weekly class sessions for each TA for 2 months of the fall 2014 semester. This class introduces (mostly) sophomore computer science students to elementary topics regarding machine language, compilers, and code optimization. The kinds of activities that commonly occur in recitations range from TAs performing worked examples and lecturing to students working on problems on their own. The recitations were small (4 to 23 students each). Each lasted for 50 minutes. I observed TAs for 6 – 7 sessions, totally about 30 hours of observation.

3.1.3 Data

Data collection included classroom observations once per week, occasional interactions with data visualizations (in paper prototype formats), and final interviews with the subjects. Data sources included audio recordings of each class, field notes and classroom observations, contemporaneous notes following post-class interactions, and interview transcripts. I generated the qualitative and quantitative field notes through a computer system called Look Who's Talking (Chen et al., 2014). This research tool/note-taking software was designed to give observers a way to quickly catalog spoken transactions in a classroom. The interface (Figure 3.1) has a seating chart, class log, metadata editor, and app functions. The blue boxes in the seating chart represent each desk in the room. Those that are labeled represent students in the class. TA events are either T1 (direct lecture) or T-Demo (on-screen coding samples). The observer's position was logged in the pink box. The rest of the boxes represent when "Multiple Students" spoke at the same time, any periods of "Silence," and other "Event" types that were worth noting but did not fit any other predefined speech act.

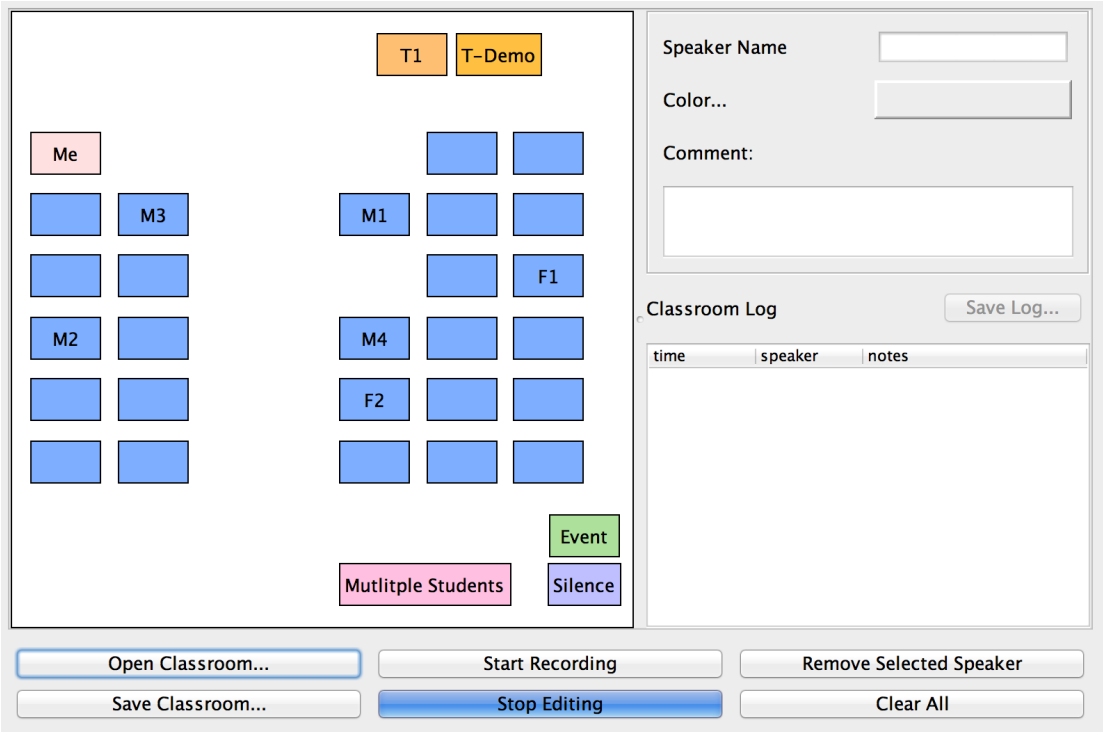


Figure 3.1: Screenshot of a typical Look Who's Talking session.

When an observer clicks on any of the boxes in the seating chart, the program generates a line in the Classroom Log. The line includes the amount of elapsed time since the beginning of the class ("Start

Recording"), the text within the selected box, and any qualitative notes the observer chooses to type. Table 3.1 shows a sample table generated on October 27, 2014 in one of the TA's recitations. Note the inclusion of a Duration column that produces a length of each event.

Table 3.1: Transcript excerpt of a classroom observation.

Timestamp	Speaker	Duration	Notes
22:21.2	Silence	00:04.6	"Any questions?"
22:25.9	T1	02:37.5	
25:03.4	Silence	00:02.0	"Any questions on this?" I should really make a button for this on this guy.
25:05.3	T1	00:54.6	
25:59.9	Silence	00:03.6	Content question.
26:03.5	T1	00:18.9	
26:22.4	Silence	00:02.1	Here he asks a content question. The following student answers. (It's only one word)
26:24.5	F1	00:01.8	
26:26.3	T1	00:11.0	Here he asks another content question

The categories of questions TAs asked were classified as Closed-Ended ("Is it physical memory or virtual memory, or is it both?"), Open-Ended ("Can anyone tell my why this is the answer?"), Any Questions ("Are there any questions?"), and Administrative questions ("Has everyone finished their homework?").

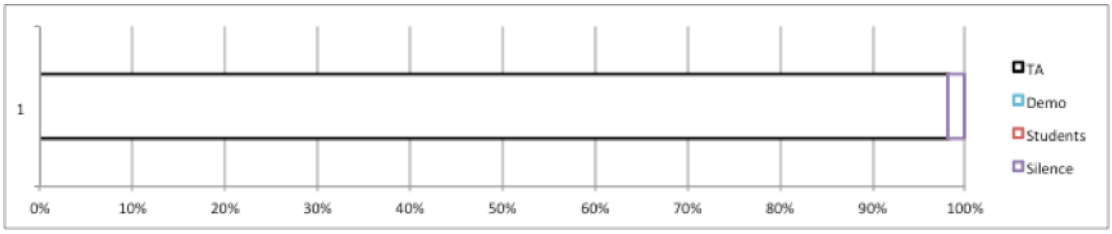


Figure 3.2: The percentage of TA talk, Student talk, and Silence for TA1 during class on October 27, 2014. Note that there was no demo that day, and that there was so little student talk that the sliver of red is imperceptible.

During one-on-one interviews with each TA at the end of the semester, I showed a larger set of data visualizations. These included representations such as how many students talked, the length of different students' speech turns, and timelines of events from each class. They also saw representations of multiple classes in a single graph. I included graphs that were intentionally difficult to read in order to test if participants would be honest about poorly designed visualizations.

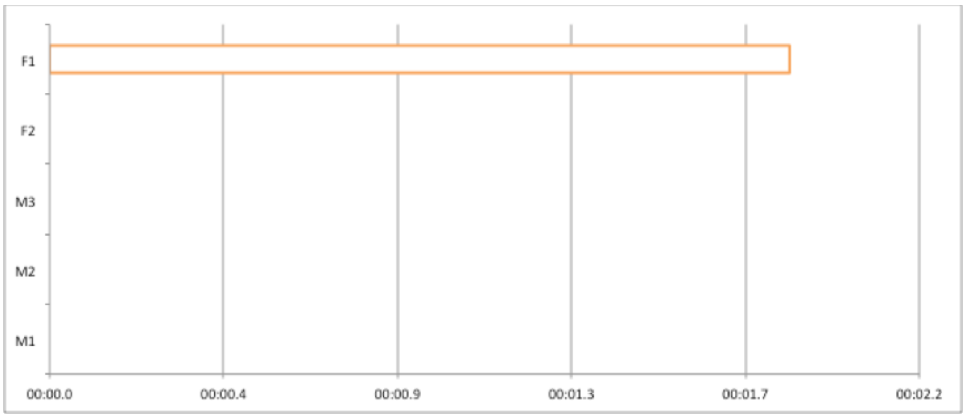
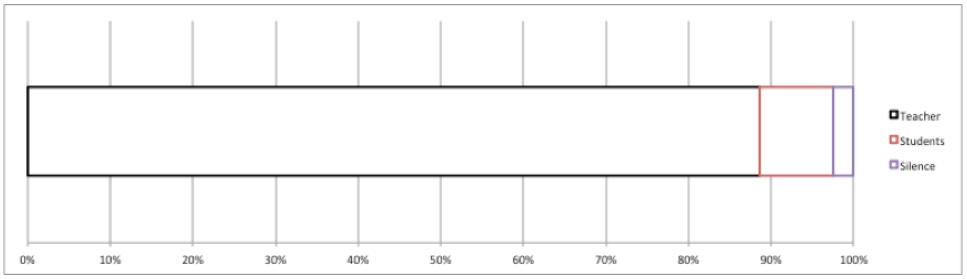
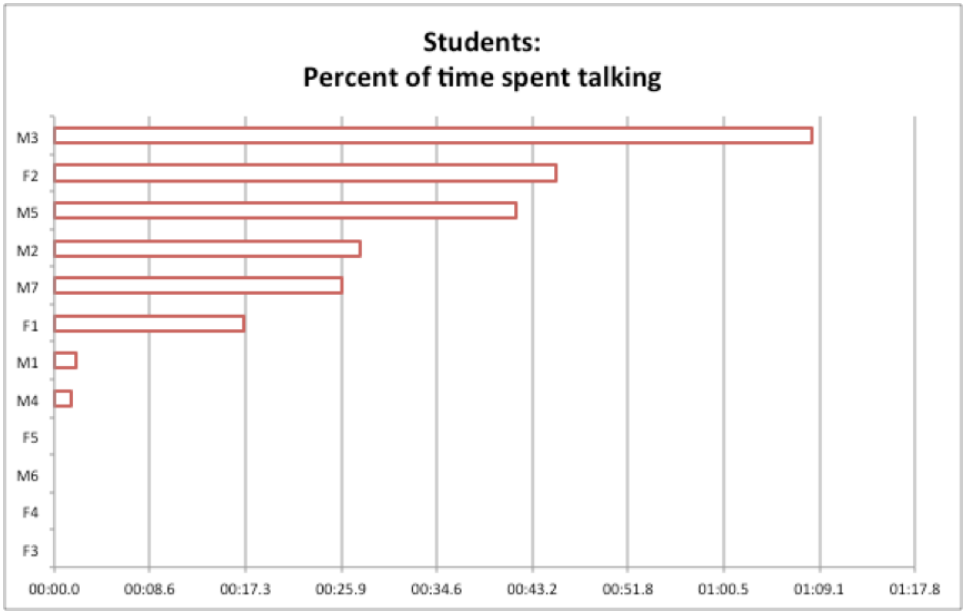


Figure 3.3: TA1's student participation graph from October 27, 2014.





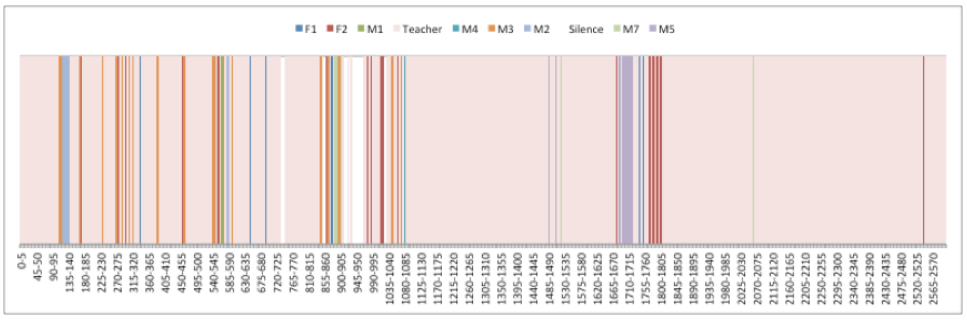


Figure 3.4: Ratio, student breakdown, and timeline visualizations for TA2 on Sept 9, 2014.

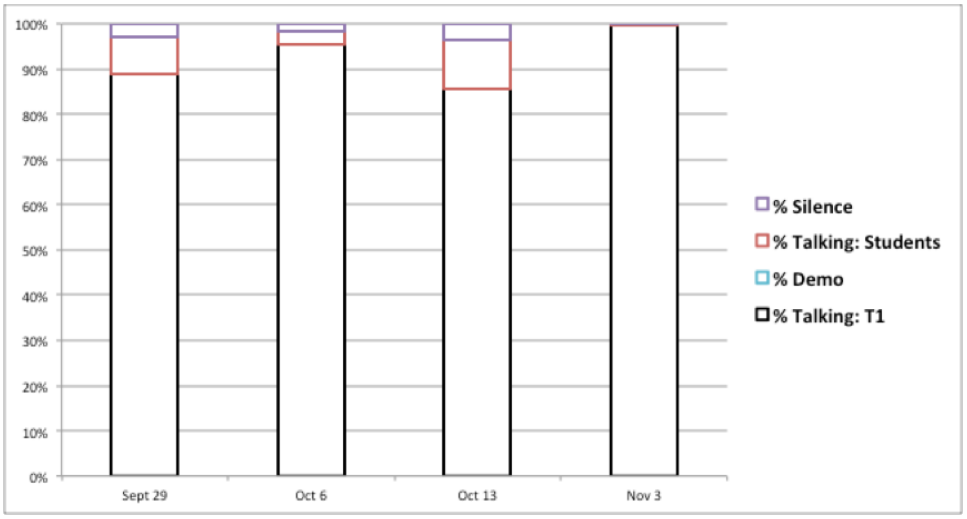


Figure 3.5: A graph of multiple days taught by TA2.

In addition to the ratio graphs, I included experimental visualizations during each interview. The goals of these visualizations included probing TAs on specific aspects of their teaching behaviors, testing for surprising reactions to the data, and to test whether TAs would identify that graphs were hard to read (Figure 3.6).

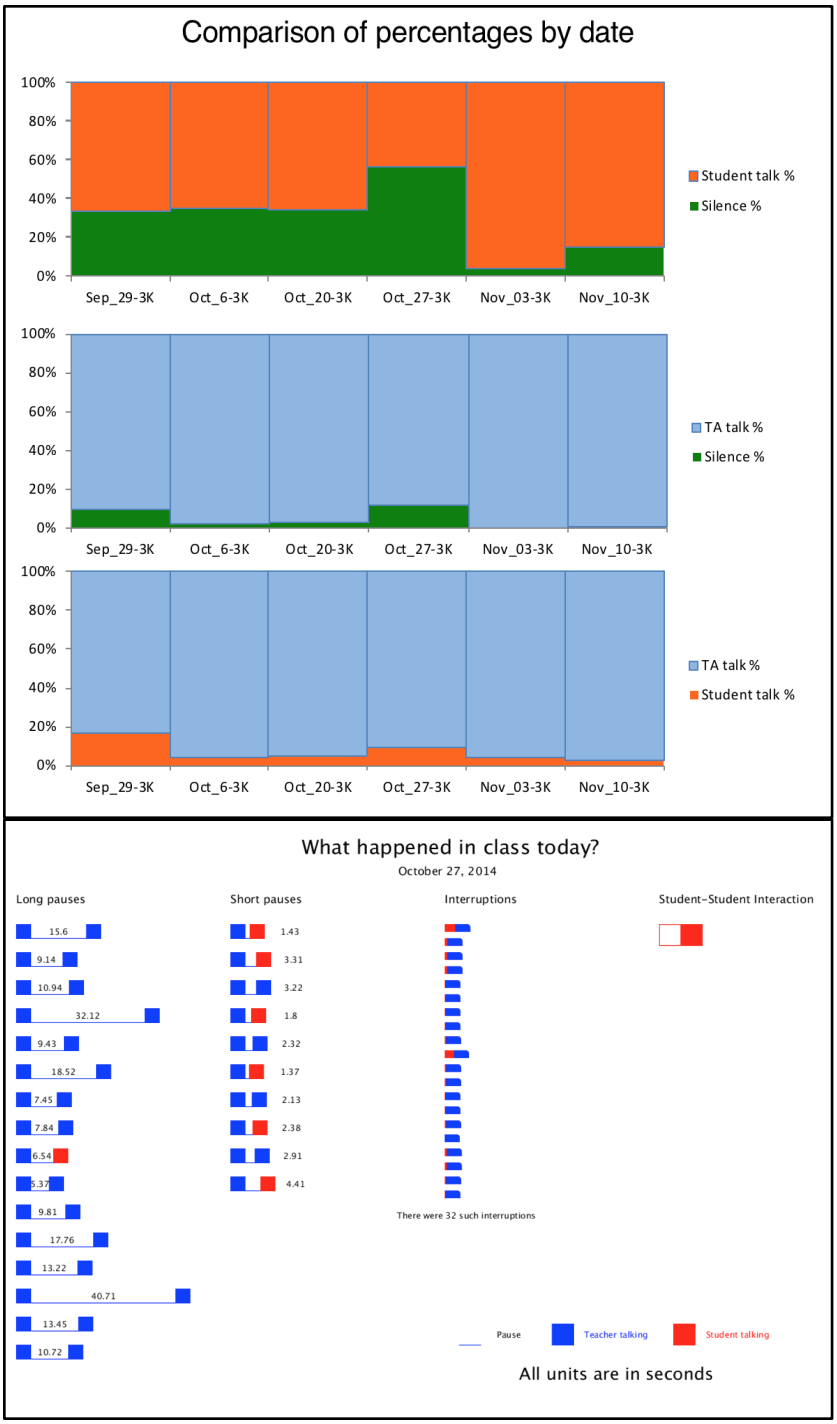


Figure 3.6: Examples of intentionally difficult visualizations. Percentage comparisons (top) were designed to isolate two of the three relevant variables (TA talk, student talk, and silence), but they are difficult to compare. Visualizations of wait time (bottom) are isolated and free of context, making them hard to understand. These graphs were meant to elicit negative responses when shown to participants during interviews. They are objectively difficult to use, making it possible to test if participants were comfortable criticizing designs in front of the researchers.

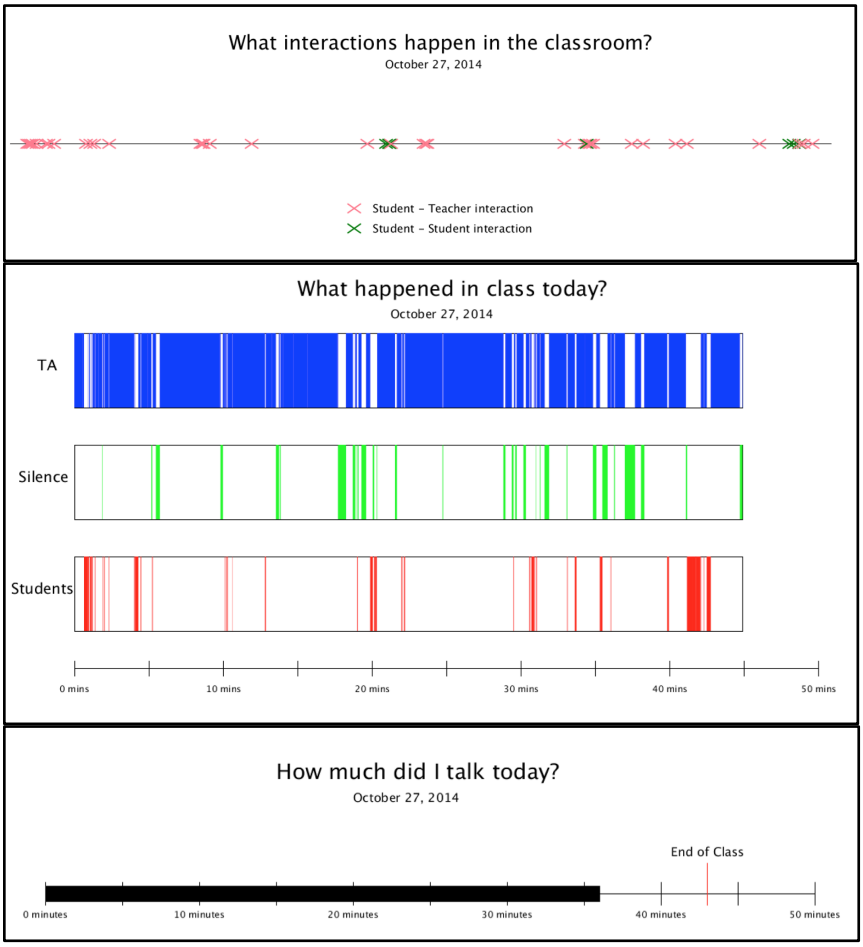


Figure 3.7: The extra visualizations (beyond ratios and timelines) used in TA3's final interview.

The final interviews were semi-structured explorations into more than just reactions to visualizations. These conversations probed TAs on their beliefs about teaching and learning, their experiences in past teaching or training, their likelihood of teaching more in the future, and their perspectives on the use of discursive teaching techniques.

3.2 Data analysis

Using the observations/field notes, I used exploratory data analysis (Behrens, 1997) to uncover trends about how the TAs were teaching. I looked within subjects at potential changes in behavior over time, such as the numbers and types of questions they asked, and the length of silence they allowed after each question before speaking again.

Qualitative data sources included observational field notes and quotes/impressions from the interviews. I used the interview data to build an "affinity diagram," as described in contextual design (Beyer & Holtzblatt, 1998). This is an inductive process that surfaces themes of perspectives across participants. To do this, I transcribed each interview in its entirety. I then segmented each conversation based on every discrete participant idea or statement. I labeled each of these statements with a numbered

participant ID, scrambled them, and built an inductive clustering of themes by iteratively arranging and rearranging the statements. Reflecting on these data sources produced the following findings.

To reduce any unintentional error possibly introduced by the live question cataloging system, I analyzed the audio logs from a random sample of 5 minutes of audio from every class (about 10% of the audio data) to calculate the number of questions that TAs asked. From this sample I found that TAs asked students about 26 questions per class (SD=9.7). Approximately half of these were closed-ended, content-specific (e.g., “Is it physical memory or virtual memory, or is it both?”). Roughly one-quarter were the TA asking, "Are there any questions?" The remaining questions involved administration such as whether or not the students had completed their homework. I observed only a small number were open-ended questions, such as "Can anyone tell me why this is the answer?"

3.3 Findings

3.3.1 Mixed methods reports

TAs had little to no training. They did not think that improving their teaching was a goal of their TA experience, nor did they think they had time within the performance of their TA duties to work on improving their teaching. One consistent theme across the TAs’ experiences was a lack of training experience. They did not believe that they were expected to take steps to improve their teaching, nor did they feel they had the time to pursue it. They did not take advantage of training opportunities available on campus. The sample of TAs were all teaching as a cohort of instructors who shared teaching materials and resources for a single class. Interestingly, they did not perceive themselves as belonging to a community of people working on their teaching. Rather they viewed themselves as domain experts who were improving their own mastery of the domain content by teaching it.

The recitations were almost entirely lecture. TAs mostly described theory, led demonstrations of algorithms, and occasionally gave students a problem to attempt silently on their own. There was very little discussion between the TA and the students at the class level, and there was never any discussion among students. Averaging across TAs and weeks, instructor speech filled 92% of class time (SD=3.6%); student talk filled 5.25% (SD=2.3%); and time where no one talked with no talk filled the remaining time (M=2.75%). Instances of student talk had an average duration of 6.2 seconds (Median=3.4, SD=12.6).

LeGros & Faez (LeGros & Faez, 2012) observed a similar pattern of questions asked by TAs. I found that students replied to approximately half of the content-specific questions, but only approximately 10% of the time when TAs asked if they had any questions, replicating prior findings suggesting this is a poor technique for encouraging participation (Rounds, 1994). In an interview probing the use of questions to engage students, TA-1 shared, "Probably, I should, like, ask more times if they have questions." This response followed a class where this approach constituted 13 of the 15 total questions asked, and where only 1 student contributed during the entire class.

Waiting after asking a question has been shown to increase student participation and improve student learning. Across the four coded classes, students responded to 15% of TA questions, and TAs spoke first after 85% of their questions. In terms of wait time, TA3 was an outlier, averaging 6.6 seconds before speaking again. He did not appear to use the long pauses in conjunction with any social pressure to answer, however, such as looking directly at the students. All other TAs averaged below 2 seconds. TA2 justified this lack of wait time as a result of students not knowing the answer: “If a teacher asks a question in the class, you should be able to answer.”

There were no further techniques to get students talking. Following student speech events, TAs averaged below one second pause (M=.94, SD=1.1). Only 2% of all student speech acts were followed by another student, rather than by the TA. This was not evidence of students being unwilling to discuss the material, however. I observed a large amount of student to student interaction in the few minutes before class started. Most consisted of animated discussions about the course content. Students spoke about their progress on homework, described their approach to the material, and discussed challenges in understanding the material. In most cases these discussions ended abruptly when the TAs asked students to stop talking in order to begin the lecture. Over the course of the semester, these pre-class discussions diminished.

3.3.2 Reactions to data visualization

After reviewing student-TA talk visualizations, TAs expressed surprise at the amount of time they lectured. They said that they wanted to engage students in discussion, but that they did not know how. They were discouraged by how difficult it seemed. TA1 said he would, "… keep asking if there are any questions but … no one speaks, so I cannot help this one." They generally interpreted students' lack of response to their prompt "any questions?" to mean that students were aware of what they did not know, or else they would have asked.

While TAs had an intuition that it was beneficial for students to speak up, they also expressed a sense of conflict. They felt pressure to cover all of the material, and worried discussion might prevent this. TA2 reported, "Maybe I might want to involve their participation a bit more than what it is, but I also fear by doing so, if they'll be able to complete the [assignment]…" There were also cases where the TAs did not empathize with what their students might not know, e.g., when TA4 described his own classrooms learning experiences: "… participation was not that important. I used to get what [the instructor] was teaching."

3.3.3 Reactions to teaching strategies

When prompted about trying specific discursive techniques, TAs said they would feel uncomfortable using students' names, and they were unlikely to call on specific students to answer questions. The one technique they seemed willing to consider was increasing wait time after asking a question, but they had no ideas about how long they should wait. At a general level, they did not see much value in asking students questions. For example, TA2 reported that, "The recitation is supposed to cover what the professor taught and not ask too many questions to the students."

TAs acknowledged that they had little intuition about how to teach, and they welcomed the idea of improving their teaching through objective feedback about what they were doing. TA5, for example, said "I would keep track of a lot of things when I'm actually lecturing because there are so many things to be worried about. But I would definitely want to be conscious about how to make my teaching better; how to make my class go better."

Although some of the TAs had acknowledged early in the semester that they wished their students would speak up more, several of their attitudes changed by the end of the semester. The accumulation of failed attempts to get students to ask questions, combined with a lack of guidance on how to improve, may have led to some TAs dismissing their original admissions of talking too much, e.g., when TA1 said that he would, "… keep asking if there are any questions but … no one speaks, so I cannot help this one."

3.4 Insights

3.4.1 Reversion to familiar strategies

With so much time spent in the formal class lecture, continuing that pattern of teaching in recitations seems like a lost opportunity. TAs exhibit teacher-focus and patterns of information transmission at the risk of losing their students. These student instructors have such small classes that they could easily incorporate student participation in active knowledge construction rather than fall back to boring patterns of lecture. And yet this is what they do, perhaps because of their lack of knowledge about how to teach.

TAs needed cognitive support to learn about techniques they could use to increase student participation. These techniques should be introduced to the TAs throughout their teaching experience. They should encounter new ideas about teaching in a meaningful and useful order. For example, tactics drawn from the literature such as "wait X seconds," would not address the TAs persistent use of shallow questions or the felt need to cover material through lecture. TAs would need explicit training on the value of deep questions and how to integrate them into their class structure.

Post-class feedback through behavioral visualization seemed to elicit some awareness of a problem, but it did not seem to motivate a desire for behavior change. In our approach the TAs only saw these graphs twice throughout the semester, and they were given no guidance on how to interpret the information. It seems likely that ongoing attempts to guide TAs through how to teach should provide more consistent feedback with clearly stated goals of what the TAs should be trying to achieve.

The TAs' repeated statements about not knowing what to do indicated not just a lack of pedagogical strategies, but also a lack of confidence. On the surface, TAs expressed frustration at not knowing how to engage their students. However, when I gave them strategies to consider during the interviews, they immediately dismissed most of them. They did not want to ask unfamiliar types of questions, like asking students to elaborate, because it might make them look unknowledgeable. They did not want to call on students, because they might seem intimidating. The only strategy they were willing to consider was increasing wait time. In the studies that follow, I will attempt to increase the number of questions that TAs ask as well as the wait time they employ. One goal of the work will be to address low confidence by promoting behaviors that novices might be more comfortable starting with.

3.4.2 Revising Personal Informatics

This study attempted to use the techniques of PI to change teaching behaviors. Previous PI research had improved users' attitudes (Lin et al., 2006) and raised awareness (Comber et al., 2013). In this study, however, simply seeing data about the behaviors that needed to change did not motivate TAs to try anything new. These users did not describe an increased appreciation of the variables of interest.

Although PI relies on reflection-heavy theories (e.g., Epstein, Ping, Fogarty, & Munson, 2015; Rivera-Pelayo, Zacharias, Müller, & Braun, 2012), the interventions do not often push users to reflect deeply. This is a serious limitation to the theory, or at least to its typical application. Most PI artifacts have very little scaffolding for reflection. They tend to operationalize that part of the theory through nothing more than a display of graphs. Perhaps this is sufficient for runners or money savers, but it does not seem to bridge the gap for TAs. The reason might be that PI interventions are usually based on topics that users are already aware of and already care about. Going forward, PI toward belief change needs to find a way to introduce topics that may be unfamiliar or even uncomfortable for users to adopt.

Chapter 4: PD with an instructional app (Study 2)

In this chapter I explore the use of explicit topic introduction and justification with scaffolded reflection in a partial PI framework. I leave out the personal data visualizations in order to avoid undermining TA confidence and amplify the possibility of increasing knowledge and risk taking. I return to using personal data in the two studies that finish out the document, but only after having confirmed that topic exploration and performance reflection can produce behavior and belief change. This second study builds an initial, iterative intervention in the learning environment, building on the observational findings in a DBR cycle, moving from problem identification to initial prototype solutions.

This cycle typically involves a research team reviewing the roadblocks that emerge in an exploratory study, brainstorming sketches of solutions, and building one or more low-fidelity prototype solution. The roadblocks that emerged in Study 1 include the following:

1. Giving TAs access to their data alone does not help them
2. Failure to engage students, even when they want to, seems to overwhelm any ambitions to improve.
3. TAs want to do better, but don't know how.
    a. They need help identifying and setting goals (**metacognitive strategy support**).
4. They exhibit a lot of teacher-focus, and shy away from student-centered practices.
    a. They need help understanding the value of student-centered strategies (**cognitive support**).
5. They need explicit exposure to specific teaching skills worth developing, especially behaviors around questioning their students, i.e., types of questions to ask and how to ask them, with opportunities to practice (**procedural strategy support**).

In this work I needed to avoid building a narrow training app that would help a specific set of TAs in my target domain. This might improve the outcomes of that small group, but would have low impact, as it is impractical to design different interventions for every single domain that different TAs teach. Instead, DBR suggests building prototypes that produce design reflections and transferable principles (McKenney & Reeves, 2012). From those principles, future designs ought to be able to scale to a wider population of TAs teaching different topics.

This is challenging because designers cannot necessarily know what "better" looks like for every different context of teaching. A general system lacks the human intelligence of a professional consultant and cannot easily tailor its instructions to the specific needs of each client. In this study, I ask users to provide the content and context for their own classes. I explore whether doing so might engage them more in the process.

To work toward this goal, I built a system that asks users to provide the concrete material relevance of their own context while they learn about general pedagogical principles. For example, I ask TAs to explicate the specific learning objectives for the class which follows each interaction. The intention is for users to cast the particulars of their circumstance onto the general tasks presented by a generalized training system that is blind to the individual. I want to know if it is possible to substitute prompts for TA reflection in place of human-guided feedback—and would this approach lead to substantive changes in the teaching and learning environment?

More specifically, this study tests two big ideas around personalization and viability.

1. How should designers build a PD app that implements algorithms for personalization?
2. Would TAs adopt new teaching behaviors by using a PD app that provides explicit topic instruction, support for practice, and prompted self-reflection?

4.1 Methods and implementation

To answer questions regarding scalability, the research team produced a set of design parameters based on the findings from Study 1. The design goals for our prototype technological intervention included:

1. Let the users produce the context particulars in order to personalize the experience
2. Let the app define goals on low-level features, i.e., talk less and ask better questions
3. Guide the users in assessing their own performance

With these goals in mind, we built a framework curriculum for what the TAs would learn over the course of a semester. This included identifying the high level learning goals for participants in the study, as well as a loose set of hypotheses about what conceptual units comprise the larger goals. The primary learning objective was how to get students to engage more with in-class discussion.

One effective approach to increase engagement is the use of "deep questions" during class. These are open-ended inquiries that encourage interpretation, analysis, and elaboration. They can help students construct new knowledge as they analyze and synthesize information, build on their assumptions, and generate new ideas (Bolen, 2009; Chen et al., 2014; K. Ellis, 1993). As students respond to these prompts and participate in class discussion, they exercise critical thinking and improve fact retention (Weaver & Qi, 2005). Even quiet students can benefit from classroom discussion, because talkative students are more likely to ask questions that are relevant to everyone (Howard et al., 1996).

The evidence that deep questions are producing the intended result lies in active student participation (Oliveira, 2010). One common approach to increasing student-centered practices in college classrooms, therefore, is to consider the kinds of questions students are encountering, and the opportunities they have to elaborate on what they are learning (Bolen, 2009; K. Ellis, 1993; Galloway & Mickelson, 1973; Oliveira, 2010).

Using backward curriculum design (Wiggins & McTighe, 2005), we created interaction activities that require TAs to define and implement different types of questions, and learn methods for encouraging students to answer the questions.

Following the iterative nature of DBR, we implemented a weekly design cycle in order to respond to the emerging needs of the population. While the large-scale goals did not change, low-level tactics for reaching those goals needed to adapt to the reality of the classrooms involved. For example, we left the overall number of training sessions open-ended, and designed each instructional module in response to the data from the previous week of classes. The framework curriculum guided the high-level instructional objectives. The actual lived use of the intervention and any changes in teaching behaviors guided the details of each subsequent iteration.

4.2 Protocol

4.2.1 Recruitment and data collection

Like Study 3, this research recruited from teaching assistants for an introductory computer systems class at Carnegie Mellon University. The study took place during the spring 2016 semester. None of the TAs involved in Study 1 were still teaching the course. There were 10 TAs who volunteered to participate. From this group 5 were sharing teaching roles for the same class sections. I recruited the remaining 5 who were the sole instructors for their recitations.

There were 2 women and 3 men in this sample. Of the 5 participants who participated in observations and digital professional development, only one had ever received training from the campus's teaching and learning center. 3 TAs had taught previously. TAs 2, 3, and 4 were present for all classes. TA5 missed one session (week 7) due to a job interview. Students were sent to TA1 that week. TA1 and TA5 both had sparse attendance. No students came to weeks 5 or 6 for TA1. That section was canceled after week 7 and TA5 acquired the students who had attended TA1. The research team performed 10 weeks of observations and produced 50 hours of audio recordings and live-coded classroom behaviors.

With external programming support I built an application to simplify in-class coding of behavioral data, and called the app the *TA Dashboard*. It functioned by allowing in-class observers to use single keystrokes to catalog well-defined classroom variables on the fly. TA Dashboard logged the time of each keydown and keyup to build a database of classroom behaviors. Each database saved to the coder's system. Following class, observers uploaded the logged data to a repository for full analysis. Each data structure produced metadata about each class, as described in Table 4.1.

For each class we recorded audio and logged variables through TA Dashboard. Research assistants ran this program on laptops during each recitation, and wrote down low-level qualitative observations, such as the number and perceived gender of students who attended. We performed observations for ten weeks, producing 50 hours of audio-based and live-coded classroom data, and pre- and post-class check-ins. We logged in-class behaviors for the first three weeks to provide a baseline of talk time, question asking, and use of any discursive tactics.

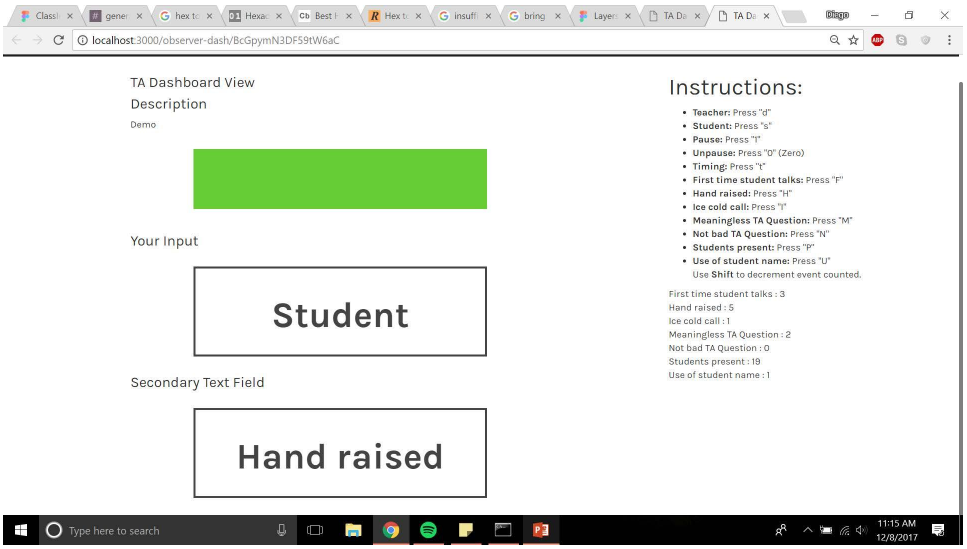


Figure 4.1: User interface of the TA Dashboard.

Table 4.1: Metadata produced by the TA Dashboard.

Variable	Description
coder	The name of the researcher/observer.
Section	The recitation section, includes A, B, C, E and I
Date	An Excel formatted date of the section.
st-talk	The total number of milliseconds where coder held down 'S' key, representing student talk.
ta-talk	The total number of milliseconds where coder held down 'K' key, representing TA talk.
overlap	The total number of milliseconds in which ‘K’ and ‘S’ were pressed simultaneously.
silence	The total number of milliseconds in which neither ‘K’ nor ‘S’ were pressed.
ta/st	The ratio of ta-talk to st-talk, or total TA talk to total Student talk.
time	The time of day when the class begins.
attend	The total number of students in class.
stTurns	The total number of students talk turns.
turns/st	The ratio of student talk turns to total students in attendance.
men	The total number of men in class.
women	The total number of women in class.
u-st	The unique number of students who speak during class.
r-st	The number of spoken contributions made by students who had already spoken.

On week 4 TAs started using the online training system. They interacted with it until week 10. For each training interaction, TAs received an email link guiding them the next lesson. There were either 1 or 2 new training interactions per week. I sent prompt messages personally, and each new activity had unique wording in the email. The links opened our training tool which we meant for use while they planned their upcoming class. We developed the modules using the commercial service Qualtrics. Use of a commercial service allowed rapid creation of weekly modules in response to the TAs online interactions and their in-class behaviors.

Each TA’s interactions with Qualtrics generated a rich set of text-based and multi-select self-reports. We also interviewed each TA at the end of the semester for over an hour each, probing them on their experience with the training and their beliefs about student participation. Following the study, the research team reviewed all of the audio records from each class and transcribed every question that the TAs asked.

4.3 Training artifact

In weeks 1-3 we took baseline measurements of classroom behaviors. TAs received no training modules. From weeks 4 to 10, they received a standard issue of training modules that addressed the high-level system requirements. As new insights emerged each week, design changes cascaded through the following modules. Here we list only the unique goals, activities, rationale, and assessments that emerged.

We emailed TAs a link to the first module about four days before each class. We sent reminders nearly every day to participants who had not accessed the training.

Weeks 1-3: Baseline measurements of classroom behavior only. There were no training modules.

Week 4—Share with colleagues: In the first training, TAs were prompted to write concrete learning goals and questions they might ask in class. We then collected and shared those entries back to the group. After seeing each other's learning goals and questions, they were prompted to compare colleagues' goals with their own. PD research recommends collaborative learning with colleagues (Brinko, 1993). Exposure to what others write promotes additional reflection, as well as questioning their own assumptions. For assessment, the research team reviewed the submitted learning goals for face validity, and they collectively coded questions as deep or shallow. This helped assess the TAs' initial conceptual knowledge.

Week 5—Practice: In the prior week, when given a prompt to write an open question for class, only half of the TAs succeeded. We therefore designed opportunities to learn more about this concept. The PD literature suggests that instructors practice classifying questions amongst different types, such as deep and shallow learning (Bolen, 2009; K. Ellis, 1993; Galloway & Mickelson, 1973; Oliveira, 2010). Our system asked TAs to classify questions as "open" or "closed." (We used these terms assuming TAs would be more familiar with them than "deep" and "shallow.") For assessment, we analyzed their accuracy, and also asked them to define the meaning of each question type in their own words. Finally, we gave them our own definitions of the terms, inspired by the literature.

Week 6— Enact behavior change: In the previous week, the prompts for question examples produced thoughtful answers. TAs improved accuracy on deep questions. Our in-class observations of week 5 showed that the TAs were better at writing questions. However, they did not ask more questions during class. We added an explicit suggestion to include questions on the slides. For assessment, we examined their slides and counted the number of asked questions.

Week 7—Share and discover strategies: Our in-class observations of week 6 showed an overall increase in both question asking and number of questions written on slides. This suggested that TAs were coming to trust the system's suggestions. In the original observation study, TAs had expressed an unwillingness to try discursive tactics. However, given this evidence of trust, we moved forward with training on general discursive tactics.

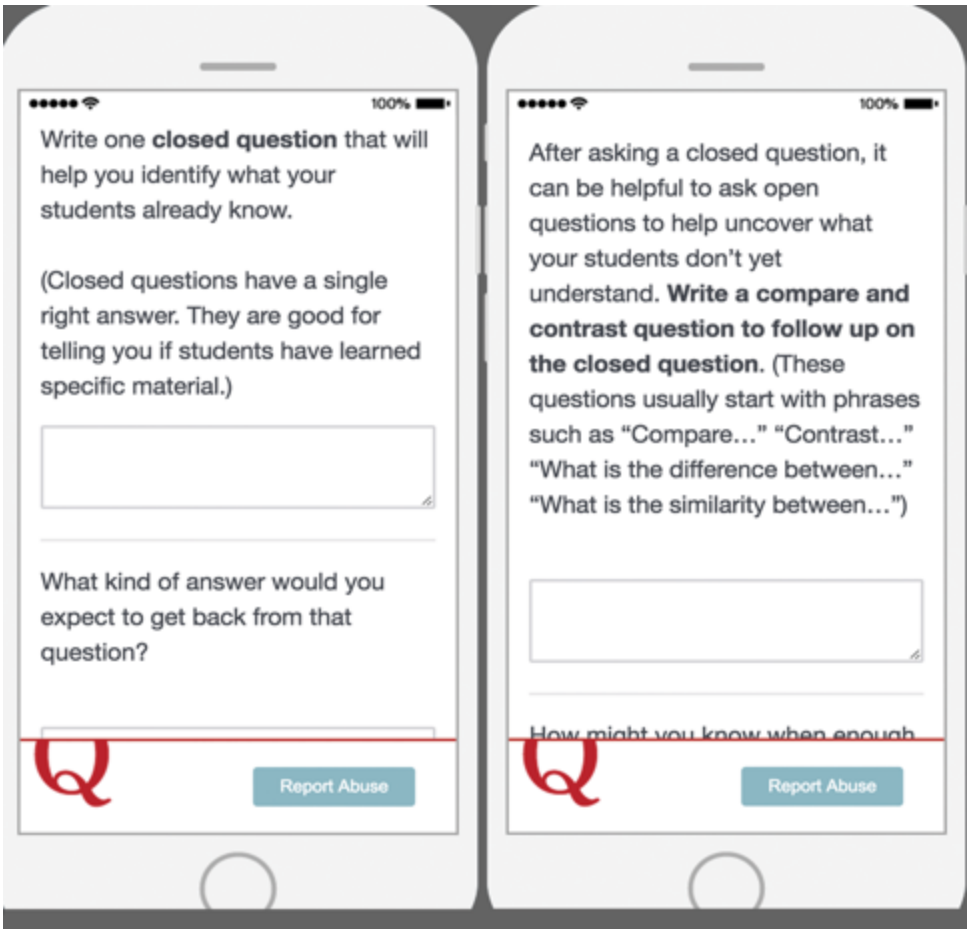


Figure 4.2: One of the interactions from second training. The left screen shows a scaffold for writing closed questions. The right screen shows a scaffold for writing open questions.

The system asked TAs to reflect on student participation, and to suggest ideas for how to increase it. It then provided suggestions of research-supported discursive tactics. Asking for them to contribute their own solutions acknowledged their personal lived experience, and tapped their specific craft knowledge. It also enabled the system to coordinate sharing resources between participants without needing an internal model of the pedagogical content.

For assessment, we reviewed their suggested solutions for face validity. We also examined in-class behavior to track any changes in discursive tactic use.

Week 8—Goal-setting: From PD literature we knew that instructors are more likely to change if they set specific goals (Brinko, 1993). In week 7, TAs became familiar with discursive tactics. In week 8, we reproduced that full list of tactics and asked TAs to select 3 they would attempt in their next session. The system asked them to write a question that they would use in conjunction with each selected tactic. For assessment, we compared their selected tactics to those observed in class.

Week 9—Review/Reflection: Several TAs dramatically changed their normal teaching routine at this point in the study. They began using live demos during the second half of their class. We had no pre-planned formula or learning objective that would address this outcome. In response, we embedded an exploratory survey in the training to find out why they had changed teaching styles.

Week 10—Experience sample: Nearing the end of the intervention, we embedded an exploratory survey in the training to probe TAs' perspectives on any changes they may have experienced throughout the course of the study.

After the ten weeks of deployment we performed exit surveys and interviews for all ten participants. Similar to the first study, participants came to the lab for a semi-structured interview regarding their beliefs about teaching, and also their impressions of the intervention.

4.4 Findings

4.4.1 Data coding and analysis

The external behavior of interest for this study had to do with behaviors around TA question asking. I was also interested in the perspectives of the TAs throughout the course of the semester, and how they interacted with the training. I analyzed the following data sources:

1. Audio logs of each observed class
2. TA Dashboard output
3. Online tool use and interactions (Qualtrics logs)
4. Field notes
5. Interviews

To learn about the questions that TAs asked, four members of the research team coded 2490 questions as Deep, Shallow, or "Any" Questions. From the system logs, we extracted the tactics they selected and the questions they wrote during in-app interactions. We search the defined data for evidence that the system impacted TAs' in-class behaviors, teaching beliefs, or dispositions and motivation.

From the audio of the observations, we first extracted each question asked by each TA. We created a coding manual based on a deep review of questions type literature (e.g., Gall, 1970; Graesser & Person, 1994; Redfield & Rousseau, 1981). Four members of the research team coded the resulting 2490 questions as Deep, Shallow, or "Any" Questions, after achieving pair-wise kappa calculations greater than .70 on 5% of the data. We then calculated the length of time each TA waited for a response following each question. We assessed how many total questions TAs asked per class, how many of each type of question, and how long they waited after each question.

From the system logs, we extracted the tactics they selected and the questions they produced in the planning stages of training. We compared these to our notes on tactics used and questions asked in class. The end of each module also included free-response questions from which we drew qualitative data about positive and negative impressions of the intervention.

The research team followed an iterative, structured reflection method common to DBR (McKenney & Reeves, 2012; Reymen et al., 2006; Visscher-Voerman & Gustafson, 2004; Wang & Hannafin, 2005), and reviewed the interview data for each TA. Specifically, I used 'Line (Quality) Norms' (McKenney & Reeves, 2012) as an analytical framework. We attended to examples where training may have influenced the TA's behavior or beliefs.

Although there are obvious limitations of comparing the first, observation study with this study, qualitative contrasts are worth noting throughout the findings, given that each sample is drawn from a similar population of TAs.

4.4.2 Behavioral outcomes

Conceptual Changes: Each week TAs responded to prompts for writing "open" and "closed" questions. Their accuracy on closed questions was high for the duration of the study (average 90%). Open questions were mixed, with an average of 55% accuracy without scaffolds. With scaffolds, however, their accuracy was 85%.

Impact on Class Preparation: Interviews revealed various changes to the ways that TAs prepared for class during the training. TA3 told us that the online training introduced the concepts of "open" and "closed" questions, and using these concepts during preparation improved their questions.

TA2 and TA4, who had taught before but had never received training, both described pre-intervention preparation strategies that included reading through the slides the night before class and taking note of any tricky parts that might require rehearsal. Neither reported changing this set of strategies during the training, but TA2 additionally made sticky note reminders of questions to ask. TA4 reviewed the community feedback from fellow TAs supplied by the training activity, and integrated it into planning the class.

TA1 and TA5, both brand new to teaching, each described spending 2 to 3 hours preparing for class prior to the training. They read through the slides, looked at the textbook, reviewed personal notes from when they had taken the class, and thought about the topics from the students' perspective. We expected that the training system would add additional preparation time for the TAs. However, during the intervention they each reported spending up to 45 *fewer* minutes preparing. They reallocated their efforts, spending less time reading the material and more time thinking about how questions could drive the focus of the lesson. TA5 began associating strategies to each question, saying, "It has shown me a systematic approach to preparing for teaching a class by listing down goals, classifying questions, picking strategies and picking the most important questions to be asked."

Use of system-suggested tactics: The system prompted TAs to ask more questions in general and more open questions in particular. Table 4.2 shows the mean number of questions asked during baseline (weeks 1-3) and during training (weeks 4-10), and a percentage change between baseline and training, averaged across TAs. These descriptive statistics are not meant to claim a causal relationship, but rather illustrate changes in behavior that are worth considering given the research goals.

TAs collectively increased the total number of questions they asked by 36%. The average number of deep questions increased across all TAs by 47%, and shallow questions increased by 43%. These increases were slow to materialize, as TAs did not begin asking more questions until the fourth week of intervention. On that week, the system had prompted them to include questions directly on the slides.

Table 4.2: Average number of questions across TAs. Baseline = weeks 1-3, Training = weeks 4-10, % Change = (Training – Baseline)/Baseline.

	Deep	**Shallow**	**"Any"**	**Total**
Baseline (3 sessions)	6.2	21.7	4.1	32
Training (7 sessions)	9.2	31	3.4	43.5
% Change	**47%**	**43%**	**-17%**	**36%**

The training suggested that "Any questions?" was an ineffective way to elicit student participation, and the average number of times TAs asked if students had "Any questions" fell by 17%. This result may show that TAs began to think about this question differently than they had during my original observations. Previously, TAs said that if students did not respond to the prompt then it meant they understood everything.

Use of TA-selected tactics: In Week 8, the system prompted TAs to select from specific tactics to try in the following section. We reviewed and compared their selections to our field notes and audio records to determine which tactics they enacted in class. Of all the tactics TAs reviewed, the ones they said they were most likely to attempt were *waiting longer after asking questions*, *calling on students by name,* and *encouraging students*.

Increasing wait time was the only tactic that TAs from Study 1 said they would be willing to try. It was also the most popular technique selected during Study 2. TAs 1-4 all indicated that they would try it. We found that when subsequently prompted to reflect on their actions, TAs did not accurately estimate their own wait time. TA2 claimed to have waited longer, while our analysis revealed the wait time following deep and shallow questions actually *decreased* (1.8s to 0.4s, and 2.1s to 1.3s, respectively). We observed that when asking a question, TA2 would downplay the importance of trying to answer, apologize for the "bother" of asking questions, and tell students they probably would not know the answer. TA3, on the other hand, reported, "I don't think I did a very good job of increasing wait time. As usual, I was trying to pack a fair amount of material into a short amount of time…." Their average wait time actually increased slightly for open questions (7.7s to 8.3s) and substantially for closed questions (3.7s to 7.2s). The only TA with an accurate estimate was TA4 who increased from an already high wait time of 6.5s to 17.4s for deep questions, saying: "The other thing that actually did help a lot was like wait longer to the point where it actually becomes uncomfortable for the students and somebody eventually decides to raise their hand."

TAs 3, 4, and 5 all said that calling on students and using students' names was a valuable strategy, and we observed each of them using it. This stands in contrast to our findings from Study 1, where TAs received no support for reflection or planning, and where they overwhelmingly rejected the idea of using these tactics. TA5 began calling students by name, and told us that it, "kept them [the students] alert." They even poked a student with an eraser when the student had fallen asleep. We observed TA3 begin using names while calling on students in response to raised hands.

TA-initiated behavioral change: During Week 8 we began to see evidence that the goal of increasing student participation was becoming salient for the participants, as several of the TAs began using a tactic not suggested in the training: live in-class demos. Asked how this approach was conceived, several participants reported that TA2 had proposed using demonstrations because students were not talking enough, and in conversation with the experimenter after class, TA2 reiterated that there was "too much lecture" in recitations. TA2 designed two demos with the goal of programming part of a server live in front of the class, while soliciting supportive input from the students. All but one of the observed TAs then used the idea in the following class.

TA2 and TA4 each ended up speaking *more* during the demo classes while explaining what they were coding. We observed an overall drop in student speech events. TA3, instead, designed their own unique demo. They coded half of the server while students watched, and then asked the students to work together to program the rest. This led to a very active class with students speaking together in small groups while the TA moved through the room answering questions and providing encouragement. TA5 decided not to use a demonstration and instead continued using the discursive tactics discussed in the training.

4.4.3 Teaching belief outcomes

My original observation study had revealed that TAs thought it was more important to cover material than to engage students, a common belief among teacher-centered instructors. By contrast, by the end of this study, 4 out of 5 of the observed TAs said they were confident that it is better to probe students' knowledge rather than try to cover everything. TA5 reported adding more time in class to thinking about core concepts rather than reviewing all of the available material.

The TAs reported valuing student talk during class. TA3 engaged in critical self-reflection with respect to student participation, stating, "If students don't talk in class, it's probably not an effective learning environment … Unfortunately, I recently realized I've made it possible for a couple of less confident (or maybe just shyer?) students to slip through the cracks and avoid participating, so my sessions probably weren't very helpful to them."

We also asked TAs to describe their experience with asking questions during class. TA1 said, "By forcing myself to ask the students more questions, I am hopefully engaging them more and they will remember the material better." This TA also mentioned that the process of thinking through question types helped reduce "feeding" answers to the students. TA1 also described an unintended consequence of the training on their own *learning* process, "I notice myself paying less attention in lectures/recitations where they basically just talk to you the entire time without you doing any work."

4.4.4 Dispositional and motivational outcomes

Acquisition of self-efficacy: Several TAs indicated a belief in their ability to become better instructors following the training. This stood in contrast to participants' self-efficacy following Study 1, where most TAs said there was nothing more they could do to get students to talk. TA2 said, "I have done a better job of asking students questions and making eye contact with them. I have also done a better job of asking questions in a way that makes it more likely for them to answer, like relating it to other things they have done already." TA4 said that teaching was, "… as much of a learning experience for me as it is for the students. I love seeing how my teaching changes semester by semester." TA5 reported feeling more comfortable with teaching: "TAs can be scared about delivering content, and this helped keep things on track."

Motivation to participate in training: We observed a high participation rate throughout the study. Each module was completed by at least eight of the ten TAs. Non-participation was evenly distributed; no single participant showed repeated inactivity. Even after TA1's class was cancelled, this participant continued the weekly training. Each week, about 1/3 responded to the first email invitation, 1/3 to the follow-up reminder (typically a day later), and the rest to the second follow-up (typically a day after that). One said it was easy to use: "I think I actually filled out a couple of them riding the bus, on my phone."

Without prompting, three TAs reported that they liked seeing the contributions of their colleagues, and some said that they went back to the group's questions to prepare for their own class. Some TAs asked for access to the material their colleagues had produced online so they "wouldn't have to take so many notes during training" (TA3).

In the final interview we probed participants on the utility the training. TA3 told us that it varied because some parts took too long to do, but that formulating learning objectives, and learning about question types was valuable. This TA put more effort into the training than the other participants, writing very thoughtful questions and long feedback nearly every week. TA2 said that it was

"Somewhere in between," and that what helped was the need to read the class slides early and formulate questions. This was a result of receiving our training prompt four days before class. Four TAs said the training was valuable because it motivated them to prepare for class earlier than normal.

Intention to Continue Learning: There was some evidence to suggest that TAs had adopted new perspectives on behaviors that they would carry forward. TA3 described an intention to do better next time, saying "I've learned numerous strategies for making sessions more interactive, and have experimented with using some of them. I didn't manage to apply everything, and some of my implementation didn't work, so there's room for improvement when I TA in the future." TA4 described an aspiration to improve performance for future teaching appointments, "Students have evaluated me well in the past, but I think I could have done better by asking more questions in class. I frequently see students on their laptops not paying attention, and I hope that I'll get feedback from them on how I could be more engaging with them."

4.5 Summary of results

4.5.1 Scalability

Returning to the research questions, I first asked: How should designers build a training app that could scale to a wide population of TAs?

DBR recommends that the answer to the first question comes from artifact reflection. This is a critical step in order to glean transferable insights. Reviewing the design constraints and solutions from each weekly iteration of the technology probe, a series of training modules or interactional phases emerged, revealing a general training system that matches stages of learning. The following modules detail how the system evolved to match the learning stages.

Module 1. Instruction: Each cycle of conceptual training should begin with a short instructional module that assesses the instructors' knowledge of the relevant pedagogical concept (e.g., differences between shallow and deep questions). This could also be a prompt for the user's solution ideas for a common problem. This produces a list of terms or ideas that the system can redistribute amongst the group.

Module 2. Reflection/Collaboration: Allow instructors to review and assess each other's submissions from the first module. This helps maintain quality and relevance of the material, and potentially reveal misconceptions by drawing on community knowledge. This provides a programmatic way of reviewing data quality without domain expertise. It also supports a sense of community and collaboration amongst the instructors.

Module 3. Planning: Ask instructors to select a specific solution or set of solutions they are willing to try. Their options should draw from the list produced in Module 1 and refined in Module 2. Request that they attempt their selection in their next enactment opportunity. Produce an internal list of popular selections in order to assess intentional states. Additionally, use this module to promote any other relevant planning and practice opportunities.

Module 4. Reflection/Review: After their enactment opportunity, ask instructors to recall and review their decisions from Module 3. Request a review of their progress. Return to Module 1 for the next conceptual lesson.

This research question also included questions about the viability of requiring TAs to do some of the work of a human consultant by providing the context of their own classrooms. We found that TAs were able to produce relevant context from their own domains. In fact, TAs produced relatively rich reflections that seemed to impact their decision-making, in spite of their inaccurate self-assessment.

4.5.2 Viability

My second question asked: would TAs adopt new teaching behaviors by using an app that provides direct instruction? We found that some TAs adopted new behaviors that the data-only approach of Study 1 did not produce. Even the TAs that did not adopt new behaviors at least perceived themselves as having changed, as exemplified by TA2. Addressing this accuracy problem may rely on re-introducing personal teaching data in the training. It is possible that constraining feedback within a supportive framework would overcome the diminished motivation that TAs in Study 1 exhibited.

Moving forward, the study also shows a training cycle starting from the point where TAs use the artifact as they plan their coming sessions. This reveals that training should attach training modules to phases of reflection and planning.

Chapter 5: A framework for SmartPD

This chapter summarizes the first two exploratory studies, investigating how TAs teach and the opportunity for improving TA teaching via a new socio-technical system. The result of these two studies is a framework describing the process by which TAs can learn to teach. I describe the details that the framework must consider. I then walk through the construction of my framework, which I call SmartPD.

5.1 The need for a framework

Study 1 revealed very little interaction between students and TAs in recitations. TAs wanted to generate more student participation; however, they did not have effective techniques to call upon in order to achieve this goal. Asking students questions appeared to be ineffective. TAs only asked convergent, shallow questions, or they asked whether or not students had any questions. Study 1 also showed that delivering data about the TA's performance may build frustration rather than change behavior. I suspect this happened because the information provided showed only their performance, and it did not suggest ways they could improve.

As an investigation into how to design contextually rich instruction within a socio-technical system, Study 2 uncovered important design principles for this population. Through this iteratively designed prototype, interesting design themes emerged that can help guide the next stage of building a scalable training artifact. These themes related to the viability of the system. For example, it showed promise in terms of the likelihood that TAs would use it. This may have related to the use of algorithmic approaches for including user input. Other design themes emerged relating to practical concerns, such as the length and timing of instructional modules. Earlier in this thesis, I described the vision of SmartPD as a possible solution teaching TAs how to teach. My first studies showed promise that the technology would be worth developing and implementing.

Recall that SmartPD is a research space that combines Personal Informatics, Smart Classrooms, and traditional PD. Through the use of sensors and a database of each instructor's real-world actions, I envision a training system that implements the direct instruction, feedback, and goal-setting strategies of PD while leveraging the reflective advantages of data-rich PI. Future potential for this line of research include compiling data from these disparate interaction spaces and building models of instructor knowledge and skill. Models of this type may eventually help researchers in SmartPD determine the appropriate learning pathway for novice teachers. They might even provide live hints or recommendations (much like modern Intelligent Tutoring Systems do).

The primary objectives for SmartPD in its current stage are to:

- Gather behavioral data from real-world classrooms
- Discriminate and classify pedagogically meaningful behaviors
- Support grounded reflection on users' implementation of pedagogical tasks
- Provide direct instruction on new pedagogical ideas
- Motivate changes in beliefs and actions

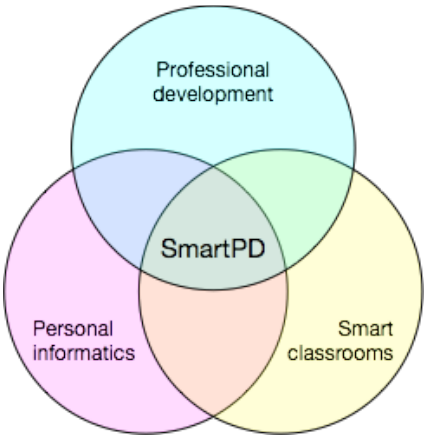


Figure 5.1: The intersection of disciplines needed for SmartPD (reprint of Figure 1.1).

No one has built a socio-technical training system that meets these objectives. In order to do so, I need a way of planning each detail of the design. Returning to descriptions of how to carry out DBR research, I found some guidance on how to do this in the stage labeled "Building tentative products and theories" (McKenney, 2001). *Tentative products* are those which compile the current findings into a testable, deployable artifact. The artifact should operationalize design intentions and produce a way for examining expected outcomes. *Tentative theory* comes from specifying the overall process in clear and potentially testable terms.

The tentative product must be an artifact that operationalizes explicit and specific design intentions. It must also produce evidence of the efficacy or validity of each operationalization. Insights from the DBR work implied that these novice instructors required feedback that includes goals and acknowledges context. Insights from related literature indicated important design recommendations to include. The following list compiles all of the design insights, a plausible operationalization of each, and a source of evidence that the operationalization supported the design intention.

5.2 Design goals for a tentative product

Here I describe the necessary design objectives as revealed by the prior studies and literature review. These design goals are for building a tentative product; a testable training system. The features are grouped within important features of learning to teach as described in the literature and confirmed or discovered in the research so far. Each design goal is described by its concrete operationalization and the evidence required to evaluate the design. Following this section, I describe how I matched each of these features to a tentative theory.

5.2.1 Feedback and reflection

Learning a new skill, including learning to teach, requires grounded feedback (Brinko, 1993). Study 2 offered insights into this phenomenon by providing contextual instruction without any performance feedback. TAs seemed unclear as to their ability, and they exhibited inaccurate self-assessment. The tentative product should operationalize this finding by explicitly visualizing the variables that TAs should address. This might mitigate issues around overconfidence and underconfidence. To test this design objective, the tentative product should prompt TAs to speculate on their own performance before providing visualizations of what they actually did. This test of perceived accuracy avoids

hindsight bias. Allowing TAs to attempt this prediction over the course of the semester would likely show if their self-perception improved.

Study 1 revealed that providing grounded feedback through data visualization promotes curiosity and self-reflection. It also showed that some visualizations work better than others in terms of how TAs interpret and make sense of the data. This finding suggests a need for continued refinement of visualizations. Effective visualizations should be easy to understand and actionable. TAs reflection will reveal when these visualizations work, and efficacy will be evident in how they reflect. As the system provides feedback, it should also gather evidence that reflection occurred. Also, there may be different levels of reflection—some contexts will be richer than others, and users should be able to choose to give shallow or deep reflection.

Reflection-on-action helps to produce changes in teaching behavior (Gormally et al., 2014). This is a particularly helpful intervention when provided in concert with expert scaffolding (Bell & Mladenovic, 2008). The system can operationalize this by prompting the user to reflect on specific features of the data, or about their unique experience teaching this class. These prompts should highlight some critical aspect(s) of the data. For example, a training module that focuses on getting students to answer questions might prompt the TA to reflect on the ratio of questions that they asked vs. how many students answered. The TA's response should be gathered as a free-text interaction type after each visualization. This would provide the data to evaluate the presence, depth, and relevance of the reflection.

5.2.2 Timing of feedback

The time between action and feedback impacts learning. Retention of information is improved through active retrieval processes (Roediger & Butler, 2011). The exact timing of feedback in a PD context is unclear. Research in feedback on memory shows that a delayed response of an entire day can help improve longer term memory (Butler, Karpicke, & Roediger, 2007). There is some uncertainty regarding the exact timing, length, and frequency of feedback to improve teaching, but the PD literature suggests that longer delays diminish its effectiveness (Brinko, 1993; Ilgen et al., 1979). The system may operationalize this by contacting the TA shortly after each teaching session to review their data. The system can evaluate the efficacy of this approach by measuring the time that passes between the invitations and the user's viewing of their feedback. These response data might also provide hints about TA perceptions of engagement and relevance.

A single point of reflection within a semester-long course would not likely produce lasting behavior or belief changes. Effective PD offers teachers repeated opportunities for review and reflection over time (Brinko, 1993; Ilgen et al., 1979). To operationalize this notion, the system should provide repeated reflection opportunities after teaching sessions. In order to evaluate this design, metadata about user interactions could show whether the TA engages in repeated review and reflection. The content of the reflection would help to evaluate the presence of ongoing intentions to change. Data analysis could include an assessment of reflection over time, asking, for example, if the nature of the reflections remains consistent, or if it increases or decreases in level of detail.

Along with repeated reflection opportunities, data visualizations should show accumulated efforts over time. This allows users to perceive changes (or lack of changes) in their behavior throughout the course of the intervention. Unfolding, time-based reflection can support a growing sense of self-efficacy (Young & Bippus, 2008) or at least support self-reflection and decision making about future teaching actions (Duval, 2011; Govaerts, Verbert, Duval, & Pardo, 2012; Rivera-Pelayo, Munk, Zacharias, & Braun, 2013). The system can operationalize this by using visually coherent graphs that

help TAs see similarities between day-level statistics and compilations of statistics over time. To evaluate the impact of this design, interactions following accumulated data should allow TAs to indicate the presence of trends or changes through multi-select or free-text options.

Study 1 and 2 showed that TAs have very limited time. This could be one reason for low participation rates of the causes of low rates of volunteering for in training in the first place. If users perceive the new training technology as a burden, it could lower the impact of the tool, or require more external rewards or pressures to encourage use. If users are to adopt this new training technology, it must not be perceived as a burden. To operationalize this constraint, the system might distribute the training over time in small chunks or training modules. The length of training in Study 2 set a benchmark for ongoing design, yet its prototype modules had inconsistent and unpredictable lengths. Some TAs took as long as 20 minutes to complete a single session. Future designs should attempt to keep users online for a more consistent amount of time, attempting to keep users online for fewer than 10 minutes per session. There are at least two ways to evaluate this: (a) does it actually happen that users take fewer than 10 minutes, and (b) does this support improvements in teaching? In the former case, the system can gather a record of how long TAs take to complete each training session. Assuming the goal is to maintain the timing target, then the design should adjust its content as necessary to address average completion times. To address the latter, the current research should gather qualitative assessments of how users perceive the burden and efficacy of module timing. As the design solidifies over time, future research should perform randomized-controlled trials of actual efficacy.

5.2.3 Context

Participants in Study 1 exhibited little interest in using the discursive teaching tactics that I suggested during the final interview. In Study 2 I implemented several design changes, including direct instruction on the value of discursive teaching strategies, involving TAs in the generation of new tactics, and including research-supported suggestions that were sensitive to the TAs' context. Addressing why these ideas matter, including the TAs voice, and expressing expertise may all be good to include in the design of a system.

The tentative product should ask TAs to provide context about their teaching environment, and then reflect that knowledge back to them in following interactions. To test the success of this approach, the tentative product needs to gather evidence regarding users' openness to suggested tactics. This could follow a design which allows users to freely read about, ignore, or adopt suggestions related to new ideas. Their level of openness might correlate to the number of new tactics they select to try, how concrete their selected tactics are, whether the tactics are relevant to their individual classes, and whether or not they actually attempt the tactics. Interviews would be another valuable source of data about this feature of the training.

Research on feedback for higher education instructors should come from a trusted expert (Brinko, 1993; Henderson et al., 2011). The "voice" of the system, therefore, should inspire confidence. Referencing the empirical evidence behind each recommendation offers one way to operationalize this goal. When describing a specific tactic, the tentative product should summarize *why* that tactic helps students learn. Another way to increase trust in the system may be to offer feedback on concrete, measurable, behavioral data rather than directly focusing on behaviors with poor definitions or general abstractions. Evidence for the creation of a trusted "voice" may lie in whether users attempt to follow suggestions, and whether they report the suggestions as relevant for their own classes.

PD engenders trust for a consultant by treating the instructor as the expert of his or her own classroom (Wergin, Mason, & Munson, 1976). One way to operationalize this finding is to gather personal

reflections from the TAs about their own experiences as well as ask them for suggestions they might recommend. The tentative product should then take these comments from prior modules and reproduce them later as elements within a list of possible tactics to try. Successful operationalization might mean that users recognize their own contributions when they reappear. Furthermore, users might show a growing sense of agency in their teaching as they learn more from a trusted source (Bandura, 1977, 1997). Post-intervention interviews could potentially reveal if they saw and recognized their own suggestions and comments, if they felt that they had control over the methods of their teaching, how much agency they felt while selecting from an externally generated list, and whether they perceived the list as being from an external authority, from themselves as users/learners, or a combination of both.

Instructors often benefit from belonging to a community of practitioners with shared goals and values (M. D. Cox, 2004; Hill et al., 2008; Prytula, 2012). Allowing instructors to share ideas and solutions with each other may improve the chances that they will begin to experiment with novel teaching approaches (Bolen, 2009). The system may operationalize this by asking its users to make suggestions about how they typically attempt to address problems in the class, then share those back to the group. Evidence of successful collegial support may exist in whether TAs perceive and value the commentary from their peers.

5.2.4 Planning

Study 1 revealed that TAs had some sense of how they should interact with their classes. When they saw visualizations revealing high rates of TA talk, they talked about the desire to engage students more. At the same time, they also said that they did not know what they should do, nor what those rates should be for their particular class. They subsequently did not change teaching behaviors. During the spot checks early in the semester they described ambitions to include students more. In their final interviews, they generally justified their unchanged behaviors as the best they could do, or else the best one could expect from their class.

Study 2, on the other hand, gave TAs substantial support for planning their upcoming classes. Providing explicit advice about what to try helped avoid confusion about tactics. Providing accountability mechanisms helped to close the planning loop by giving TAs space to reflect on what their plans achieved.

The design of the training system should operationalize these findings and support various aspects of planning. These aspects include supporting tacit beliefs that students should speak up more, explicit advice on what to try, and help setting performance goals.

In Study 1, TAs were also uncertain of how to change. Subsequently, most of them did not. Few had any prior teaching or training experience. They may have attempted ineffective pedagogical strategies. Study 2 showed that direct advice can translate to new behaviors. The training system should provide explicit advice on how to change, delivering concrete suggestions of tactics to try. Evidence that these tactical suggestions are of use to TAs will emerge from whether or not they actually include the tactics in their practice. Additional data regarding their intuitions and beliefs about the strategies could come from including multiple opportunities to reflect within the system and in post-intervention interviews.

TAs did not have explicit performance goals in Study 1. They saw graphs revealing how much they and their students talked. They saw nothing about what “good teaching” looked like. The data visualizations should include goals directly in their visual design. The use of color or lines in a graph could signify behavior targets for the user. There are at least two ways to evaluate the use of these designs. Do TAs agree that the goals are meaningful, and do they attempt to reach them? In order to

evaluate the value users have for the goals, quantitative evidence should indicate whether TAs change behaviors over time. Qualitative evidence should indicate how they feel about the goals. A direct analysis might include asking TAs to describe their thoughts on the goals. It may be best to ask for these reflections after the intervention in follow-up interviews. Probing participants about these topics during the intervention might introduce an avoidable state of premature skepticism.

When it comes to using PD to help teachers plan, there can be tradeoffs in terms of useful support. Asking instructors to make explicit commitments to try specific strategies can help increase follow-through (Brinko, 1993). This is particularly true for novice instructors. Instructors in K-12 often need to feel a sense of autonomy over their own class in order to support their sense of self-efficacy (Locke & Latham, 2002; Pearson & Moomaw, 2005; Tschannen-Moran & Hoy, 2007). It might be counter-productive to promote behaviors that instructors do not believe in. To balance these concerns, the system might give specific goals to try, but let the user select whichever they think would be relevant for them. While preparing, instructors might see a list of various tactics, and then select as many as they would like to attempt. This provides the psychological benefit of allowing them to make their own selections. It may also produce implicit commitment to attempt the selected tactics. To evaluate this complex design choice, the research should track how many and which tactics users select, and it should ask how much autonomy users felt during the process. Data gathered in-app would address the first concern. Post-intervention interview data might produce the latter.

5.2.5 Timing of planning

In determining how to design this framework, it is important to consider when to support planning. The performative aspect of teaching is sufficiently demanding that instructors typically create a detailed plan for each class session. Following Study 2, it is clear that members of this population are likely to prepare for class anywhere between days before teaching up to preparing on the day that they teach. There is a gap in the literature regarding when instructors *should* be planning their pedagogical tactics. For the time being, the training should settle on a specific time to prompt users to log in and make plans. Over repeated applications of the intervention, evidence about the appropriateness of this timing should emerge. Specifically, the length of time between prompting users to plan and how long it takes them to login may help to adjust the timing.

5.2.6 Belief change

Studies 1 and 2 showed some cognitive roadblocks that TAs might face when they consider taking new risks and trying unfamiliar teaching strategies. These roadblocks have to do with what the TAs believe. The first challenge, emerging in the interviews of Study 1, is that a TA should believe that a suggested strategy is meaningful. This means that the suggestion must be practical and effective at producing a desirable result. If TAs are skeptical then they will be unlikely to take direction. The second major belief challenge, emerging in both studies, was that the TA must have some sense that they will be able to perform the action. A suggested strategy could make perfect sense to a TA, but if they lack confidence in their ability to enact it without embarrassment, they are unlikely to try it.

To address the first belief challenge, the system design should provide clear explanations of the overall strategy the training promotes. In the case of this research, the overarching goal is to bring students further into discussion with the TA and each other during class. The value of this activity is not necessarily self-evident to TAs. The training should take steps to point to the empirical basis of this approach to teaching. Each subsequent strategy toward that larger goal deserves similar treatment. It might be useful to remind TAs about the scholastic value of student-centered practices throughout the

longitudinal training. However, persuasion is a delicate challenge. It is important to explicate the risks involved in any new strategy, to limit the amount of theoretical exposition, and to emphasize the practical benefits (Tschannen-Moran & Hoy, 2007). It is also important to avoid directly condemning teacher-centered practices. Not only are these practices likely to be deeply internalized for many participants, there are certainly times when lecture and direct instruction are beneficial approaches to teaching (Hmelo-Silver, 2004). It would not help to make users defensive by undercutting those efforts.

There are several sources of data to gather while weighing the impact of this approach. The Approaches to Teaching Inventory is a validated instrument for measuring the attitudes teachers hold regarding teacher-centered and student-centered practices (Prosser & Trigwell, 2006). Users should take this instrument before and after the intervention. To frame the self-report nature of this instrument, the system should also gather behavioral data regarding the variables of interest. Actual enactment of suggested strategies would likely imply some level of agreement with the suggestion. Additionally, open-ended reflection prompts (in text and in person) at the end of the intervention may uncover complex beliefs the user holds. (Repeated reflection prompts may have an added benefit of helping the TA develop the ability to notice what they do while they teach.) At a qualitative level, observers in the field should gather notes about the specific tactics that the TAs enact.

Addressing the challenge of self-efficacy is more complicated, and requires a more nuanced approach. Self-efficacy for teaching is generally good for students (cf., Schwarzer & Hallum, 2008)., Study 2 revealed that TAs can hold inaccurate self-assessments. An instructor who overestimates their abilities may fail to recognize where they fall short. In such cases, self-efficacy is unlikely to support learning. Underestimating what they can do may hold a talented TA back from taking beneficial risks in teaching. The system should not necessarily be designed as a mechanism for simply increasing self-efficacy. Instead it should both be a tool for increasing accurate self-assessment, and a tool that helps to support a growth mindset toward changes in behavior. These changes would be reflected through increased self-efficacy.

To address the improvement of self-efficacy accuracy, the system should provide data visualizations of concrete in-class behaviors. To get the most out of this approach, the design should include a predictive element, asking TAs to recall their performance on each metric before seeing the actual results. This design element should help to avoid hindsight bias and increase accuracy for self-assessment over time.

To address the improvement of self-efficacy as a mediating variable for change, the training should deliver tactics that are measured in what the TA does rather than what the students do. It is likely true that an improvement in student outcomes would lead to an increase in TA confidence. However, this would require limiting the training to only address actions that TAs can easily produce from students. The truth is that meaningful changes in teaching practice are likely to take a long time to make. It is unlikely that such changes would occur for an unpracticed novice in the first semester of training. Therefore, the training should focus on actions the TA makes, and assessing whether or not the TA reached personal performance goals. This keeps the evaluation on the TA and internalizes their locus of control, an important element of building self-efficacy (Bandura, 1997).

Evidence for changes in self-efficacy are likely to be found through self-report in surveys and interviews. The Teacher Sense of Efficacy Scale (TSES; Klassen et al., 2009) is a validated instrument for assessing general self-efficacy for teaching. Users who engage in training should take the instrument before and after the intervention. Additionally, it would be useful to know users' confidence toward specific tactics. The system might ask TAs to predict their confidence in enacting

concrete actions. Post-intervention reflections and interviews are also a useful source for information on general and specific feelings of self-efficacy.

5.3 Building toward a tentative theory

DBR is largely about contributing to theory while impacting practice (Barab, 2014). It identifies and examines multiple interacting variables, producing "system-level understanding." To be valid, DBR requires evidence-based claims about learning. The prior section identified many of the relevant variables for this research domain, as well as the evidence each feature requires. To be meaningful, DBR must advance theoretical knowledge of the field.

As I mentioned above, building a tentative theory means specifying the overall picture of what I am trying to build in clear and potentially testable terms. As a starting point, I turn to the Interconnected Model of Professional Growth (IMPG; Clark & Hollingsworth, 2002). The goal of this empirical model is to articulate how training for teachers interacts with the teacher's beliefs, actions, and outcomes to describe multiple pathways of teacher development. Instructors are individuals with varied and different experiences. Their backgrounds and motivations are different, and their beliefs impact their approaches to professional development.

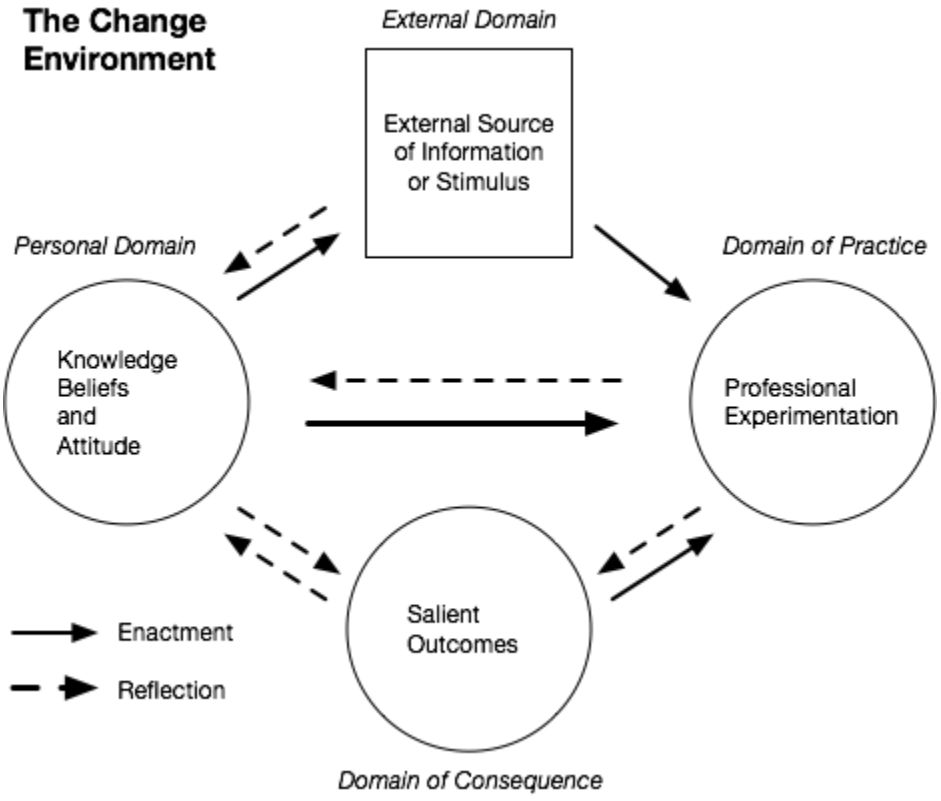


Figure 5.2: The IMPG is an empirical model that allows for multiple pathways of instructional development Clark & Hollingsworth, 2002.

The model describes the entire "Change Environment," which is the collection of four domains within which training and its outcomes emerge. *The External Domain* is the source of professional development, be it formal or informal. This might be a group seminar for or one-on-one observation

and feedback session. *The Personal Domain* is the collection of personal states that make up the instructor's approach to teaching. This includes what the instructor's pedagogical content knowledge, beliefs about teaching and students, and attitudes about teaching and learning. *The Domain of Practice* is the instructor's "experimental" space, which is usually the classroom. This domain is the learning environment. *The Domain of Consequence* is the collection of outcomes that emerge as a result of the instructors teaching "experiments," e.g., how students respond to enactments.

Enactment and reflection describe the types of interactions that occur between the domains. Professional growth emerges (or not) as domains interact with each other through a teacher's beliefs and actions. Reflection is any deep and meaningful consideration the instructor gives to concepts, events, and conditions within the change environment. The depth of reflection is typically related to the quality of a developmental experience in the external domain. Enactment is any attempt the instructor makes to put a new idea, practice, or principle into action. The source of these ideas also typically derives from the external domain. The model shows that reflection and enactments, regardless of their conceptual inspiration, frequently occur between the practical domains of teaching and not always directly from the training itself.

PD often works to impact the actions teachers take through a direct intervention on what they believe about teaching (McShannon et al., 2006; Tillema, 2000). There are multiple pathways of practical change, however. Those pathways might not depend directly on intervening directly on beliefs about teaching and learning. Different outcomes are possible depending on the ways in which instructors reflect and enact. For example, in one case (graph a), a TA reads about a pedagogical idea in a training manual, believes it is good, and then tries it in class. The TA notices a positive student response to the enactment, but does not need to update beliefs. This is an example of training leading to new behaviors via reflection, without impacting beliefs. In another case (graph b), a TA reads the same idea and tries it in class before being convinced that it is efficacious. Perhaps they are motivated to enact the behavior due to a sense of duty to the source of the training. In this case the TA notices a positive student response and then updates beliefs about the pedagogical idea. This is an example of training leading to new behaviors through enactment and new beliefs through reflection.

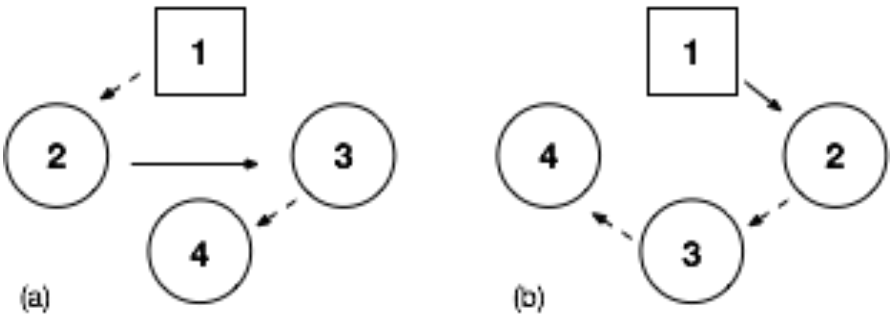


Figure 5.3: Different pathways of development.

These different pathways of development provide a useful way of explaining the different types of experiences instructors have while learning to teach. The IMPG outlines stages of transition, providing clues about how to deploy the tentative product. Teachers change after reflecting on (a) their enactment of new teaching strategies (the "domain of practice") and (b) what their students do in response ("domain of consequence"). The tentative theory should highlight these stages of development.

Applying the IMPG to the domain of STEM TAs, a major limitation for the population emerges. Because there is no training for these instructors, the external domain is devoid of any support for

impacting reflection or enactment. This is exacerbated by an institutional lack of incentive for reflecting on in-class teaching practices or student behaviors. This limitation cascades throughout the model, removing reflection and enactment opportunities as they relate to direct interaction with a training unit. Furthermore, the lack of pedagogical training limits the likelihood that TAs would reflect on events from the domains of practice or consequence, as they lack a mental model of what productive learning might look like in those domains.

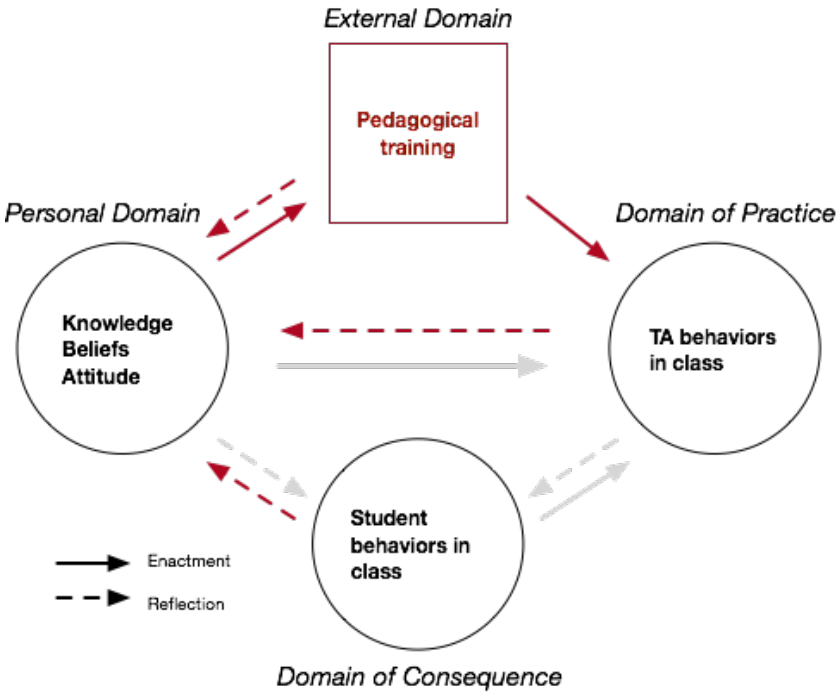


Figure 5.4: The red objects are those that do not exist for most current STEM TAs in higher education. The population is left with an impoverished and limited model of change.

Given this explanation, it would seem that introducing a training process would solve all the problems. In terms of analyzing the process of change for TAs, that may be the case. However, as I mentioned above there is no clear model of how to build a training for TAs that is sensitive to their unique environments. Seminars, workshops, and expert oversight do not address all of these gaps. This is what leads to the need for the current research. Therefore, a new theoretical model is required in order to direct the design of the training. The IMPG is limited in this regard. It does not clearly reveal the repeated cycles of learning and practice that teachers in general, or TAs specifically experience every day. It does not provide a roadmap for how to build SmartPD. For that reason, adaptations to the model are necessary in order to highlight (a) iterative practice cycles, and (b), the algorithmic aspects of design where the system becomes "aware" of what happens in each of the domains.

5.3.1 Compiling a new framework

This section describes the stepwise construction of a framework for building SmartPD artifacts. It draws inspiration from the IMPG as a general model of learning to teach while also including the need for repeated cycles of training. I do not call this new framework a model because it does not yet have any predictive mechanisms. It may be a proto-model, as future research could perform quantitative tests of its components.

To begin, I identify the basic stages of ideal TA practice. The relevant literature and the studies so far indication that an iterative framework should include the critical stages of planning to teach, the act of teaching itself, and reflection on the teaching interaction.

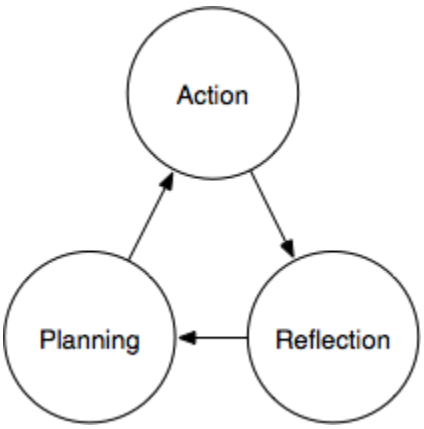


Figure 5.5: Setting the concrete phases of PD as an iterative cycle of experience.

Drawing from the IMPG, I then discriminate external environments from internal environments. External environments include those that describe events that happen in the world and can be described by objective data. These data comprise the domains of practice and consequence, as well as all of the instructor's enactments. The internal environment is made up of the personal domain (knowledge, beliefs, attitudes), as well as all of the instructor's reflections.

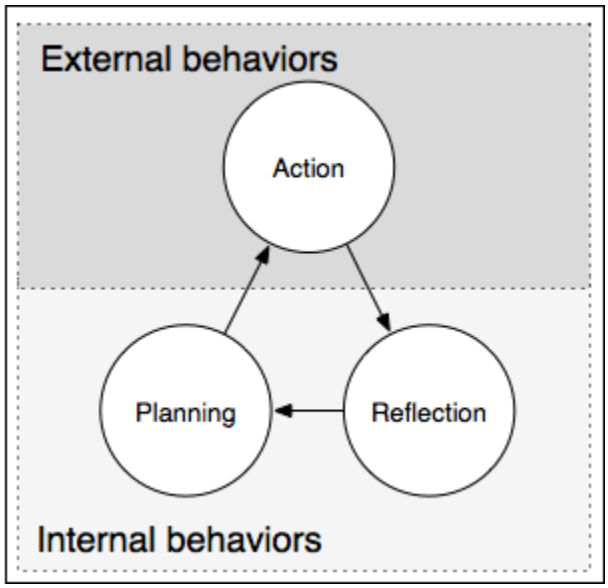


Figure 5.6: Differentiating between external and internal behaviors.

This approach will be useful in that it naturally lends itself to the use of digital sensors that detect objective, quantitative behavioral data. This is the "Smart classroom" section of Figure 5.7. An added benefit of this approach is that addresses a limitation in prior research in computational support for instructional development. Those studies typically focus on external behaviors. This framework acknowledges the value of including computational support for internal behaviors. This is visualized in the discrete boxes around planning and reflection in the "Training system" part of Figure 5.7.

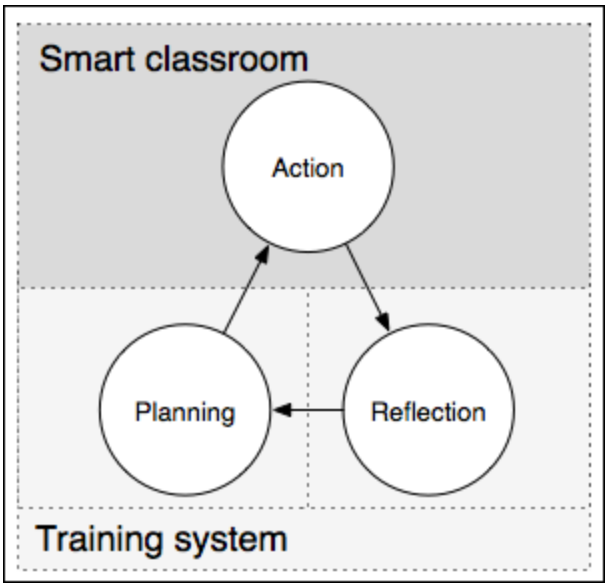


Figure 5.7: Distinguishing where to apply computational systems.

Within the discrete support stages for planning and reflection, the framework designates different types of support for each internal behavior, as well as an order for when those supports should appear (based on prior data and related work). M.1 and M.3 (Figure 5.8) signify unique scaffolds for reflection, an initial review of external behaviors prior to direct instruction on a pedagogical concept (M.1), and a post-action review after attempting to change the Domain of Practice (M.3). M.2 signifies scaffolded planning for each upcoming action opportunity, i.e., preparing for class

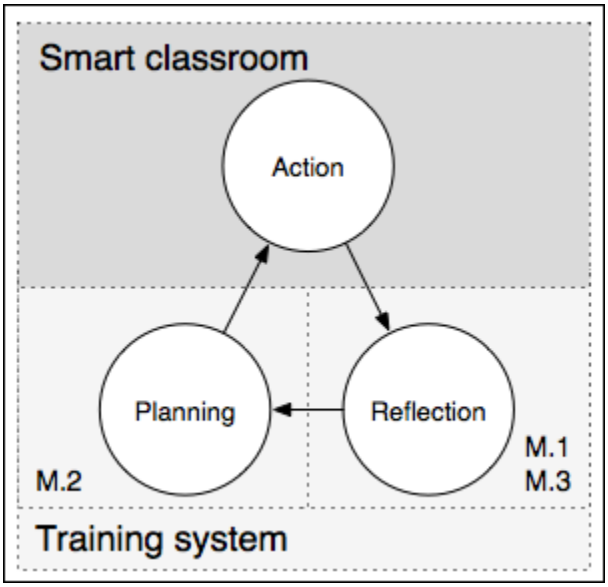


Figure 5.8: Outlining the process of training.

In the studies that follow, I will show how this general framework can apply to specific training topics and teaching environments. When using this as a guide, curriculum designers must answer specific questions about how to address their particular design goals. For example, it outlines the order of

training routines (M.1-M.3) as they relate to events within the environment, but it requires the designer to set the frequency based on the regularity with which the trainee encounters the Action phase.

The framework is designed to be applicable to any training topics. A pass through each module within an entire training routine should cover a single concept. In the next chapter I will describe a specific example of how to design each module for a given topic. Interested designers should be able to apply those steps to their own pedagogical concepts and instructional domains.

The framework outlines when researchers should support reflection and planning. As such, it can also help identify output variables of interest. This helps to envision the design of the necessary components to a smart classroom that addresses the particular needs of the research. Recall that for the current research, I address discursive teaching as an entry into helping TAs learn to engage students. Reviewing the operationalizations mentioned previously, consider the following list of variables. These outcomes, categorized within each domain of the IMPG, function as indications of intervention efficacy:

- External domain (the training itself)
- Domain of practice (what the TA does in class):
    - Asking questions, asking more questions, asking deeper questions
    - Waiting longer after asking asking questions before speaking again
    - Using the students' names
    - Calling directly on students to answer questions or share perspectives
- Domain of consequence (what the students do in class):
    - Responding after TA asks a question (Answering questions, asking clarifying questions, etc.)
    - Making statements (clarifications, elaborations, explorations, explanations, etc.)
    - Increasing length and frequency of spoken turns
    - Raising hands
- Personal domain (the knowledge, beliefs, and attitudes of the TA)
    - Changes in self-efficacy and attitudes about learning (TSES and ATI)
    - Changes in ill-defined attitudes as shown through repeated reflection prompts

Flowing from this framework and these variables, I can now build a prototype intervention that provides conceptual support for new ideas and performative support through reflection and planning. The IMPG outlines the various pathways that an instructor can travel throughout a PD regimen. My new framework shows how an iterative training program within the context of a smart classroom might be designed. In the following chapter I will apply the design goals mentioned so far to the new framework to produce a testable SmartPD artifact.

Chapter 6: Testing the framework, including data visualizations (Study 3)

The following study tests the framework described in Chapter 5. Following the DBR process, the phase of the research at this point is to test an artifact prototype and assess preliminary products and theories (McKenney, 2001; Wademan, 2005). This work tests the design goals outlined in the PAR-TS framework for SmartPD. I evaluate those design decisions based on evidence within subjects where I use "Point (Quantity) Induction" and "Line (Quality) Norms" as my analytical framework (McKenney & Reeves, 2012). The explanation of these terms follows in the method section of this chapter. In short, these methods are appropriate for exploratory work where the research should remain open to what *might* happen with the given design for the sample of participants. This is different from determinative research that attempts to explain what *would* happen with a given design for a specific population.

6.1 Design of the artifact

This section describes the development of the deployed system. It is an in-the-wild simulation of a prototype commercial system. Consider it a "good enough" system that simulates technologies that might exist in a future smart classroom in the service of a dedicated training application.

There are two parts to the system: the TA-facing artifact that performs the feedback and planning functions, and the behind-the-scenes behavior tracking system. As in Chapter 4, I once again used *TA Dashboard* to simulate a smart classroom. RAs gathered these data while TAs conducted class, producing immediate counts of classroom events such as spoken turns and the presence of TA questions.

The TA-facing artifact is a simulation of an app that would offer training and feedback to TAs while supporting reflection and planning. As with the system in Chapter 4, I built this series of interactions in the commercial platform Qualtrics. There are some limitations incurred by using a commercial platform rather than one built in-house. For example, data types are defined and enforced by Qualtrics, and they are stored on inaccessible servers. At this stage of development, however, the benefits of using a polished development platform, even if external, were higher than the costs. It was very easy to build each of the training modules quickly because every interaction can, more or less, fit within a survey framework. Furthermore, the platform kept overhead costs low in terms of building a working interaction space. There is also a practical benefit to using a platform that is usable across multiple platforms, such as desktop and mobile. For all these reasons, this is a good approach for testing a brand-new framework.

Following the PAR-TS framework, the app simulation guides users through an introduction to a pedagogical concept and initial exposure to baseline performance on relevant behaviors (M.1). It grounds the user's reflections in actual in-class behaviors (from the smart classroom simulation). In the later reflection module (M.3), the prompts are also grounded in things the TA said in M.1 and M.2. The release of each module is timed to support iterative cycles of planning, actions, and reflections.

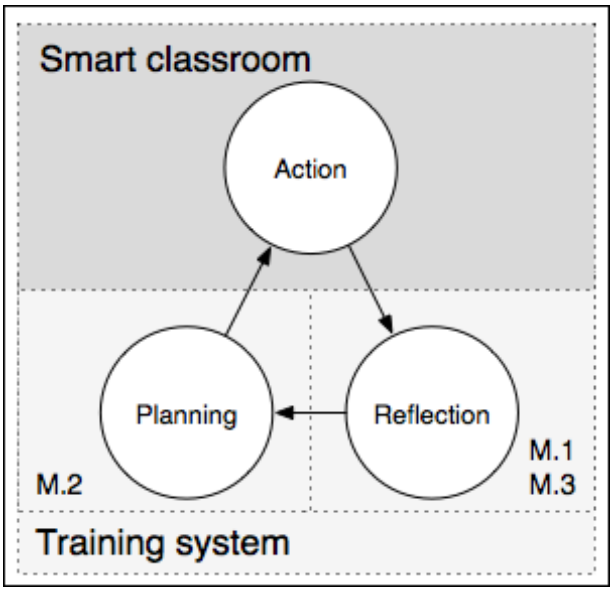


Figure 6.1: PAR-TS Framework (reprint of Figure 5.8).

6.1.1 Action and Reflection

From a researcher's perspective, the variables of interest that influence the artifact design draw from those explained in Chapter 5. Some of the most important include in-class behaviors that relate to discursive teaching practices, the perspectives participants have about teaching and learning, and any changes that occur in these variables over time. From a user's perspective, only a subset of the total collected variables are relevant for designing the reflection goals. These are a small number of variables that highlight the overall goals of the training (e.g., TA/Student talk ratios), as well as the specific goals of each unit (how many questions the TA asks, when, and of what type).

The TA training in this study continued to focus on discursive teaching practices. It began the first training interaction by describing the concept of eliciting student participation in class, and how doing so improves student outcomes, i.e., when they talk more they tend to do better. Following this explanation, users saw a visualization of their most recent class and how much they had talked. Study 1 showed us that this visualization worked well for calling to attention a limited amount of student participation.

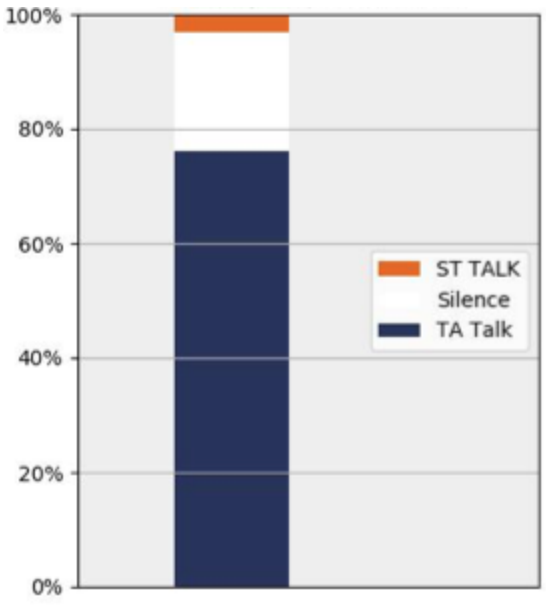


Figure 6.2: A participant's first view of talk ratios (M.1).

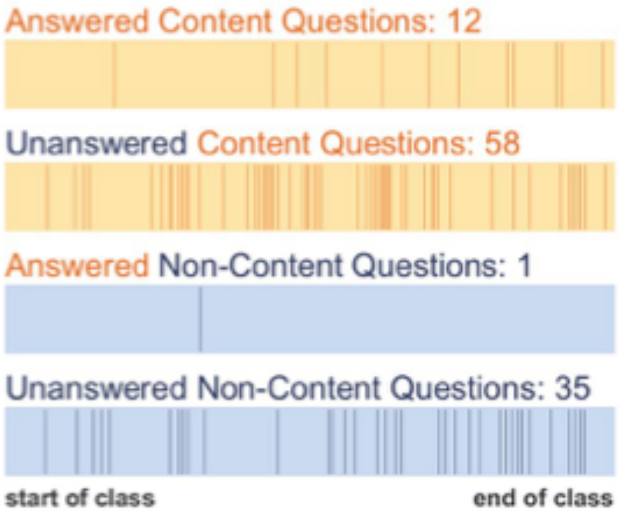


Figure 6.3: A participant's first data exposure for questions asked and answered. This visualization shows the number of questions the TA asked in a single class.

I use ratios of talk between Students and TAs to operationalize the overall goals of the training, i.e., to get students talking more and TAs talking less. This visualization shows how much a TA spoke during a single class relative to how much students talked. Silence included the amount of time spent in Wait Time, writing on the board, or other silent activities. It did not include independent work time, where students studied alone or in groups.

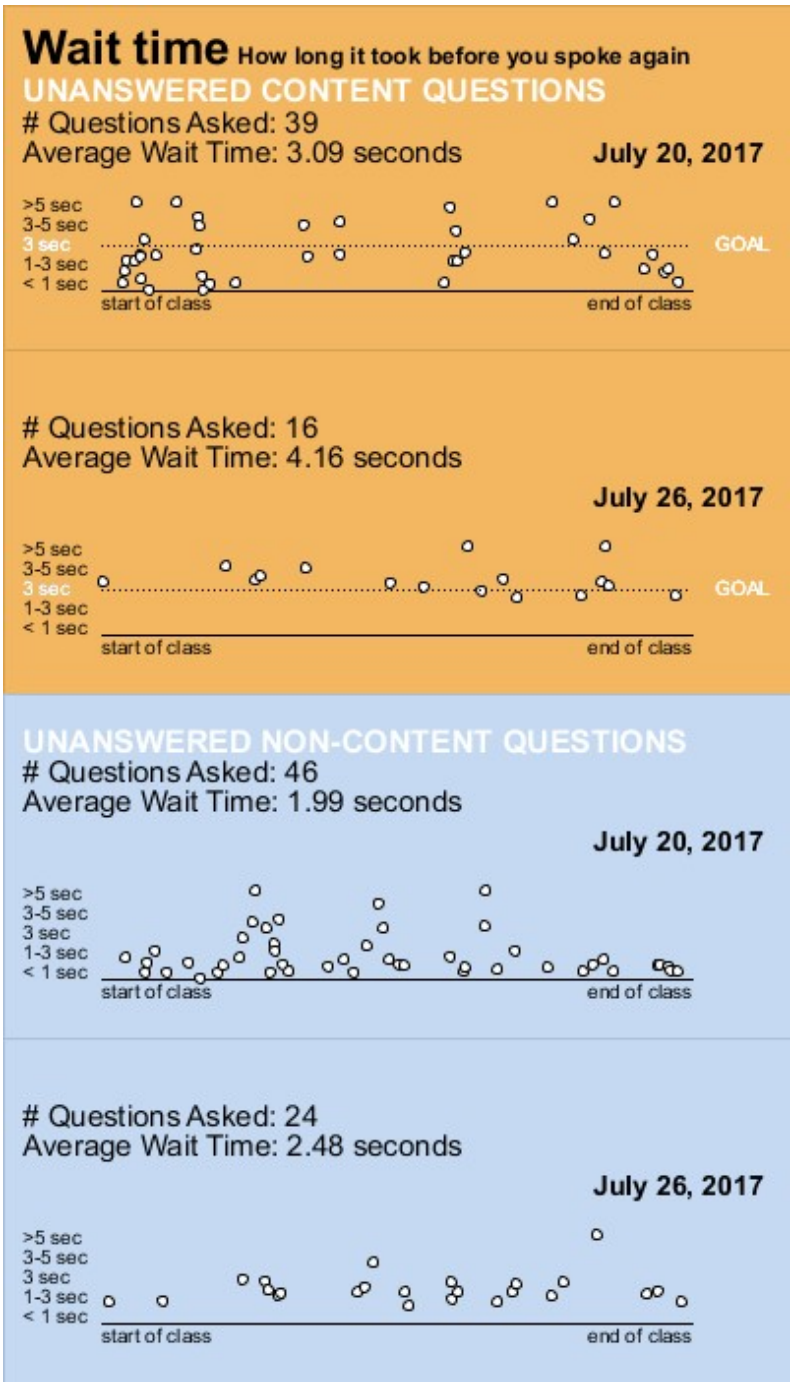


Figure 6.4: Visualization sample of Wait Time in Unit B.

The training itself comprises multiple conceptual units around this theme. A sequence of modules from M.1 to M.3 make up a single conceptual unit of training. Unit A in this study explained the use of content questions during class and how they were important to use instead of just non-content questions. In this study the participants saw not only the number of content/non-content questions they asked, but also how many were answered (see Figure 6.3). Unit B addressed the use of Wait Time, and how it is important to give students enough time to answer questions. As shown in Figure 6.4, content questions have a 3-second goal line, whereas non-content questions have no specified goal. This

emphasizes that content questions are part of a productive approach to questioning, and de-emphasizes the importance of non-content questions.

Following each of the visualizations, participants were asked to write a text response in support of reflecting on their data. Questions included:

- "Do these numbers align with what you thought was happening?"
- “Are you waiting as long as you thought? Longer? Shorter?”
- "Are you surprised by the numbers? Why or why not?”

Participants responses to these questions were stored for later analysis using “Line Norms,” as described in the data analysis section below.

Following these data reviews, participants answered questions about what they would try to do if they wanted to change the highlighted behavior. Their text responses to this prompt produced some of the suggested strategies that they would see in their planning module, described in the next subsection. The TAs returned to these themes in the last module of each Unit (M.3) after their next teaching opportunity. The final module issued prompts for reflecting on the outcomes of the unit-level suggestions and goals.

6.1.2 Planning

Following the initial reflection module (M.1) and preceding the next teaching opportunity, the TA received access to a planning module (M.2). Study 2 revealed that TAs typically prepared for class 1 or 2 days before teaching. Therefore, these modules arrived about 48 hours before their next class.

The first step of the planning module was to ask TAs what their learning objectives were for the next class. Specifically, they answered the question: “Before we get started, what are some things you want your students to be able to know or be able to do by the end of your next session? What's the purpose?” Study 2 revealed that some TAs valued this task as it helped them to summarize what they were going to teach in the next class.

Following the learning goals, participants see a list of tactics designed to address the unit topic. The prompt was, “For your stated goals, which, if any, of the stated strategies make sense for you and your class?” This is followed by a multi-selection interaction for making concrete plans for actions to take in the next class (Figure 6.5).

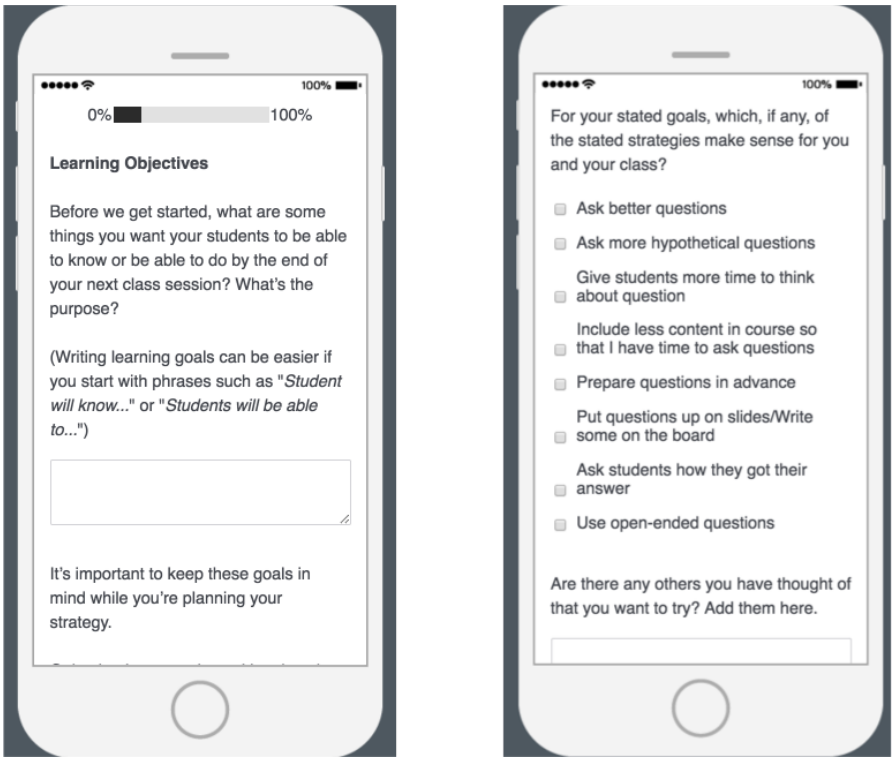


Figure 6.5: Some sample pages from one of the planning interactions.

In order to support commitment to the goal, the last questions of the training ask TAs to predict their performance on the unit metric. They are also asked to report their confidence in reaching that number. Note that in the following reflection module (M.3) they have an opportunity to recall their selected tactics. Their attempt to remember is logged as a recall opportunity for later accuracy analysis.

6.1.3 A sample encounter with the first unit

This section describes a typical user's experience with this study. After volunteering for to participate, the TA signs the consent form and fills out a pre-survey. The survey gathers basic information about teaching experience and self-report instruments for self-efficacy, and approaches to teaching. After successful study recruitment, an observer begins attending classes and gathering baseline data about in-class behavior metrics through TA Dashboard. Observations continue for as many class sessions as possible throughout the semester.

The first invitation to interact with the training arrives after the baseline behavior measures, usually 2 or 3 class sessions. This is a personal email from the researcher letting the participant know that their training is ready to begin. The first module, called "Module 0," is a simple introduction to the training with a description of discursive teaching, what to expect in the study, and an introduction to the Qualtrics format. Shortly thereafter TAs receive a new invitation to complete Module 1 for Unit A.

The following outline summarizes the design of a single unit of the training. Each step is repeated in the subsequent training, Unit B, which addresses Wait Time.

1. **Unit A/M.1**
    a. Within 1 hour of the completion of the participant's next teaching session, M.1 arrives

b. The TA sees an overall measure of classroom participation in the form of a stacked ratio graph of TA talk, silence, and student talk
c. The TA is asked to recall the number of questions they asked today
d. A graph appears to reveal the number of actual questions asked (and answered), separated by whether they were content-based or non-content, with definitions of those terms
e. The app asks the TA to note any discrepancies between what they thought happened and what the data reveal
f. The app asks the TA what they might do differently if he were to want to change their results the next time
g. The app offers the TA access to supporting research papers about how to improve questioning, but does not require an affirmative response before closing

2. **Unit A/M.2**
    a. M.2, oriented toward planning, arrives two days before the next teaching session
    b. The training reminds the TA of the overall topic, and provides a goal to increase content questions and reduce non-content questions in future classes
    c. The app presents a list of representative responses to the question from M.1 ("What would you do differently") and provides a list of research-supported tactics that address the same goals
    d. Tactic options include, "Ask more questions," "Prepare questions before class," "Put questions on slides," etc.
    e. The TA is asked to select any number of provided tactics that they would be willing to attempt to enact in the next teaching session
    f. The app asks the TA to produce a content-level learning objective for the coming class
    g. The app asks how confident the TA that they will be able to enact each of the selected tactics
3. **Unit A/M3**
    a. Unit A M.3 arrives within 1 hour following the next class session
    b. The app asks the TA to recall how many questions they asked before it reveals the data
    c. The app asks the TA to point out any discrepancies between their recall of what happened and what the data show
    d. The app asks the TA to reflect on what was effective and what could have been better
    e. The app enters a multi-path response loop for each listed tactic:
        i. "Did you try [tactic]?"
            1. If yes then "Did it work?" and "Who gets credit for the success?"
            2. If no then "Why not?"
        ii. Each question has pre-populated response options, e.g., options for "Why not?" include "It was too hard," "It wasn't relevant," "I didn't know how," etc.

A similar design follows for Unit B. Following the conclusion of B/M.3 the intervention comes to an end. The researcher then asks the TA to come in for an interview.

6.2 Methods and implementation

For this study I recruited a small number of TAs in order to gather initial data about the fit between the framework and the protocol design. 5 TAs from three departments at Carnegie Mellon University responded to the recruitment effort and were available for observations. There were four PhD student instructors and one undergraduate. They all happened to be male. Each taught either a recitation section or a stand-alone course. Average attendance per class was about 20 students. The classes were

heavy in procedural knowledge (i.e., math, engineering, and accounting), all of which traditionally follow a teacher-centered pattern of instruction but would benefit from student participation (Rocca, 2010). The research team observed a total of 80 class sessions totaling 89 hours of observation across the 5 instructors, and collected approximately 150 individual interactions with the system.

The team conducted a set of baseline observations of about 6 sessions per participant before they were offered any interaction with the training system. TAs then received their first prompt to log into the Qualtrics-based training system. Participants interacted with the system seven times throughout approximately three weeks of intervention. Each training session was short (no more than 10 minutes). Sessions that included reflection arrived within 1 hour after teaching. Sessions which involved planning arrived 2 days before the next teaching session. These timings co-occurred with the TAs' class review and preparation schedules.

6.2.1 Data collection

Observers gathered data about TA and student classroom behaviors using TA Dashboard (Chapter 4). The tool allowed them to record the length and order of speech turns (instructor or student), the pause between speech turns, the presence of questions (from instructor or students), the type of question (either content or non-content), whether a question was answered, the use of group work, the number of students in attendance, the ratio of students who spoke, the number of times students raised their hand, and the number of times the instructor called on a student.

TAs used the training tool and produced data about their reflection (M.1/M.3) and about their planning (M.2).

Using free response text inputs in the reflection modules, the system asked TAs to recall details of their most recent teaching performance before showing them their actual data. Studies 1 and 2 showed that this population of TAs can have inaccurate assessments of their real performance, often overestimating their student engagement. The design goal here was to force TAs to put a specific number to how many questions they had asked or how long they thought they had waited after asking questions. The "pre-reflection" was meant to overcome any hindsight bias that might emerge from taking TAs directly to feedback.

Following the guesses about how well TAs thought they had done, the system revealed their actual performance on the variables of interest for the given Unit. It then presented prompts for reflection through the use of free text response interactions in conjunction with these visualizations. This way the system gathered qualitative evidence of how TAs responded to visualizations immediately following feedback. Such reflections not only allow TAs to think about the teaching experience, but it helps the research team determine various useful analyses, e.g., whether the TA articulated any critical self-reflection, or whether the TA noticed when their performance prediction was different from the actual results.

The planning modules used multiple-selection interactions and asked TAs to pick from a list of suggestions for tactics to try in their next class. Examples of tactics the TAs saw during the first Unit (on questioning strategies) include, "Write questions up on the board rather than just say them out loud," or "Count to 10 silently after asking a question" (Freeland, 2007).

Participants filled out a teaching experience survey at the beginning and end of the study. These self-report instruments gathered pre- and post-measures of teacher-centeredness (the attitude that students need only direct, didactic instruction) via the Approaches to Teaching Inventory (Prosser & Trigwell,

2006), and self-efficacy for teaching via the Teacher Sense of Efficacy Scale (TSES; Klassen et al., 2009). I use these instruments to operationalize beliefs (about self-efficacy) and attitudes (about approaches to teaching) for the duration of this thesis.

The research team interviewed each TA for 1 hour at the end of the study, followed by an interpretation session in which we produced summarized notes for each idea expressed in the interviews (Beyer & Holtzblatt, 1998).

6.2.2 Data analysis methods

To analyze beliefs, attitudes, and perspectives I used the surveys and the interview data. For surveys, I scored the instruments and compared pre- and post-scores for each participant to see if individuals signaled any potential changes in beliefs. I conducted each interview with a silent research assistant who took hand notes during the conversation. Following each interview, the research assistant and I held an "interpretation session," and produced summarized notes for each idea expressed by the participant. I labeled and segmented these ideas. After conducing all of the interviews I randomized the labeled, segmented ideas and produced an Affinity Diagram (Beyer & Holtzblatt, 1998). This is an in-depth, iterative modeling approach for surfacing general themes across participants through a process of inductive qualitative clustering. In this case I spent 10.5 building, reviewing, and rearranging the notes and their collective themes. The goal in each case was to reveal unexpected findings about TAs' impressions of their experiences during the observations and interventions.

To analyze in-class actions I calculated discursive teaching behaviors from the class logs (i.e., TA Dashboard output). These data included information about TAs and students at the class level, e.g., number of questions asked and answered per session. I also gathered meta-level data about how TAs used the system, e.g., the length of time between sending each prompt to use the system and when TAs logged in. I used exploratory data analysis (EDA) to uncover behavioral trends at an individual TA level, both across each TA's teaching session, as well as averaged between periods of teaching during baseline and teaching during the intervention. This helped to surface general trends in how a TA taught, what possible impacts may have emerged of training, and to surface unique session-level outliers.

To analyze reflection, I used log data from Qualtrics to assess TAs' reactions to feedback they received in the simulated app. Quantitatively, I calculated across sessions the recall and precision for TAs in terms of how well they remembered what they did in their most recent class and which tactics they had selected in the app. Qualitatively, I read through each of their responses to the prompts and looked for depth of reflection and changes in depth over time.

I combined each dataset and conducted a mixed-method analysis across every dataset resulting from the field trial. The goal was to form a holistic picture of each participant's experience.

As in the previous studies, I used a process of structured reflection for analyzing data within a DBR context. I used "Point (Quantity) Induction" and "Line (Quality) Norms" as the analytical frameworks (McKenney & Reeves, 2012). Following these processes, the research team identified specific data points or sequences of events that revealed an unplanned insight or new hypothesis. We then discussed the context of the finding and formulated interim hypotheses and questions. We then returned to data in search for answers, and to produce more questions. We repeated this process for each participant until we reviewed every piece of data. The most informative answers and theoretically interesting questions constitute the bulk of our findings and discussion, where we explore three emergent themes of how participants interacted with the intervention.

6.3 Analysis

The intervention seemed to work in that TA behaviors changes throughout the course of the training. It did not seem to work in that there were no clear signals of TAs going through cognitive changes about what constitutes quality teaching. However, a clear measure of that particular variable would take a very long time to emerge under even the most rigorous and balanced study design. This work follows a "good enough" philosophy in order to test the face validity of the framework and justify ongoing development of the overall training approach.

6.3.1 Framework findings

In general, the TAs all used the training as designed. They engaged in a structured cycle of action, reflection, and planning. None abandoned the system. They typically responded to each emailed prompt to use the system within one or two days, at which point they completed the module. All participants described the system as "useful" for improving their teaching. TA1, TA3, and TA4 all exhibited positive change over the semester. TA2 and TA5 exhibited ineffective teaching strategies from the start, and these persisted throughout the semester.

In terms of initial reflections (M.1), TAs expressed curiosity and surprise when viewing visualizations of in-class behaviors for the first time (M1). They generally seemed to appreciate the objective nature of the data. TA1 and TA4 appeared to rationalize what they viewed as negative aspects of their data by explaining some challenge they faced in class. In answering the reflection prompts, TA1 focused on a high number of unanswered questions. He shared that he had a "bad habit" of "chaining" several questions together, which seemed to not offer the students the time or opportunity to answer. TA4 focused on his first visualization, which showed that he talked for 75% of class time and the students only talked for 5% of class time. He explained that this might be happening because he tried to "prioritize delivering all the materials on time..." He felt more pressure to cover all of the content as opposed to creating time for students to participate.

When viewing the data again (M.3), all TAs acknowledged the validity of the data and tried to explain some aspect of it. Part of the training asked TAs to review their attempts to enact new tactics. Some TAs expressed positive self-evaluations of themselves as instructors, pointing to increases in student participation/responses or describing their skill with a new tactic. For example, TA4 shared that preparing questions before class helped him ask more content questions than he had previously been asking. In the interview, he shared that he continued this practice throughout the rest of the semester. TA1 shared that he "paid more attention to what [he] was saying question-wise," as he began to notice going "on autopilot." TA4 reflected on the fact that he was not able to increase the percentage of student talk. Interestingly, he gave himself credit for trying. "I set goals for myself after seeing the data, like when I saw the data where I waited too short. So, I set the goal to wait longer."

Not all the TAs felt they had improved. When prompted to recall their goals, neither TA2 or TA5 could recall the techniques they had planned to try. When asked to reflect on his goal attainment in Unit A/M.3, TA5 stated he could not remember the class in question. The system prompts TAs to review their data soon after teaching. However, in these cases, TA2 waited 3 days before logging in, and TA5 waited 9. Later, in Unit B, TA2 reflected that waiting longer after asking questions, "Did not seem to make any particular difference, as students could not answer the question anyway." TA5 reflected that he should probably wait longer in the future, despite waiting longer than the target of 3 seconds. He seemed to fail to notice a larger issue, that he asked very few questions.

In terms of planning (M.2) and taking action in the classroom, TAs were response to the support for goal-setting. When TAs saw a list of tactics related to the Unit goals, they each selected at least on to try in their upcoming class. During Unit A, all TAs increased the number of content questions they asked in those following classes.

TA1 selected the most strategies during the planning stage, setting goals to slow down during question asking, ask a wider variety of question types, and wait longer before giving hints or answering a question. TA4 set a more conceptual goal, describing how the questions should support the need to cover material. TA3 said he planned to ask, "a few more questions than last time." After planning, TA1 avoided rapidly repeating/rephrasing questions, and his students responded to a greater ratio of the questions he asked. TA4 increased his use of content questions and reduced non-content questions. TA3 increased his use of content questions, and this persisted for the rest of the semester.

Alternatively, TA2 and TA5 showed only isolated teaching improvement. TA2 jumped from 1 content question at the beginning of Unit A to 19 after the relevant planning module. TA5 went from 6 to 15. Both then reverted to very low numbers of content questions after a single class. Goal-setting for these TAs seemed vague or disconnected from their practice. TA5's only goals in Unit A/M.2 were to "Ask better questions" and "Give students more time to think about questions," (despite already averaging over 3 seconds of wait time). TA5 never explained the dramatic rise and drop of content questions. Importantly, during the class where he asked 15 content questions, he averaged only 1 second of wait time after each question, and no students responded to any of the questions.

6.3.2 Interaction Types: Learning and Behavioral Changes

Most of the TAs had no training to teach prior to the intervention. Unsurprisingly, initial observations revealed that they relied on information transmission via lecture and the use of shallow questioning strategies.

Over the course of the training, most TAs became better at predicting what they had done in class prior to seeing a visualization of their data. They had five opportunities to do this (three times for number of questions asked and twice for length of wait time). Four of the five TAs started with a low accuracy for reporting how many questions they had asked (between 7% and 55% recall), and improved by their third attempt (between 64% and 80% recall). The exception was TA2, who accurately recalled asking 1 question on the first and third opportunity.

TAs were responsive to requests to set goals and enact new behaviors in class. Each set and achieved goals for at least one strategy (more content questions or more wait time), and used the app to reflect on their attempts. They evaluated their success (or lack thereof) by reporting either on students' responses to their new actions, or by describing their skill at trying the new behavior.

Beyond these commonalities, analysis of the surveys, log data, and in-class behaviors surfaced three themes in terms of "interaction types":

Productive self-doubt: TAs who acknowledged suboptimal teaching strategies and provided explanations. Goal-setting to address specific concerns. Tactic openness and cautious optimism. Repeated changes in practice. Mixed student measures. Increased self-efficacy.

Confirmatory assessment: TA who acknowledged suboptimal teaching strategies but gave no explanations. Goal-setting to address specific concerns. Tactic selectivity and confidence. Repeated changes in practice. Improved student measures. No change in self-efficacy.

Shallow dismissal: TAs that did not acknowledge their suboptimal teaching strategies. Vague or general goals disconnected from specific context. Tactic fixation and lack of confidence. Erratic changes in behavior. Decreased student measures. Decreased self-efficacy.

6.3.2.1 Productive self-doubt: Critical self-reflection, positive changes

Throughout the cycles of teaching, reflecting, and planning, a theme of critical self-reflection emerged for some TAs. When presented with data about their teaching performance, they each worked to explain some challenge they had faced in class. They selected strategies that would address what they viewed as problems with their teaching behaviors. They then enacted those plans and altered their teaching approach. These TAs expressed "productive self-doubt" in that they voluntarily focused on the limitations of either their own actions or their teaching environment, and they used this reflection to generate achievable, concrete goals that would address their concerns. This theme was typified by the experiences of TA1 and TA4.

This interaction type exhibited critical self-reflection. They began the semester with high numbers of questions and very low student participation. During Unit A/M.1, their visualizations included representations of unanswered questions and non-content questions (see Figure 6.6). In response to the data visualizations, each TA answered the reflection prompts by explaining their assumptions of what had gone wrong. TA1 focused on the high number of unanswered questions, interpreting this as a result of going "too fast," and describing a "bad habit" of "chaining questions" rather than giving students time to answer. TA4's first view of the data, on day 8, revealed a low percentage of student talk (Figure 6.7). TA4 reflected that he had been trying to "prioritize delivering all the materials on time..." and how this makes it, "...so there isn't a lot of room to engage with the students."

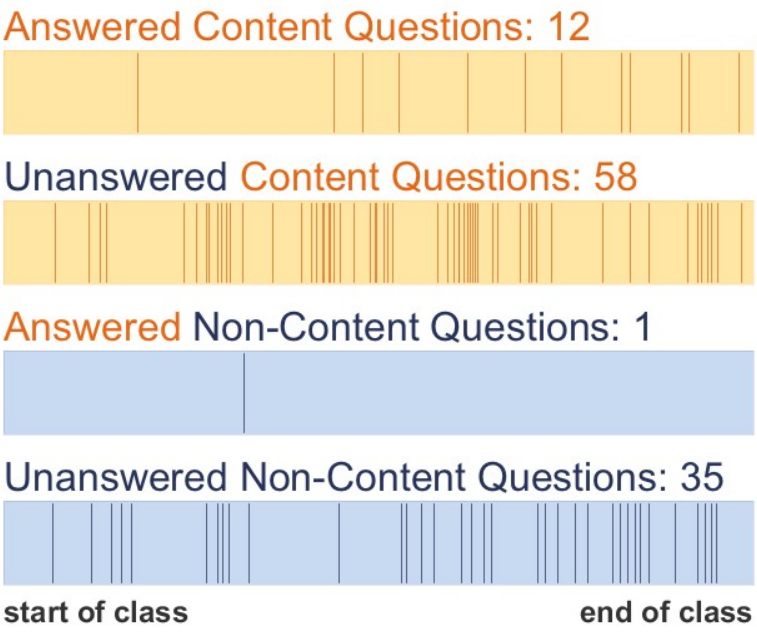


Figure 6.6: TA1's first exposure data for questions asked and answered.

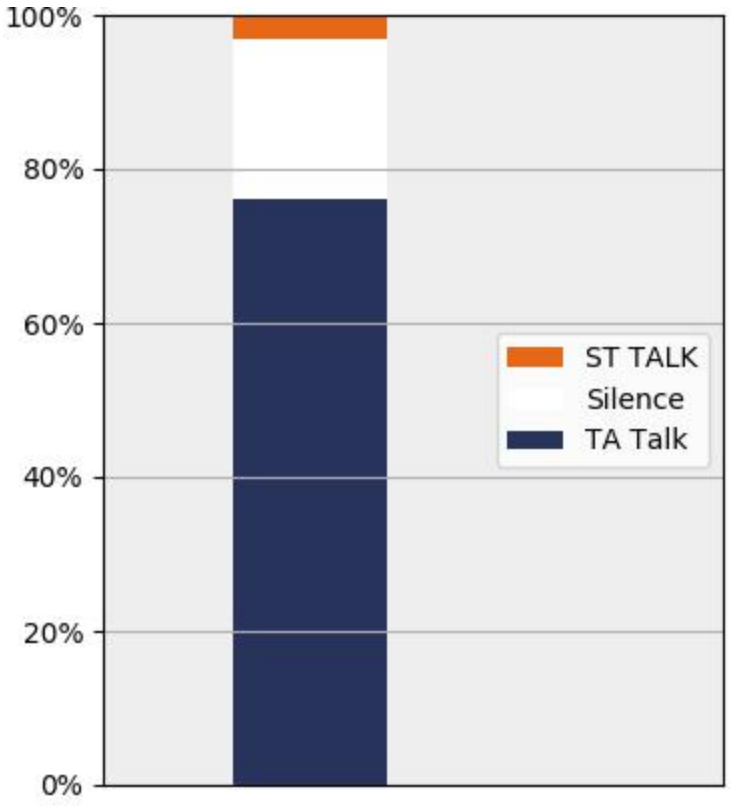


Figure 6.7: TA4's first view of talk ratios.

This group set goals to change the results of their teaching, and achieved their objectives. In Unit A/M.2, TA1 described an intention to slow down during question asking, ask fewer questions, ask a wider variety of question types, and wait longer before giving hints or answering the question himself. TA4 described how the questions should support the need to cover material. He selected goals for asking a wider variety of question types, preparing questions in advance of class, and putting questions up on slides. Following A/M.2, TA1 decreased the number of unanswered questions and avoided rapidly repeating/rephrasing questions. He asked fewer questions overall than he had previous, but he enacted his goal and increased wait time to about 3 seconds per content question. TA4 increased the number of content questions and reduced non-content questions. He reported that enacting a goal of preparing questions before class helped. His wait time had always been above three seconds, but it increased even more as he moved into Unit B.

In reviewing data following enactment opportunities, TAs who had originally exhibited productive self-doubt noticed concrete changes in their own behaviors. Comparing visualizations for talk percentage from before and after goal-setting, TA4 said, "I'm a little disappointed that I wasn't able to encourage students to talk more in class. However, I'm glad that the orange bar [indicating student talk] on the third day improves compared to the second day. I intentionally asked more questions on that day." TA1 noted that students' answers had improved, reporting that classes following goal-setting had gone well. "...not all of my questions were answered, but the answers I did get were great!" His students showed improvement as well. As his wait time got longer, the percentage of questions his students answered increased.

At the end of the intervention, users in this interaction type showed an increase in self-efficacy. TA1 succeeded in waiting longer and, possibly as a consequence, achieved a higher percentage of student

responses. This seemed to lead to a sense of increased ability. For example, he told us that over time he "paid more attention to what he was saying question-wise," because he became more aware of when he goes "on autopilot." TA4 focused on the most concrete of his selected strategies: preparing questions before class and putting them on the slides. This seemed to encourage his sense of change, but did not produce the higher student talk that he had hoped for. Regardless of this outcome, he successfully enacted his goals. For example, "I set goals for myself after seeing the data, like when I saw the data where I waited too short. So, I set the goal to wait longer."

6.3.2.2 Confirmatory assessment: High confidence, little reflection

The confirmatory assessment interaction type describes TA3, the only participant who had previous instructional training; a semester of instruction on collaborative learning techniques. When presented with data about his teaching, this TA either produced terse reflections that the data were unexpected and informative, or else that the data were irrelevant. In neither case did he attempt to explain his reasoning. TA3 enacted a plan to address the unexpected data by increasing the number of content questions during Unit A, but dismissed suggestions to wait longer during Unit B.

The confirmatory assessment theme is most clearly exemplified through comments which acknowledge the legitimacy of data representations without expressing any deep reflection. Responding to data in Unit A.1 (16 non-content questions vs 11 content questions), TA3 reported, "I thought there would be more content questions." Unlike TA1 and TA4, he did not attempt to explain the context. At no point did he express self-critique. This was most notable in Unit B, regarding wait time. TA3 had an average 1.9s of wait time before speaking again for content questions throughout the course. He reflected, "I think what I'm doing is okay."

Goal-setting in this interaction type was marked by high levels of confidence. In Unit A/M.2, TA3 said he planned to ask, "a few more questions than last time," and rated himself as "highly confident" in doing so. He went on to increase his content questions, and explained his approach in Unit A/M.3 as simply, "I was trying to ask more content-oriented questions." This improvement lasted for the rest of the observations, with twice as many content as non-content questions in each remaining session. In terms of wait time, TA3 was the only participant to set a goal *lower* than the suggested length of 3 seconds, which he achieved. At pre-test TA3 was near ceiling in self-efficacy, and did not change at post-test.

6.3.2.3 Shallow dismissal: Almost no reflection, declining self-efficacy

The third interaction type to emerge involved dismissiveness toward constructive feedback. TAs in this group neither acknowledged ineffective teaching strategies (as indicated by the training in Qualtrics), nor did they exhibit critical self-reflection. The experiences of TA2 and TA5 typified this group. When presented with data, they gave no indication that their practices were misaligned with effective teaching practices. When the data contradicted their inaccurate recall for recent teaching behaviors, they did not reflect on the discrepancy. In terms of goal setting, this group followed the suggestions from the system, increasing questions or wait time, but their enactments of one strategy were usually out of sync with the other, indicating a lack of ability to connect low-level tactics with overall strategies.

Shallow dismissal emerged most clearly as a lack of acknowledgement when data did not match the stated objectives of the training. TA2 and TA5 began the semester asking very few content questions at baseline (M = 2.6 and M = 6.8, respectively), and both TAs had almost no student talk. Unlike the other groups, these TAs made no reflective comment on the feedback in Unit A. The closest they came

to acknowledging ineffective teaching emerged when they described feeling compelled to cover the material. When responding to data that disconfirmed their recall (e.g., number of questions asked), they responded that the data matched what they had just reported doing, even when it did not. Interestingly, however, in Unit B this group did notice that their wait times were longer (i.e., better) than they had expected, indicating a willingness to accept data that appeared complimentary.

Goal-setting for this group seemed disconnected from the context of the classes. TA5's only goals in Unit A/M.2 were to "Ask better questions" and "Give students more time to think about question." This indicates that without critical self-reflection, TAs might only produce vague or irrelevant goals. Furthermore, when asked to recall their goal selections a few days later, neither TA could accurately remember what they had said they would try. This may have been partly due to the fact that each TA responded late to the prompt to log into *ClassInsight* for their performance reviews. TA2 waited 3 days before logging in for Unit A/M.3, and TA5 waited 9. It is likely this made meaningful review difficult. When asked to reflect on his goal-attainment in A/M.3, TA5 admitted that he could not remember the day in question.

In terms of putting goals into action, this group saw improvement, but it tended to be transitory or ineffective. They fixated on employing singular tactics in isolated contexts. For example, each had a large increase in the number of questions they asked between A/M.2 and A/M.3 (TA2: from 1 to 19 content questions, TA5: 6 to 15 content questions). During the review section (A/M.3), TA2 described an attempt at a new method of asking for student input while doing board work, rather than solving the problems himself. His students answered 74% of his content questions that day, and he said this approach gave him new insight into students' misconceptions. However, he asked only 5 content questions the following session, and averaged 3.4 per session thereafter. In reviewing these data in Unit B/M.3, he reflected, "Session last week was review for midterm, while session this week was introducing novel content. Thus the latter had significantly less questions for the class." The importance of student participation did not transfer to the new context.

TA5 also reduced content questions after one session, from the high of 15 down to 2 the next session, with an average of 3.0 questions per session thereafter. He never reported a reason for this change. However, his average wait time was less than 1 on his day of high questioning, and there were no student answers. In Unit B, TA2 and TA5 both improved on wait time for content questions (TA2: 3.9s to 5.9s, TA5: 3.3s to 4.7s). However, as the number of questions declined, student participation disappeared. TA2 reflected that waiting longer, "Did not seem to make any particular difference, as students could not answer the question anyway." TA5 reflected that he should probably wait longer in the future, despite waiting over 3 seconds for the majority of the few questions he asked.

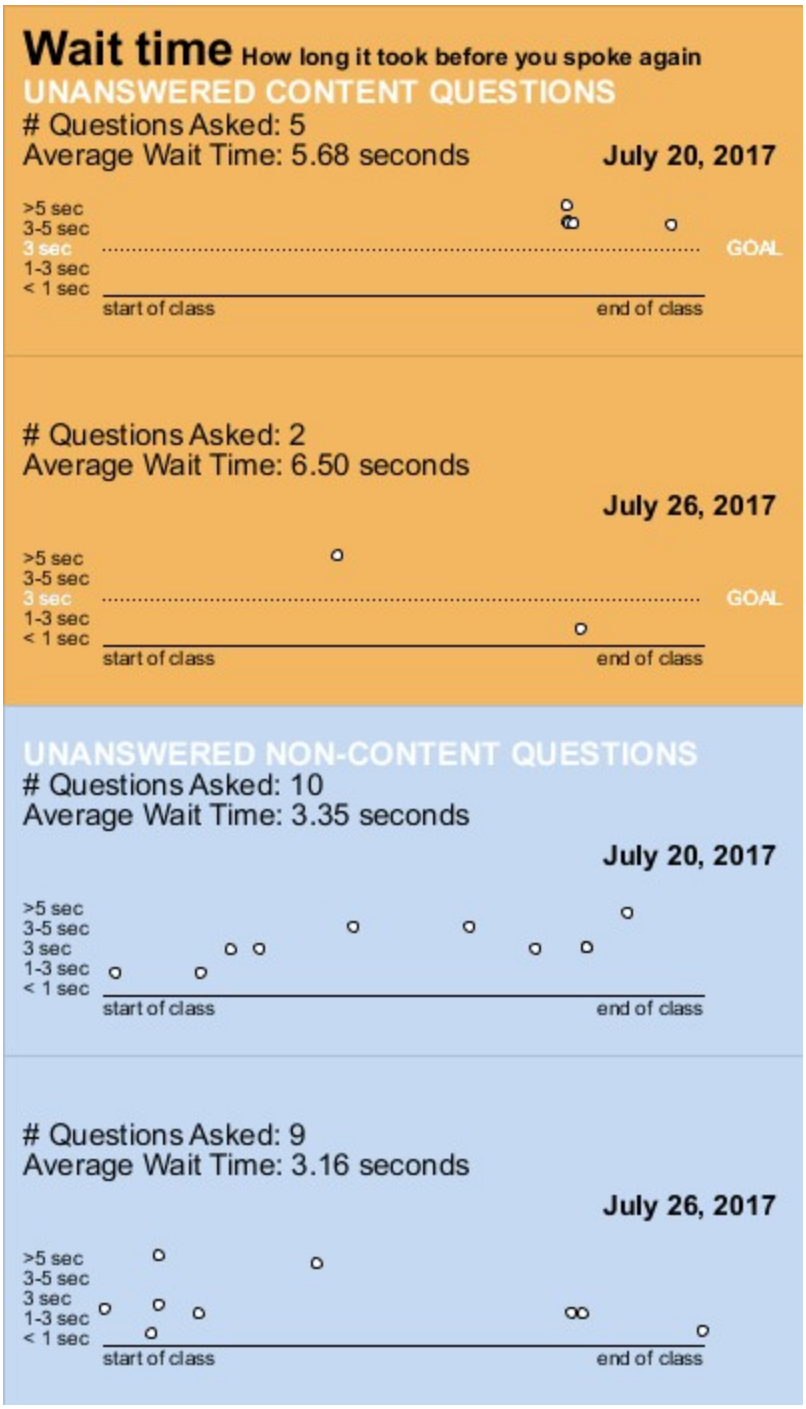


Figure 6.8: A typical visualization that emerges for TAs in Shallow Dismissal. Note the low number of content questions, especially in relation to the number of non-content questions.

Shallow dismissal stands out from the other interaction types in several ways. One important feature was in contrast to the fact that other TAs successfully reached their behavior goals, but TA2 and TA5 struggled to even remember which tactics they had selected. Their goals seemed isolated, and their attempts to change failed to produce useful outcomes. While they both volunteered for the training,

and reported it was useful, in the end, they each seemed to abandon the effort to engage students, and both showed a modest decrease in self-efficacy.

6.4 Discussion

The PAR-TS framework seemed to promote some reflection and planning for all TAs. Three engaged in substantive self-critique, and these instructors seemed to improve on the measured variables. Each TA made plans, and all of them attempted at least one new strategy. All of the TAs received the same type of feedback, and all saw that there were areas where they could improve. Although two of the TAs did not engage in meaningful reflection, concrete planning, or beneficial changes to their teaching, they did still engage with each stage of the training. The content of the training may not have been relevant for these participants, but the stages of the training at least kept them engaged and returning to Qualtrics.

However, while all TAs attempted strategies, they did not all maintain the use of these strategies. Each TA who made concrete plans and performed more thoughtful reflection seemed to continue challenging themselves as the course progressed. The findings indicate that what the individual TAs brought to the table changed how they responded to the feedback.

The study provides evidence that there may be at least three clusters of responses ("interaction types") to this socio-technical training system. Productive self-doubt, confirmatory assessment, and shallow dismissal may be useful ways to describe sets of behaviors, beliefs, and attitudes for determining how the next stage of technology-enhanced training might adapt to the different needs of the instructor.

The three interaction types indicate potential relationships between what the TA knows about pedagogy, what they believe about their ability to enact teaching strategies, and their approaches to how students learn. Not surprisingly, teacher-centeredness was high for everyone at the beginning of the study. More interesting, that did not seem to have an impact on what TAs did. Beliefs, particularly self-efficacy, seemed to have a stronger influence. This indicates that if a TA's self-efficacy has an impact on how they reflect, set goals, and apply their emerging knowledge, it is important to explore how this belief may relate to other individual variables.

Confirmatory assessment may have been due in part to the TA's prior training, his high self-efficacy at pre-test, or a combination. It is not yet clear which of these features may have contributed to his ability to change behaviors. The next study may help clarify influences on this interaction type.

One of the more interesting facts to emerge from productive self-doubt was that these TAs not only improved discursive behaviors, but in one case also increased student participation. A change of this type is very hard for most instructors to achieve. It is promising that a short, experimental intervention like this one can produce immediate student outcomes—and worth further exploration.

6.5 Summary of Results

The tentative product successfully promoted reflection and planning for all TAs—but TAs had very different individual results. The study tested the strength of the PAR-TS framework as a guide for making design and implementation decisions on building a PD app. The stages of the training kept users engaged and returning to Qualtrics, though the content of the training may not have been relevant for each participant.

The framework's predictions were preliminarily verified: each TA who made concrete plans and performed thoughtful reflections seemed to continue challenging themselves and growing. For those TAs who did not plan, there was less reflection and less evidence of change. One of the TAs in Study 3 not only improved discursive behaviors, but also increased student participation. This is a hard outcome for novice instructors to achieve; I aim to understand the TA factors that led to a positive outcome.

Inductive data analysis surfaced three qualitative groupings that describe how people responded to the training: productive self-doubt (PSD), confirmatory assessment (CoA), and shallow dismissal (ShD). These "interaction types" may be useful as a way for generating meaningful summaries of what users do when they engage in this socio-technical training system we call SmartPD.

Table 6.1: The three interaction types and their component features elicited from Study 3.

	Productive Self-Doubt (2)	**Confirmatory Assessment (1)**	**Shallow Dismissal (2)**
ATI Pre-test	High	High	High
TSES Pre-test	Mid-range	High	Low
Self-reflection	Critical	Mixed	Absent
Goal-setting	Tactical	Mixed	Vague
Actions	Deliberate	Deliberate	Isolated
ATI Post-test	Mixed	Mixed	Mixed
TSES Post-test	Increased	No change	Decrease

Qualitative assessment of the data produced in Study 3 surfaced PSD, CoA, and ShD. The following table reproduces those findings. The highlighted rows in Table 6.1 show the classification features that show separation between each of the groups.

But while we can now group TA beliefs and behaviors into descriptive clusters, it is not yet clear which features of an individual TA's attitudes may contribute to their exhibiting those behaviors. Self-efficacy seemed to have a strong influence on how TAs reflect, set goals, and apply their emerging knowledge. Individual perspectives beyond self-efficacy, such as teacher- and student-focus, did not show compelling correlations to what TAs did. The major limitation of this study, the very small number of participants, limits the lessons that can be learned from this data. In the following study I build on these preliminary findings and begin to address the limitations of the work so far.

Chapter 7: Development and formative evaluation of updated app (Study 4)

In Study 3 I described three different interaction types. These categories described (a) the orientations TAs had toward reflecting and planning via SmartPD, (b) the actions they took following these stages, and (c) changes in beliefs the TAs did or did not experience. Interaction types were based on analyses of TA behaviors within the app, in the class, and on the self-report measures of their beliefs. Some TAs showed positive behavior and belief change, while others showed less evidence of change. Study 3 did not reveal any patterns regarding teacher-focus and student-focus, neither at pre-test nor post-test. All of the participants had high teacher-focus at the beginning of the study. At post-test, attitudes toward student/teacher-focus changed in unpredictable ways. The study had a small number of participants, which may have masked attitudinal trends or effects, if any exist.

This chapter describes a follow-on study, where I continue the analysis of the previous chapter with more depth and on a larger group of TAs. This final study is partly based on the expectation that TA attitudes would have an impact on how TAs interact with SmartPD and what they do in the classroom. In addition to the behavioral data from the previous studies, I can now extend the analysis by applying what Study 3 suggested about TA interaction types. Those groupings serve as a classification strategy which may help to produce more insights about how to support TA learning.

Before moving forward, I present a visualization for the increasingly complex research space. Figure 7.1 organizes the participants, intervention interactions, and outcomes that we have investigated thus far, and proposes what the final study will help examine. The design of this model combines the Interconnected Model of Profession Growth (Clark & Hollingsworth, 2002) with a method for deriving path coefficients using partial least squares (Sanchez, 2013). I will not run a statistical test of the coefficients in this study, but the model is helpful to introduce at this point for three reasons:

- It provides visual organization of the complex research space,
- It begins to map out a testable collection of the multivariate space that these studies are defining,
- And it supports a DBR-inspired reflection on the research.

Circles in the diagram represent the domains of the Interconnected Model of Professional Growth (Clark & Hollingsworth, 2002), described in Chapter 2. These include the personal domain (prior traits/states of the TA), the external domain (TA learning intervention), the domain of practice (how the TA changes—or not), and the domain of consequence (how the students change—or not). These domains are high-level collections of variables that comprise the boxes in the model. Boxes in the diagram represent mid-level phenomena that constitute each circle. Not shown are the low-level behavioral data that comprise each box.

There are three visual aspects of the arrows in the diagram that require definition. Thick arrows between the circles represent hypothetical directions of influence. Arrowheads that are filled represent where the prior studies have shown evidence of influence. Arrowheads that are empty represent purely hypothetical relationships for which we have not yet seen empirical evidence. Thin arrows between boxes and circles represent the nature of the relationship between each mid-level phenomena and higher-level domain. An arrow that points from a box to a circle represents a "formative" phenomenon—one that has a direct influence on the domain. An arrow which points from a circle to a box represents a "reflective" phenomenon—an operationalization of some aspect of the domain. These

definitions of reflective and formative variables are drawn from a practical description of path coefficient analysis using partial least squares (Sanchez, 2013).

Putting the definitions of the visualizations together, the TAs' prior knowledge about teaching and learning, their beliefs (operationalized by self-efficacy), and their attitudes (operationalized as student/teacher-focus) should all impact what they do in the app, what they do in the class, and any potential changes in beliefs and attitudes. The "known/unknown formative variables" box represents the wide array of variables that make each TA unique. Some of these we know, such as how much teaching experience they have, but many of them we do not.

In this view of the research, app interactions are governed by interaction types. Study 3 produced the interaction types based on an inductive reflection of everything the TAs did. In this study, those interaction types are changed into a top-down rubric which allows for using them as their own descriptive phenomena for each TA. Following from these interactions, what TAs do in the app should have a direct impact on TA outcomes. Finally, what the TA does and believes should influence what students do. In return, what students do should have a reciprocal effect on what TAs do and believe.

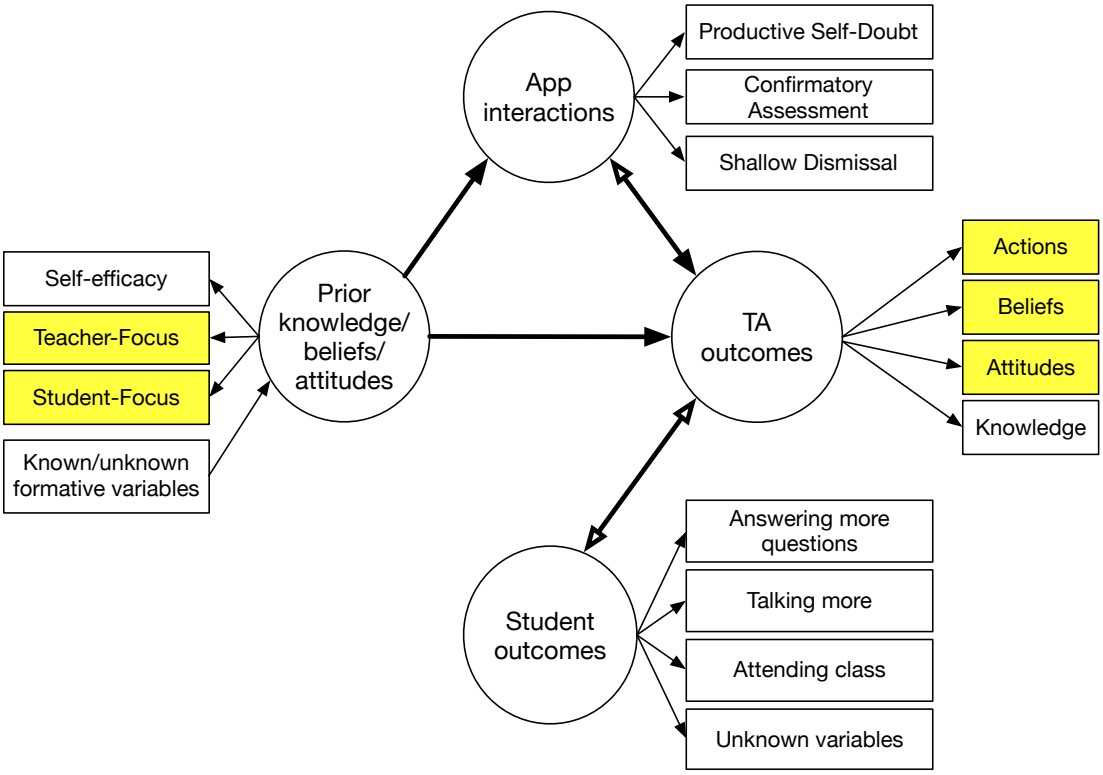


Figure 7.1: Visual model of variables within the PAR-TS framework. This chapter focuses on (a) uncovering the possible influence of teacher/student-focus on TA outcomes, and (b) delivering more information about how actions, beliefs, and attitudes may change through the PD process.

The research space as described in this model is large, and it could grow to include many more variables. For the purposes of this study, however, I primarily focus on the impact that attitudes might have on how TAs change. Teacher/student-focus is likely to have a unique impact on TA outcomes. While I want to highlight this possibility, I will also investigate the rest of the variables that are in Figure 7.1.

The high-level goals for this study are to produce more and deeper insights of SmartPD, built upon the PAR-TS framework and extended to 5 modules; to better understand how beliefs impact responses to SmartPD; to uncover how scaling up might change outcomes; and to build a SmartPD platform for this and future research.

7.1 Research questions

1. Does SmartPD help novice instructors enact better teaching practices?
2. To what extent do beliefs and attitudes contribute to behaviors?
    a. Do self-efficacy and teacher/student-centeredness predict interaction types?
    b. How do self-efficacy and teacher/student-centeredness influence classroom behaviors?
    c. How do classroom behaviors influence beliefs and attitudes?
3. What contributions would result from extending instructional units of SmartPD to include (a) an additional teaching opportunity and (b) reflection on roadblocks.

7.2 Implementation

Based on the goals of this broader study, I engaged in a field deployment of an app for SmartPD. Beyond using a new platform and increasing the sample size, the intervention for Study 4 was very similar to that of Study 3, with one major difference. In order to enhance the ability to detect changes to the variables of interest, I increased the length of each conceptual unit by an extra 2 modules (RQ3). This effectively doubles the opportunities for TAs to reflect on the teaching behaviors targeted in the unit, plan for new tactics, consider roadblocks before moving on, and enact changes.

A minor limitation of Study 3 was that the entire training was built and conducted by hand. I personally updated the third-party platform (Qualtrics) and individually contacted each participant when it was time to proceed with the next step of the training. It was not possible to re-use that specific curriculum without updating it for new participants and contexts, nor was it possible to scale up to a larger number of participants. Qualtrics required that much of the training be hard coded for every lesson. This slowed the deployment process and required considerable human oversight—barriers that we eliminated for this study. In this study users interact with a new app built specifically for the instructional purpose of increasing discursive teaching methods. This platform allows the research to leave behind the "good enough" Qualtrics platform of Studies 2 and 3.

This also serves as a transition in the social context of this series of work to more closely emulate Personal Informatics interventions in large-scale implementations. Users of SmartPD in this study, using the app, no longer have as many one-to-one interactions with the researcher. This change in social context alters retention and dropout behaviors in the study, which is examined closely in the findings.

7.2.1 App Design and Development

Over the course of 6 months (November 2017 to April 2018), the research team built an application for training TAs through a simulated smart classroom protocol, based on the objectives of SmartPD. The design of the app (called *ClassInsight*) automated features of the Qualtrics deployment that had been controlled by the researcher. This includes:

- Sending emails to TAs to alert them to new training opportunities
- Allowing user to pace themself while coordinating in-app learning activities with most recent teaching opportunities
- Unlocking training modules appropriately based on customized rules for each TA (i.e., mostly based on teaching schedule)
- Individual TA tracking of teaching behaviors, logging in to app, and module completion
- Computational visualization production from class records
- Storing and re-issuing user input data across training modules

The research previously included manually building data visualizations, tracking user progress, attending to the timing to unlock modules, and storing interaction data. A SmartPD system should be able to automate these processes. *ClassInsight* was capable of this, and produced visualizations procedurally immediately following data collection. Example output are shown in Figure 7.2.

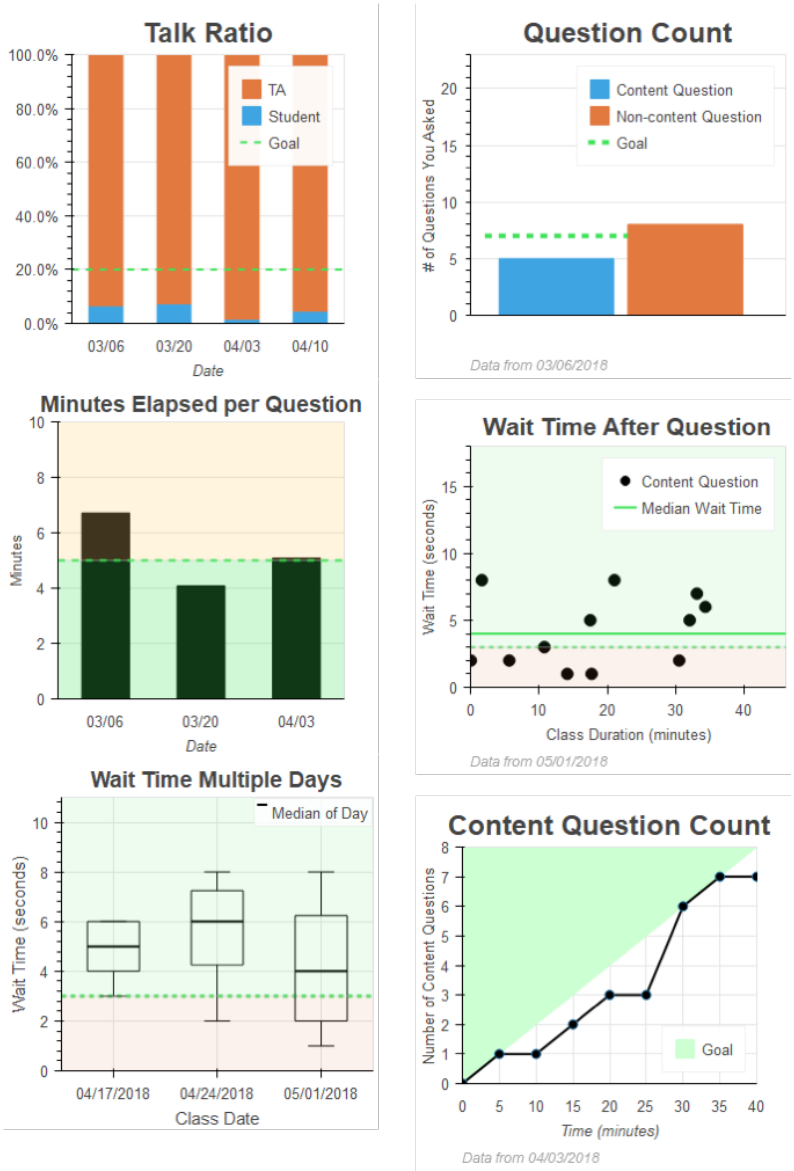


Figure 7.2: Visualizations of TA activity as shown in *ClassInsight*.

The research team decided to retain the two conceptual units from Study 3 (*asking more content questions* and *waiting longer*), but to lengthen the training on each topic. Instead of only 3 modules per unit, the new training had 5. This doubled the number of training cycles and TA teaching opportunities per unit. The second training cycle for each unit focused on asking TAs to report roadblocks they were experiencing, brainstorm solutions, and then describe their results on trying those solutions in the final module. This expanded the reflection/planning/action cycle to incorporate a meta-level of roadblock analysis in each unit.

The curriculum of *ClassInsight* in this deployment once again addressed discursive teaching habits around asking content questions and waiting for students to respond. Unit A addressed asking more content questions and fewer non-content questions. Unit B addressed waiting at least 3 seconds for students to respond to content questions before moving on and asking again or providing an answer.

Each of the modules followed a specific order of themes. The italicized sentences below identify the instructional objectives and activities that were new for this study. Note that the final visualization of unit metrics, previously part of Module 3, is now located in Module 5. This maintains its goal of delivering a summative evaluation and final reflection opportunity for the user.

- **Module 1 (Topic Introduction/Review):** Introduce the unit topic and reveal current data trends along high- and low-level performance variables
- **Module 2 (Planning):** Prior to the next teaching opportunity, ask TAs to identify concrete learning goals and to select teaching tactics and performance goals for metrics relevant to the given unit topic
- **Module 3 (Topic/Performance Reflection):** Reveal updated visualizations and request deep reflection on overall performance and individual tactics. *Ask about what might have been a roadblock*.
- ***Module 4 (Roadblocks Planning):*** *Reveal roadblocks and strategies to overcome them as produced by other participants. Ask TA to select solutions to attempt in following class.*
- ***Module 5*** **(Topic/*Roadblocks Reflection*):** Updated visualizations of unit metrics. Requests for deep reflection on overall performance *and individual roadblocks*.

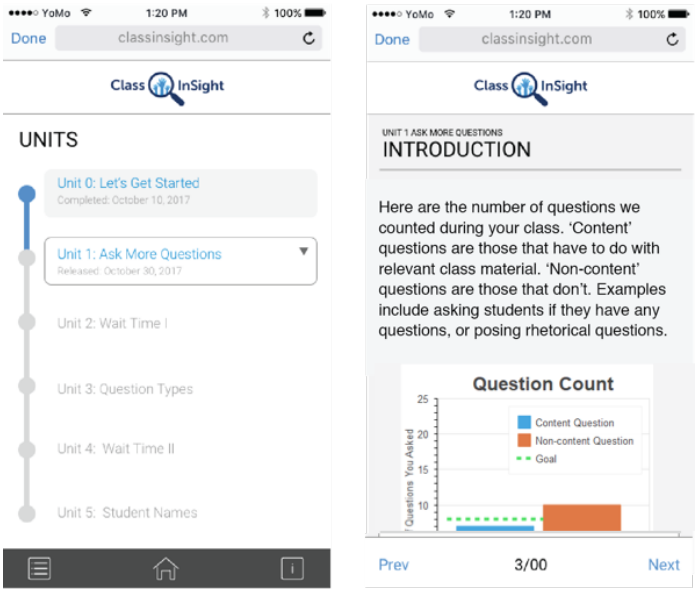


Figure 7.3: User interfaces for modules in *ClassInsight*.

7.2.2 Recruitment and Observations

In a 16-week spring semester (2018), 21 TAs volunteered to participate. Each signed a consent form allowing for in-class observations and access to the instructional app. The research team observed the majority of the classes taught by all of these TAs, equaling 236 classes over 13 weeks. TA08 taught twice per week, and the rest taught once per week. 7 of the TAs had a substitute either once or twice during the semester. Sessions taught by substitutes were removed from the final analysis. Observations began on week 2 of the semester for most TAs. The first week of observation served as training for research assistants. Data collected during observer training were removed.

19 of the volunteers responded to the intake survey, which included questions about demographics, teaching experience, and self-reports for beliefs and attitude measures: self-efficacy, teacher-focused, and student-focused approaches to teaching (TSES and ATI, Klassen et al., 2009; Prosser & Trigwell, 2006). Of the TAs who completed the intake survey, TA21 had very short class interactions (typically about 5 minutes each) and spent the rest of class time overseeing the students in quiet quiz-taking. Despite completing the training, the structure of this TA's class made it very different from the other courses in the sample. Therefore, this TA's data were removed entirely from the analyses. This left 18 TAs with self-report pre- and post-tests, and 20 with classroom data. I describe the observational data for all 20 TAs whenever possible and exclude the two non-responders where self-reports are necessary for analysis.

TAs were required to take the initial survey in order to use *ClassInsight*. There were 8 graduate students and 10 undergraduates who received access to the app. They came from five STEM colleges within Carnegie Mellon University: Chemistry (Chem), Electrical and Computer Engineering (ECE), Mathematics (Math), Mechanical Engineering (ME), and Computer Science (SCS). Interactions between academic year and STEM college are described in Table 7.1.

The professor who oversaw the two TAs from ME (TA01 and TA02) unexpectedly canceled several recitation sessions early in the semester. Because the app is designed to build on a TA's recent teaching experience as they plan their following sessions, these TAs would not have been recruited had their unpredictable teaching schedules been known beforehand. Rather than remove them from the study however, I decided to conduct the observations of their class, withhold access to the app, and include them as a small comparison group of observed TAs.

Table 7.1: The distribution of colleges and student years for TAs participating in Study 4.

College	Doctoral	Masters	Senior	Junior	Freshman	Total
Chem	1					**1**
ECE		1				**1**
Math	4		1	2	1	**8**
ME	1	1				**2**
SCS			2	4		**6**
Grand Total	**6**	**2**	**3**	**6**	**1**	**18**

7.3 Analysis Methods

7.3.1 Variables Collected

In-class observations, app interactions, interviews, and surveys conducted during this study produced a list of qualitative and quantitative variables which are explained below. Each variable belongs to exactly one of the four categories from the Interconnected Model of Professional Growth (Clark & Hollingsworth, 2002). These categories are the external domain (training), personal domain (knowledge, beliefs, attitudes), domain of practice (teaching), and domain of consequence (student outcomes).

The domain of practice includes all of the behavior variables gathered while the TAs teach. The domain of consequence includes the behavior variables gathered about students during class. The external domain in this study is operationalized as the app (*ClassInsight*) built on the framework and principles of SmartPD. I do not test for knowledge on any of the trained topics in this instance of the work, so the personal domain is limited to beliefs and attitudes of the TA. These are operationalized by self-report surveys (self-efficacy and teacher/student-focus).

7.3.1.1 In-class Behaviors

Quantitative measures were based on data gathered using the TA Dashboard, as with Studies 2 and 3. For clarity, some definitions are reprinted in Table 7.2.

Table 7.2: Data produced by the TA Dashboard (partial reprint of Table 4.1).

Variable	**Description**
ST	The total number of milliseconds where coder held down 'S' key, representing student talk (st-talk in chapter 4).
TA	The total number of milliseconds where coder held down 'K' key, representing TA talk (ta-talk in chapter 4)
Silence	The total number of milliseconds in which neither ‘K’ nor ‘S’ were pressed.
Attendance	The total number of students in class.
U-ST	The unique number of students who speak during class.
Overlap	The total number of milliseconds in which ‘K’ and ‘S’ were pressed simultaneously.

Table 7.3: Variables produced by, or induced from, TA Dashboard data.

Domain of Practice (TA behaviors while teaching)	**Domain of Consequence** (Student behaviors during class)
Number of content questions asked	Changes in # of students in attendance
Number of non-content questions asked	Number of answers to TA questions
Ratio of content to non-content questions	Ratio of questions answered
Amount of time spent talking	Overall ratio of student talk
Amount of time spent in silence	Length of individual speech acts
Ratio of talk and silence	Number of non-answer comments
Length of Wait Times I and II	Ratio of students who speak at least once
Use of student names	Number of hand raises
Number of times students are called on	

16 research assistants and 3 members of the research administration performed in-class observations of 220 classes included in the analysis (of the total 236 observed). After removing classes taught by substitutes, observations taken during RA training, and each of the records from TA21, there were 66,255 unique classroom events, recorded with TA Dashboard, available for analysis. These events produced data from the classroom regarding what TAs and students did, as described in Table 7.3.

7.3.1.2 In-app Behaviors

The app produced 1,750 unique interactions across all of the TAs that used it. Some TAs produced many more interactions than others. TAs were self-paced and allowed to choose how many of the training's 10 total modules they would complete. This produced 5 "training completion" categories (called AppGroups hereafter):

- *B_Complete*: all training finished.
- *B_Start*: at least 1 module of Unit B finished.
- *Dropout*: at least 1 module of Unit A and no modules from Unit B finished.
- *No_Start*: TAs who volunteered for the study but never used the app.
- *Unavailable*: TAs who volunteered to use the app but were not given access.

One of the primary goals of SmartPD is to increase critical self-reflection through the use of grounded feedback. As described above, the TAs saw visualizations representing certain aspects their questioning behaviors and student responses in the app. Similar to Study 3, each visualization was framed to match the specific goals of the given unit. For example, if the unit focused on wait time (Unit B), the TAs saw timeline representations of their wait time throughout their most recent teaching opportunity. Furthermore, as units proceeded over multiple weeks of teaching, TAs received visualizations of trends in their behavior over time. Reflections on these data, and the subsequent behaviors around planning for future teaching opportunities, comprise the bulk of the qualitative data from the external domain. Metadata, e.g., how frequently TAs interacted with the app, comprise the quantitative data from the external domain.

7.3.1.3 Reported Beliefs and Attitudes

Table 7.4: External Domain variables, observed from TA in-app behaviors during training.

External Domain (from TA in-app behaviors during training)	**Personal Domain** (TA beliefs and attitudes)
Categorical responses to reflection and planning prompts in *ClassInsight*	Qualitative responses to reflection and prompts in *ClassInsight*
Reaction times to login prompts	Self-Efficacy (TSES)
Number of prompts preceding login	Information Transmission/Teacher-Focus (ATI)
Number of training and teaching opportunities	Conceptual Change/Student-Focus (ATI)
	Interview data segmented by conceptual unit as determined by research team

Table 7.4 describes variables derived from interactions TAs produced in the use of *ClassInsight* as well as their personal beliefs and attitudes operationalized by self-report and interviews. 18 of the TAs

completed the pre- and post-test surveys. These included basic demographics, a teaching history survey, and self-report measures of self-efficacy for teaching (Klassen et al., 2009) and approaches to teaching (Prosser & Trigwell, 2006). Refer to Appendices A and B for the full surveys.

15 of the participants volunteered for the final interview. Each interview lasted about 1 hour and took place in private with two members of the research team. The research team conducted these interviews with one questioner and one note taker. Following each interview, the two participating researchers conducted an interpretation session to summarize every comment the interviewee made (Beyer & Holtzblatt, 1998).

7.3.2 Interaction Type Rubric Evaluation

As with prior studies, the qualitative data that the research team gathered comprised semi-structured interviews, in-app interactions, field notes, and self-report surveys. The majority of qualitative analysis in this study focused on applying interaction types to individual TAs. I formalized this process by creating a rubric to identify the interaction types induced in Study 3 (Productive Self-Doubt, Confirmatory Assessment, and Shallow Dismissal).

The rubric was based on critical components that defined each interaction type in Study 3. These components produced 14 rubric questions with 4 possible dominant interaction types for each: PSD, SHD, COA, or unknown (UNK) when data were insufficient to make a judgment. I answered each question based on a borad evaluation of the reflections, goals, and enactments TAs produced throughout the semester.

A sample of rubric questions is presented in Table 7.5 (see Appendix C for the full rubric). Unlike Study 3, where TAs were categorically labeled by interaction type, this top-down process produced a numeric value of markers for each interaction type, with the total *sum* of the three values being less than or equal to fourteen.

I created an additional rubric for adjusting the number of dominant interaction types produced by the rubric. Because TAs were assigned an integer value for each of the interaction types, it was then possible for TAs to reflect multiple combinations of types. For example, a TA might be equal parts productive self-doubt and shallow dismissal. To reweigh the integer values, I reviewed interview data and field notes. In some cases, those data produced additional evidence for or against each of the interaction types. Whenever I found concrete evidence for or against a given interaction type (PSD, COA, or SHD), I applied a doubling or halving weight (see Table 7.6) to the final tally of markers (this produces no change when an interaction type's rubric rating is 0).

Table 7.5: Abridged rubric for assigning interaction types to participants (see Appendix C for full rubric).

	Productive self-doubt	**Confirmatory assessment**	**Shallow dismissal**
Reflection			
When presented with data about their teaching performance, did they make an effort to explain some challenge they had faced in class?	They explain some challenge they faced in the class. They try to explain what they think may have gone wrong. They may indicate that their practices were misaligned with effective teaching practices.	Unlikely to mention a challenge. They may mention a constraint (different because it cannot be solved). Unlikely to elaborate or explain underlying causes, context, or conditions. They may acknowledge accuracy of data, but without any reflection to explain or interpret it.	They may mention a challenge, but more as an excuse rather than as an explanation or insight.
Accuracy of recall			
Are they able to accurately recall their selected strategies?	This TA is likely able to remember their selected strategies.	This TA may be able to remember their selected strategies.	This TA is unlikely to remember what they said they would try.
Goal-setting			
Are their goals relevant? List their strategies and goals and check for coherence with stated and derived problems.	Their goals will usually have something to do with the problems they have outlined. They are concrete.	Their goals may relate to stated problems, and will usually be concise. They may dismiss suggestions as not relevant to their particular class.	Goals may not connect well to stated problems.
Enactment			
Is their strategy selection aligned with their actual needs, or is it out of sync? Are reflections connected to the outcomes? Review their reflections, any stated needs, their goals, and their in-class performance.	Likely to show an awareness of their attempts to change, exhibit reflection in review module, and persist in new behaviors. Students may exhibit changes.	Not likely to exhibit deep reflection, but persists in new behaviors. Students may exhibit changes.	Reflection is likely to be shallow, struggling to connect low-level tactics with overall strategies. Behaviors are not likely to persist. Students unlikely to exhibit changes. Look for indications that the TA blames the students. Look for indications that the TA plans to continue to do something that isn't working
Following study			
Did TSES change?	TSES likely to increase at post-test.	TSES likely to stay high.	TSES likely to decrease.

Table 7.6: Weighs applied to interaction group rubrics based on interview data as a way to surfacing dominant themes given the additional data.

	-1 (x2 multiplier)	**0 (x1 multiplier)**	**-1 (x0.5 multiplier)**
PSD-Interview	Reveals additional evidence that PSD happens, just wasn't picked up in the app.	No additional evidence of PSD or lack of PSD.	Evidence of lack of PSD, such as anxiety about performance without addressing issues.
CoA-Interview	Reveals additional evidence of CoA, such as saying that they don't need the training.	No additional evidence for or against CoA.	Evidence of lack of CoA, such as describing a desire to do better or needing support.
ShD-Interview	Reveals additional evidence of ShD, such as saying that they don't care or just want to prioritize practical goals.	No additional evidence for or against ShD.	Evidence of a lack of ShD, such as dismissing for practical reasons or revealing interest in performance.

As an example of this process, TA03's count of markers from the rubric were 2 for PSD, 4 for COA, 6 for SHD, and 1 UNK. In reading the interview data I found evidence against PSD (e.g., they reported that their visualized performance of student talk was worse than they expected, so they decided to ignore it) and for SHD (e.g., counting seconds after a question seemed fine, "but you forget"). I did not find evidence for or against COA. By applying the weights, this changed the markers to 1 PSD, 4 COA, 12 SHD, and 1 UNK. At 3 times greater than the next closest marker type, I categorized this TA's interaction type as Shallow Dismissal dominant with sub-dominant Confirmatory Assessment. This process aids in high-level analysis and discussion. The tradeoff is that it also limits the specificity of each TA's experience.

Following the precepts of validity within DBR (McKenney & Reeves, 2012), Table 7.7 reveals the full breakdown of initial rubric scores and post-hoc weights applied to each participant. These raw data give the reader insight into the evaluation process. I invented this weighting measure myself and there does not yet exist any test of internal validity. DBR allows for creative qualitative evaluation methods so long as researchers express methodological "credibility" (McKenney & Reeves, 2012). We achieve credibility by revealing a full account of how we conduct our analyses and delivering a full explanation of the process.

Table 7.7: Initial rubric results, transformation weights from interview data, and final outcomes after applying weights for each TA participant. "N/A" indicates TAs who did not participate in final interviews. AppGroups: 1 = Unavailable, 2 = No_Start, 3 = Dropout, 4 = B_Start, 5 = B_Complete.

		Rubric markers per interaction type				Transformation applications				Transformation outcomes			
TA##	**AppGroup**	**PSD**	**CoA**	**ShD**	**Unk**	**PSD**	**CoA**	**ShD**	**Unk**	**PSD**	**CoA**	**ShD**	**Unk**
TA03	**5**	2	4	7	1	-1	0	1	0	1	4	14	1
TA06		8	2	2	2	0	0	1	0	8	2	4	2
TA08		4	4	6	0	1	-1	0	0	8	2	6	0
TA13		10	1	2	1	-1	-1	1	0	5	0.5	4	1
TA14		3	5	5	1	1	0	-1	0	6	5	2.5	1
TA15		9	1	1	3	0	0	0	0	9	1	1	3
TA18		3	6	4	1	-1	1	0	0	1.5	12	4	1
TA19		13	0	0	1	1	-1	0	0	26	0	0	1
TA20		5	6	3	0	0	1	-1	0	5	12	1.5	0
TA22		1	4	9	0	n/a	n/a	n/a	0	1	4	9	0
TA05	**4**	4	0	10	0	-1	0	1	0	2	0	20	0
TA17		0	4	8	2	n/a	n/a	n/a	0	0	4	8	2
TA07		3	0	6	5	-1	0	1	0	1.5	0	12	5
TA09	**3**	1	0	6	7	-1	1	-1	0	0.5	0	3	7
TA12		1	2	2	9	-1	1	1	0	0.5	4	4	9
TA10	**2**	0	0	0	14	n/a	n/a	n/a	0	0	0	0	13
TA11		0	1	0	13	n/a	n/a	n/a	0	0	1	0	13
TA16		0	0	0	14	n/a	n/a	n/a	0	0	0	0	13
TA01	**1**	0	0	0	14	n/a	n/a	n/a	0	0	0	0	13
TA02		1	0	0	13	n/a	n/a	n/a	0	1	0	0	13

7.3.3 Exploratory Data Analysis

I analyzed behavioral data using exploratory data analysis (Behrens, 1997) to uncover possible relationships between most of the variables identified above. One of the important themes of this work is the attempt to encourage TAs to implement new teaching strategies over time. In the prior studies there have been very small sample sizes; as a result, it was relatively easy to look at the raw counts of relevant data and produce simple conclusions about which data seemed to change during each intervention. In this study I once again reviewed the raw data that each subject produced, focusing primarily on the number of each type of question they asked each week, how many questions students answered, the length of wait times, the number of students present, the number of students who spoke,

and how long students spoke for. As in previous analyses, it was still easy to see if a TA had improved on any of these measures by simply looking at visualizations of individual data (see Figure 7.4 for an example).

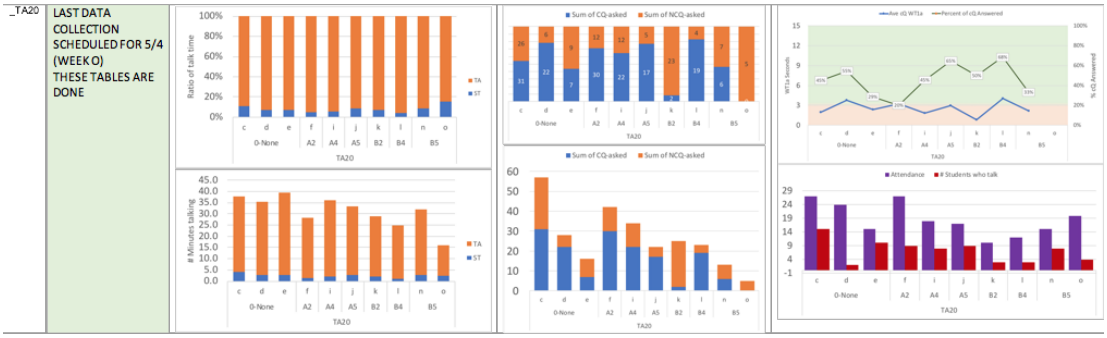


Figure 7.4: Example of raw data visualizations analyzed for a single TA participant.

7.3.4 Analyzing Change over Time

Behavioral change over time is an important variable for interpreting the findings of this study. The raw data above allow us to evaluate trends in participant behavior. They have limits, however, in that they are difficult to review across larger numbers of participants than those in the previous studies.

To assist in the analysis, I performed a simple least squares regression against the week of the semester. This variable stands as a proxy for change in each variable over time. The learning sciences and HCI communities have used slopes and intercepts to evaluate changes over time (Dyke, Adamson, Howley, & Rosé, 2013). While those studies are typically quantitative, in this work, the resulting slopes are meant to be used as guidance for qualitative analysis, to inform a model for later confirmatory assessment. Instead of issuing statistical comparisons of these change variables, I use them to compare participants in a more holistic, qualitative analysis.

To achieve this, I normalized each quantitative variable to a participant-level z-score ($Z = (i - \bar{i})/s$ for any variable i corresponding to the number of participant's sessions s) for each TA across each week of observations, using participant mean and standard deviation values for each variable across the semester. I then used the standardized measure of each week's performance to produce a linear slope, intercept, and r-square for each variable. This process allowed me to compare deltas for individual TAs based on their own emerging performance. For example, TA06 taught for 16 weeks. Observations for analysis started in week 3; there was no instruction in weeks 8 (spring break) and 16 (finals). From the remaining class periods, I calculate an attendance mean and standard deviation, as shown in Figure 7.5. From these data TA06 has an estimated attendance slope of B = -0.713, an intercept of 12.477, and an r-square of 0.590.

$$\text{TA06 } \textit{mean} = \frac{1}{n}\sum_{w=3}^{n} attendance(w) \quad = \quad 6$$

$$\text{TA06 } \textit{standard deviation} = \sqrt{\frac{\sum_{w=3}^{n}(i-\bar{i})^2}{n-1}} \quad = \quad 3.67$$

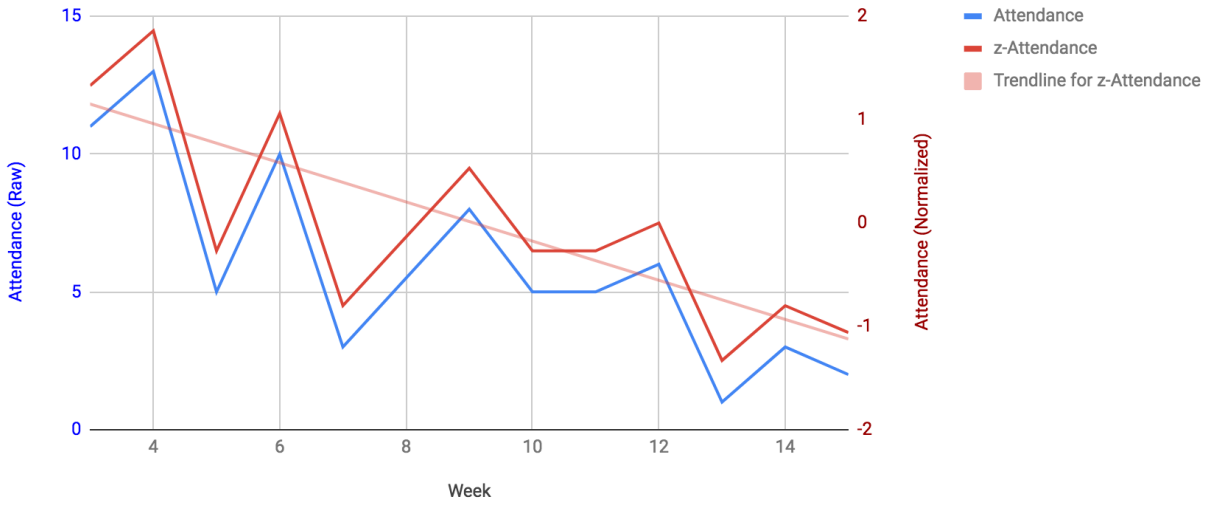
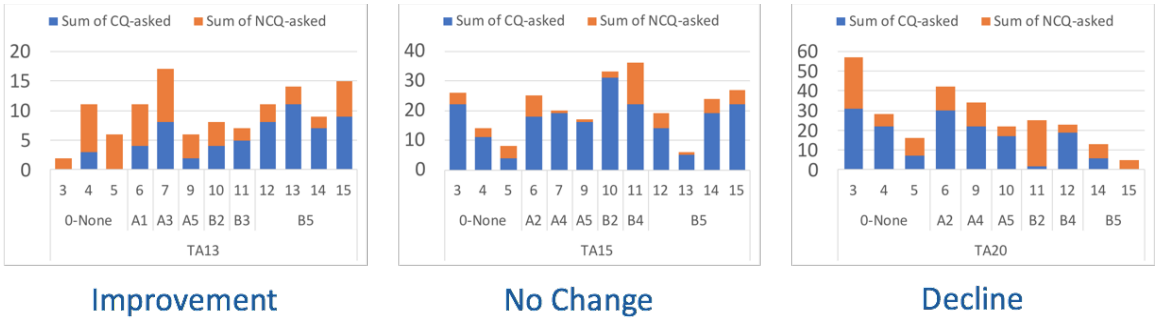


Figure 7.5: Example mean, standard deviation, and normalization of the attendance variable for TA06, to enable comparison across TAs on quantitative variables.

I conducted most of the behavior analysis of TA data using comparisons of slopes. In cases where numbers seemed improbable, I compared TAs' intercepts to see if their surprising slopes corresponded to large differences in starting points. Whenever this approach remained unclear, I returned to raw scores to verify that the slopes and intercepts made practical sense. Note that these analyses are still qualitative in nature. I am looking for general trends. The only difference with previous studies is that I have moved from reviewing arithmetic means or raw counts of variables to reviewing standardized indices of change.

In terms of classifying TAs as improving, declining, or staying the same for any given variable, I used a cutoff of .05 units change across slopes. A slope value $B \geq .05$ on any given measure would indicate "improvement" for that variable. A slope between -.05 and .05 would indicate "no change," and a slope $B \leq .05$ would indicate a "decline." These cutoffs represent fairly small changes; learning to teach takes time, and one semester is unlikely to produce dramatic changes.

In Figure 7.6 note how the three charts of raw data reveal patterns of question asking ratios that indicate "improvement, no change, or decline." Compare those graphs to the slopes for an intuitive grasp of these ranges.



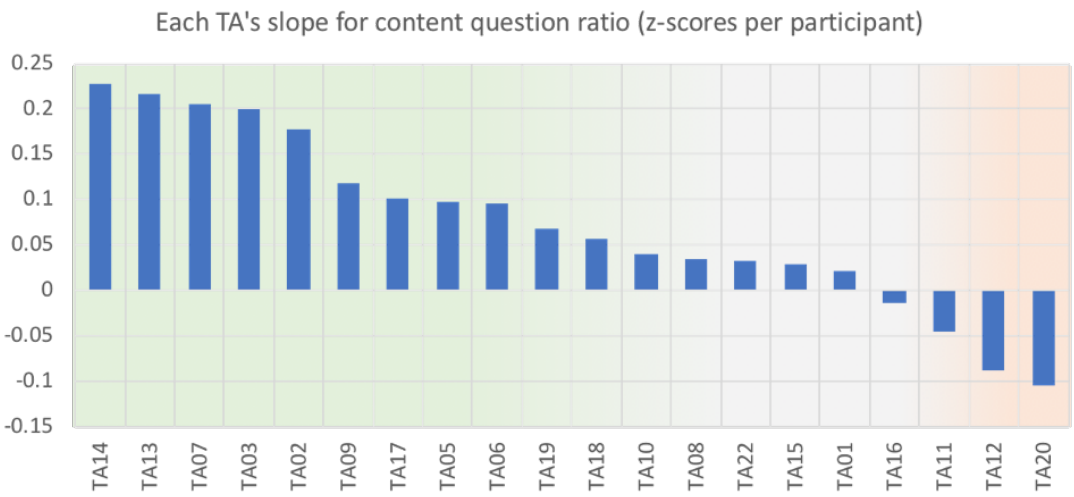


Figure 7.6: (Pg. 92) Example raw values for the content question ratio variable for three participants across teaching opportunities, and (above) their corresponding z-scores in context of all participants. Participants in the green section of the graph are categorized as improvement for this variable, white represents no change, and orange indicates decline.

7.3.5 Final Variable Set

I calculated standardized scores for each continuous variable in Table 7.8, then calculated slopes, intercepts, and r-square values. I primarily used slopes to explore changes across TAs., uncovering prevalent trends using exploratory data analysis rather than hypothesis testing (Behrens, 1997). When slopes were insufficient for explaining behaviors, I reviewed intercepts and r-squares. For example, if a TA had a very large slope in one metric of change, such as questions asked, but exhibited shallow dismissal overall, I would review their intercept and determine if the large slope came from relatively small numbers. “Improving” from asking 1 question in the first class to asking 2 in the next might produce an impressive slope but is not practically meaningful. The findings and discussion sections will clarify where these interpretations are necessary.

As a visual example of this process, Figure 7.7 shows the rate of attendance change per teaching opportunity for each TA. Analyzing standardized slopes across participants, all but one of the TAs lost students throughout the semester. This is typical for college classes in general. To compare the actual magnitude of these changes, Figure 7.8 includes the standardized intercept for each TA. Note the linear relationship between the rate of student attrition compared to the initial number of students. For attendance, the initial size of the class seems to have the strongest impact on the size of the class at the end of the semester.

Table 7.8: Complete set of quantitative variables collected per TA.

Variable	Definition
TA	Participant number
YearNo	Categorical index of TA year: 6 = PhD student, 5 = Master, 4 = Senior, 3 = Junior, 2 = Sophomore, 1 = Freshman
TSES	Pre/Post self-report averages (0 – 4 range) on the Teachers' Sense of Efficacy Scale (Klassen et al., 2009) Versions: include pretest, posttest, and post-pre (to calculate change)
ITTF/CCSF	Self-report averages (0 – 5 range) on the Approaches to Teaching Inventory (Prosser & Trigwell, 2006), indicating 'information-transmission/teacher-focus' and 'conceptual-change/student-focus' Versions: include pretest, posttest, and post-pre (to calculate change)
Week	The numbered week of the semester, 1 – 16
Opportunity	The ordered number of observed teaching session
Day	Date of observation
ModuleNo	Final app module completed, numbered 1 – 10
AppGroup	Final grouping by amount of app completed
Attendance	Raw number of students in class that session
cQask	Raw number of 'content questions' asked that session
cQans	Raw number of 'content questions' answered that session
ncQask	Raw number of 'non-content questions' asked that session
ncQans	Raw number of 'non-content questions' answered that session
allQ	Sum of 'cQask' and 'ncQask'
c-ncQratio	'cQask' divided by 'allQ' (defined as 0 when no questions are asked)
cQansRatio	'cQans' divided by 'cQask' (defined as 0 when no cQ are asked)
allQansRatio	Sum of 'cQans' and 'ncQans' all divided by 'allQ' (defined as 0 when no questions are asked)
TA-min	The continuous number of minutes observed with student (TA) talk
ST-min	The continuous number of minutes observed with student (ST) talk
SI-min	The continuous number of minutes observed with neither TA or ST talk
total-class-min	Sum of 'TA-min,' 'ST-min,' and 'SI-min'
ST-TA-ratio	'ST-min' divided by the sum of 'TA-min' and 'ST-min'
SI-ratio	Percentage of 'total-class-min' spent in 'SI-min'
cQask-per-min	Number of 'cQask' divided by 'TA-min,' representing the average number of questions a TA asks during lecture
ncQask-per-min	Number of 'ncQask' divided by 'TA-min,' representing the average number of questions a TA asks during lecture
c-ncQratio	'cQask-per-min' divided by all questions asked per minute.
PSD-trans	The transformed integer of productive self-doubt markers
COA-trans	The transformed integer of confirmatory assessment markers
SHD-trans	The transformed integer of shallow dismissal markers
UNK	The number of interaction type markers that are unclassifiable

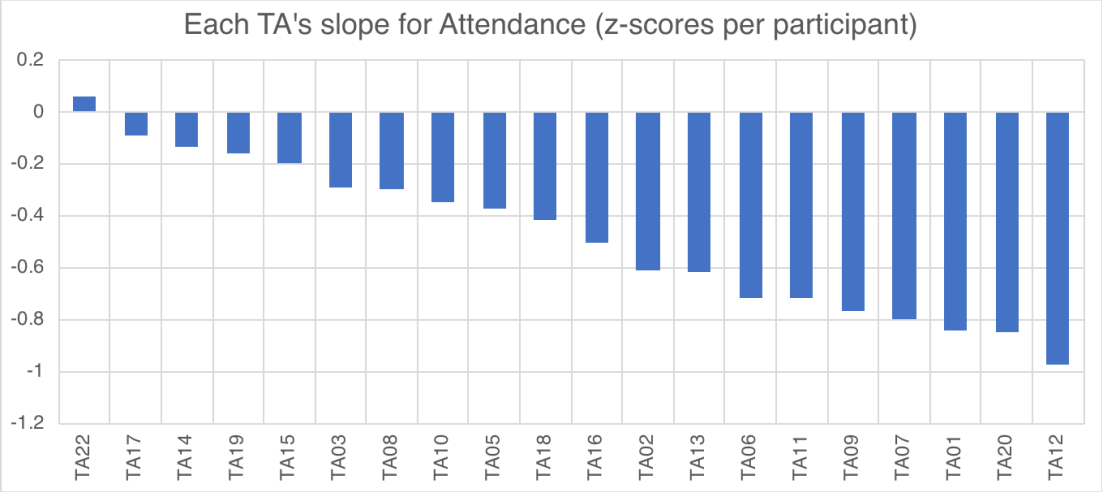


Figure 7.7: Calculated slope of standardized attendance for each TA. Slopes near 0 would be preferable. All but one TA exhibits a negative slope.

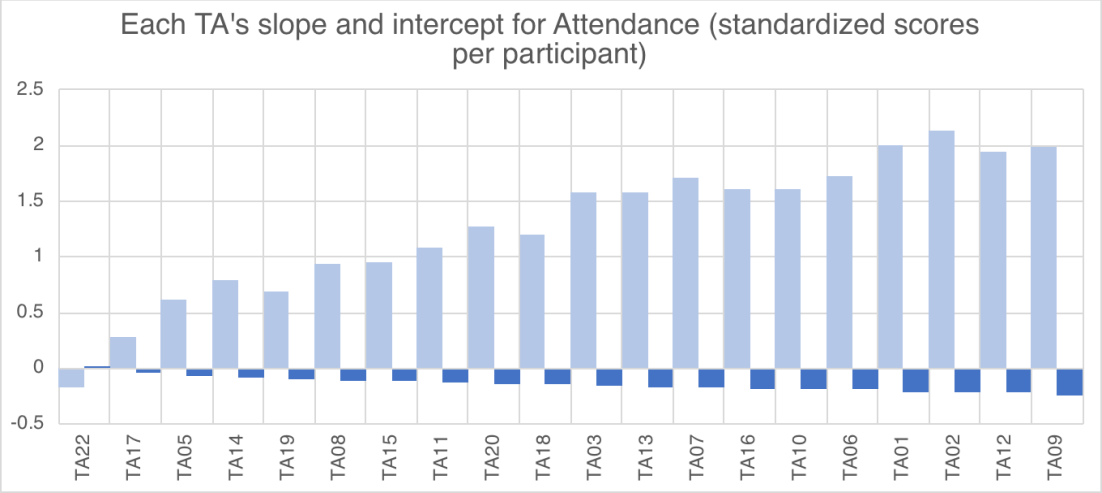


Figure 7.8: Calculated intercept for standardized number of students in attendance (light blue) for each TA and standardized slope (dark blue).

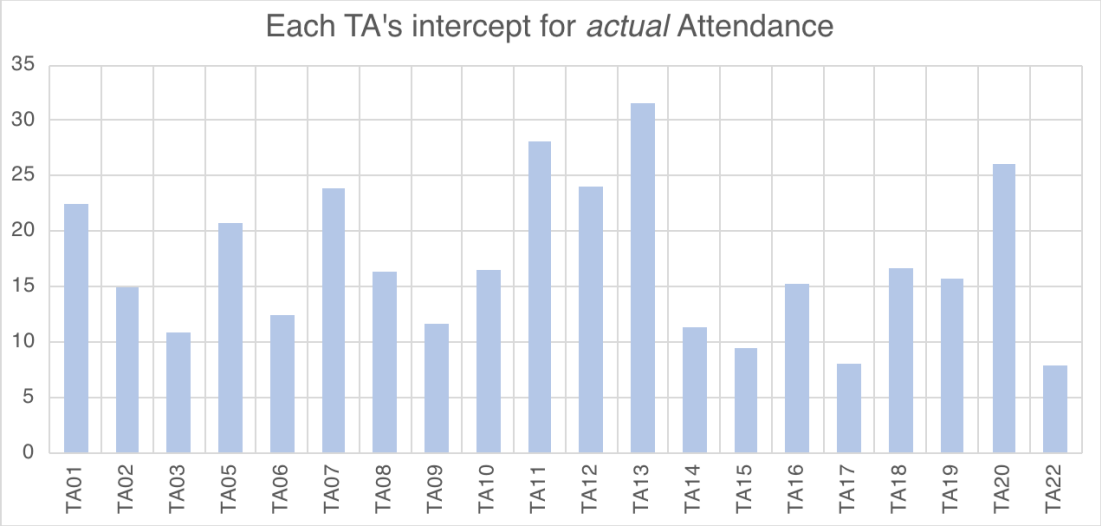


Figure 7.9: Calculated intercept for *actual* number of students in attendance (y-axis) for each TA.

After reviewing all of the raw data, then comparing slopes across TAs, I generated a final list of variables that is meant to synthesize the data across the smallest number of variables (Table 7.9). The top section of individual TA differences shows the unique measures for each TA, such as the number of opportunities they had to teach (while observed), their student year, beliefs and attitudes, etc. TA and student behavior variables are standardized slope measures that indicate changes across the semester.

Table 7.9: Final, top-level variables included in analysis report. These variables represent the synthesis of those included in Table 7.8.

Variable	Notes
Individual TA differences	
Opportunity Count	The number of classes taught by the TA during the semester (observed)
Module Number	Final app module completed, numbered 1 – 10 (*Unavailable* and *No_Start* TAs are not included in analyses of ModuleNo, due to their dissimilarity and small number)
YearNo	Index of TA year: 6 = PhD student, 5 = Master, 4 = Senior, 3 = Junior, 2 = Sophomore, 1 = Freshman
Interaction type (transformed)	4 integers that indicate prevalence of evidence for and against each type: PSD-trans, CoA-trans, ShD-trans, and Unk
TSES (Beliefs)	Self-efficacy as indicated by TSES, measured pre-and post-test; 1 – 4 scale
TSES-post-pre	Final TSES minus first TSES to indicate change direction and magnitude
ATI (Attitudes)	Teacher/Student-focus as indicated by ITTF and CCSF, measured pre- and post-test; 1 – 5 scale
ITTF-post-pre	Final ITTF minus first ITTF, to indicate change direction and magnitude
CCSF-post-pre	Final CCSF minus first CCSF, to indicate change direction and magnitude

Student behavior variables	
Attendance	Students in class
CQ Answers	The total number of content questions students answered divided by all content questions asked (cQans/cQask)
TA behavior variables	
CQ-Ratio	The percentage of questions the TA asked that were content based (cQask/allQ)
CQ-Pace	The number of content question asked during class divided by the total amount of TA talk time (cQask/TAmin)
Silence	Defined as total silence during class divided by the sum of TA talk, students talk, and silence (SI-min/(TA-min + SI-min + ST-min))
Student & TA behavior variable	
ST-TA-ratio	The total amount of student talk in a class divided by the sum of student and TA talk (ST-min/(ST-min + TA-min))

Each of the behavior variables is treated as an improvement when the slope is positive. (See chapter 2 for explanations of the learning benefits of discursive teaching, student talk, the use of deep questions, and increases in wait time.)

Belief variables were gathered before the intervention (pre-test) and at the end of the semester (post-test). In addition to reporting these metrics, I include a '*post-pre'* variable which indicates whether a TA changed orientation during the semester, and in which direction. Theoretically, self-efficacy and student-focus would be considered to have improved if their '*post-pre'* measures were positive, and teacher-focus would be considered an improvement if '*post-pre'* was negative. This may not always be the case, however, because people sometimes reorient their beliefs as they gather a more realistic mental representation of their orientations. To keep things as clear as possible, therefore, I do not reverse code teacher-focus.

Finally, for each research question I used simple correlation calculations to explore possible relationships between variables (see Appendix D for full correlation matrices). I reviewed these correlations at two levels: among all TAs who completed the self-report surveys, and separately among those who had at least made it to Unit B. The majority of my exploration involves reviewing and reporting these correlations and how they provide insights and raise interesting possibilities in the research space. I do not test for statistical significance in this case, given the small number of research participants.

7.4 Findings

7.4.1 Completion rates and attendance

As in Study 3, TAs were able to pace themselves through the training. Unlike the TAs in Study 3, Study 4 participants did not have a 100% completion rate. This may have been a result of the different, smaller number of interactions with the researcher. Each of the AppGroups had the following number of participants:

- *B_Complete*: 9 TAs completed all of Unit A and Unit B
- *B_Start*: 3 completed Unit A and at least 1 module of Unit B
- *Dropout*: 3 completed only 1 or 2 modules in Unit A
- *No_Start*: 3 had access to the app but did not use it; 2 never took the survey
- *Unavailable*: 2 volunteered to participate but were not given access to the app

To simplify results, *B_Complete* and *B_Start* can be combined into a single group (*B_Group*), representing all TAs who completed at least some portion of Unit B. Table 7.10 shows the number of observed weeks taught by each TA and the final AppGroup for each enumerated TA.

Table 7.10: TA number, the number of weeks taught & observed (OppCount), and AppGroup. Participants in the right column are included in *B_Group*.

TA	OppCount	AppGroup	TA	OppCount	AppGroup
TA02	9	*Unavailable*	TA19	9	*B_Complete*
TA01	10	*Unavailable*	TA03	10	*B_Complete*
TA16	7	*No_Start*	TA20	10	*B_Complete*
TA10	11	*No_Start*	TA18	10	*B_Complete*
TA11	11	*No_Start*	TA14	12	*B_Complete*
TA07	8	*Dropout*	TA15	13	*B_Complete*
TA12	9	*Dropout*	TA06	13	*B_Complete*
TA09	11	*Dropout*	TA13	13	*B_Complete*
			TA08	12	*B_Complete*
			TA17	8	*B_Start*
			TA22	9	*B_Start*
			TA05	13	*B_Start*

To compare change across TAs, I produced a descriptive summary of slopes (Table 7.11). Standardized scores are denoted with ‘z-‘ preceding the variable name.

Table 7.11: This table shows descriptive statistics for calculated slopes across participants for each repeated, continuous variable. 'z-' denotes a variable that is reported in its standardized form. The last column summarizes the change indicated by the statistics. Ex: average attendance is down across participants, indicating most classes lose students over the semester.

Variable	Min	Ave	Max	Var	Change over time
z-Attendance	-0.24	-0.14	0.02	0.00	Students come to class less
cQask	-1.80	-0.04	1.38	0.53	TAs change differently
cQans	-0.80	0.03	1.42	0.21	Questions answered differently
ncQask	-1.14	-0.38	0.12	0.13	Fewer ncQ asked
ncQans	-0.38	-0.06	0.26	0.03	Fewer ncQ answered overall
allQ	-2.55	-0.42	1.45	0.89	TAs change differently
z-c-ncQratio	-0.10	0.07	0.23	0.01	TAs change differently
z-cQansRatio	-0.22	0.01	0.21	0.01	Students change differently
TA-min	-1.37	-0.04	1.02	0.43	TAs change differently
ST-min	-0.24	0.02	0.55	0.03	Students change differently
SI-min	-0.37	-0.05	0.23	0.03	TAs change differently
total-class-min	-1.24	-0.08	0.80	0.37	Classes get somewhat shorter
z-ST-TA-ratio	-0.16	-0.09	-0.04	0.00	Students change differently
z-SI-ratio	-0.16	-0.03	0.18	0.01	TAs change differently
z-cQ-per-min	-0.19	0.00	0.19	0.01	TAs change differently
z-ncQ-per-min	-0.22	-0.10	0.05	0.01	Fewer ncQ asked

7.4.2 RQ1: Do novice instructors enact better teaching practices with *ClassInsight*?

Throughout the rest of this section, the following visualizations summarize the analysis described in Analysis Methods. Slopes $\geq +.05$ indicate improvement, $\leq -.05$ indicate declines, and exclusive values between -.05 and +.05 indicate no change in behavior.

7.4.2.1 Summary of Findings

Here we discuss findings by the AppGroup that instructors were in. First, the two TAs in *Unavailable* did not have access to the app. Discursive behaviors and outcomes primarily showed declines (Table 7.12). The ratio of content questions either improved or did not change. The pace of content questions declined for each TA; both TAs asked fewer content questions per minute as the semester proceeded. Each TA used less silence as the semester progressed, indicating fewer pauses in TA talk (lecturing), and less wait time. Ratios of student talk either stayed the same or declined. The two TAs differed in terms of the proportion of questions that students answered.

Table 7.12: Changes for TAs in *Unavailable* based on slope analysis.

AppGroup	TA	TA: CQ-Ratio	TA: CQ-Pace	TA: Silence	Both: ST/TA Talk	ST: CQ Answers
Unavailable	TA01					
	TA02					

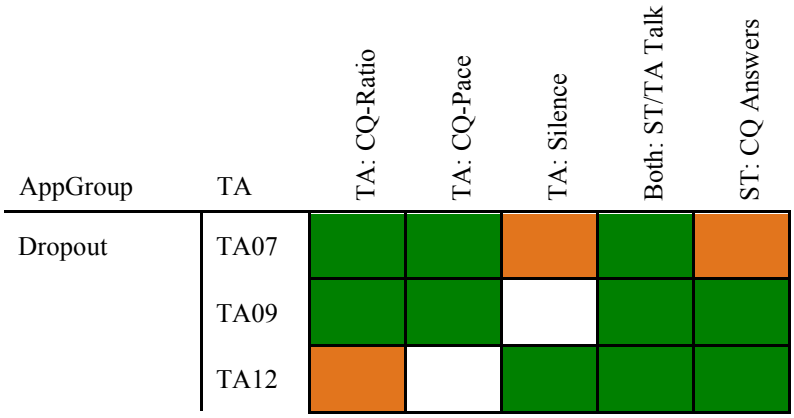
TAs in *No_Start* also showed a majority of declines (Table 7.13). Their ratios and rates of content questions mostly did not change and declined in one case. All three exhibited less silence. Student talk increased for one TA and declined for the other two. The proportion of student answers declined for two TAs and did not change for the other.

Table 7.13: Changes for TAs in *No_Start* based on slope analysis.

AppGroup	TA	TA: CQ-Ratio	TA: CQ-Pace	TA: Silence	Both: ST/TA Talk	ST: CQ Answers
No_Start	TA16					
	TA10					
	TA11					

TAs who discontinued app use after one or two modules (i.e., *Dropout*; Table 7.14) showed different trends compared to the TAs who never used the app. They exhibited a majority of improvements with only a small number of declines or no changes. Two TAs improved in both ratio and rate of content question asking, while the other declined or did not improve. Use of silence was different for all three TAs. All three TAs increased the ratio of student talk. One declined in the rate of students answering questions and the other two improved.

Table 7.14: Changes for TAs in *Dropout* based on slope analysis.



There were 12 TAs who completed Unit A. They exhibited a range of outcomes (Table 7.15). Two-thirds of TAs improved in the ratio of content questions they asked. Three TAs did not change. One declined. Half of the TAs improved in their pace of content questions. Three declined and three showed no change. Silence was less definitive. The 12 TAs were evenly distributed in terms of the three outcomes. About half of the TAs improved at the amount of student talk and about half declined. Half of the TAs improved in the proportion of questions that students answered. Three declined and three did not change.

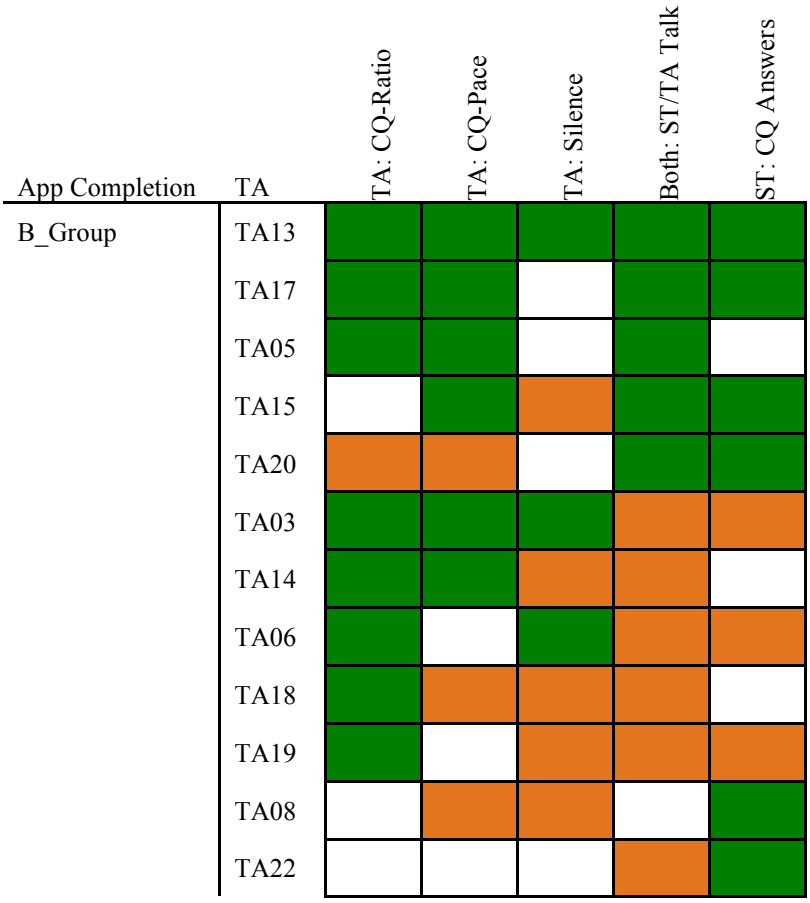
For TAs with at least some exposure to SmartPD (*B_Group* and *Dropout*), two variables related to content questions show consistent improvement. First, TAs using the app tended to ask a higher ratio of content questions over time (10 of 15 TAs improved, 2 of 15 declined). Second, they asked those content questions at increasing frequency (8 of 15 TAs improved, 3 of 15 declined). Only one TA showed declines on both measures. The other three who showed declines did so on only one of these two variables.

Comparing these two variables to *No_Start* and *Unavailable* reveals an apparent difference. Only 1 of 5 TAs improved at content question ratio. None of the 5 TAs improved at the pace of content question asking.

Silence was a relatively underutilized tactic. 4 of 15 TAs who used the app increased their use of silence. 6 of 15 used less silence. 5 showed no change. Of the TAs who never used the app, however, the rates seemed worse. 5 of 5 TAs used less silence.

The remaining two behavior variables rely at least somewhat on changing *student* behavior, not just TA behavior, which can be very difficult. 10 of the 15 TAs who used the app showed increasing rates of student participation on at least one of the two student-dependent variables. The 5 remaining TAs declined on one or both of the student-dependent variables. Of the TAs who did not use the app, 3 of 5 declined on both measures. The remaining two improved on only one measure each.

Table 7.15: Changes for TAs in *B_Complete* and *B_Start* (*B_Group*) based on slope analysis.



The net finding from measurement of these variables is that app use does seem to have an effect on TA behaviors, although one semester of training may not be sufficient for the change to consistently impact student behaviors for all instructors. In contrast, TAs who did not use the app show a consistent decline in teaching quality throughout the semester on each of the variables. When viewed through a lens of steady decline for TAs who received no intervention, the positive but mixed findings for *ClassInsight* users is promising for teacher professional development.

7.4.2.2 Beliefs and app completion

In the findings regarding RQ2, beliefs and attitudes are analyzed in detail with respect to how they interact with classroom behaviors. This short section focuses only on how beliefs and attitudes correlated to app use, because it emerged as an interesting finding tangentially related to RQ1. Table 7.16 is a table of correlations between the number of modules each TA completed and their corresponding self-reports at pre-test, post-test, and the change from pre to post.

Table 7.16: Correlations between the number of modules each TA completed and their corresponding self-reports at pre-test, post-test, and the change from pre to post. Note that self-efficacy (TSES), teacher-focus (ITTF), and student-focus (CCSF) have unique outcomes depending on the analysis view, either all TAs who completed the pre-test and post-test or *B_Group* only.

Module number completion and Beliefs	All TAs who completed surveys			*B_Group* only		
	TSES	ITTF	CCSF	TSES	ITTF	CCSF
Pre-test	-0.29	-0.25	-0.72	-0.29	-0.06	-0.20
Post-test	-0.21	-0.36	-0.72	0.21	0.03	-0.14
Post minus pre	0.14	-0.18	-0.30	0.57	0.11	0.00

Comparing module number to beliefs is one way of analyzing rates of app completion with TAs' beliefs and attitudes. Comparing the column of all TAs to the subset of *B_Group* in Table 7.16 shows that each self-report instrument has a unique relationship to how far TAs went with the training. We can consider pre-test measures as a loose prediction of how likely TAs are to complete the app, because these instruments are given before the intervention. The "all TAs" group contains everyone who dropped out or chose not to participate, and "*B_Group* only" does not. As the table makes clear, student-focus has a large negative correlation with module number for all TAs, and a small correlation with teacher-focus. This indicates that high student-focus correlates with early intervention dropout.

The post-test correlations support this finding. They show that most respondents did not change much between pre-test and post-test. The exception is in relation to *B_Group's* sense of efficacy, which showed a large improvement in correlation to app completion.

In summary, this set of correlations produces two high-level outcomes. Student-focus correlated to TAs leaving or failing to start the intervention. For those who stayed with the intervention, however, there exists a notable increase in self-efficacy for teaching.

7.4.3 RQ2: To what extent do beliefs and attitudes contribute to behaviors?

Following principles of exploratory data analysis, I use correlation matrices to investigate possible connections between beliefs and attitudes at pretest and interaction types that emerge from the TAs. For this research question, I will frequently compare the set of all TAs to the *B_Group* in isolation, as described earlier in this chapter.

7.4.3.1 Do self-efficacy and teacher/student-centeredness predict interaction types?

There is no clear indication that pre-test self-efficacy predicts interaction types as predicted by Study 3 (Table 7.17). The two relationships that emerge from All TAs is a modest negative correlation with PSD and a modest positive correlation with Unknown. Correlating with Unknown may simply be a matter of including the *Dropout*, *No_Start*, and *Unavailable* groups. As evidence, the correlations disappear when considering only the *B_Group*. Within the *B_Group*, self-efficacy does have a modest positive correlation with Confirmatory Assessment. This is internally consistent, as each depends somewhat on a user's confidence.

Evidence indicates that information-transmission/teacher-focus (ITTF) has a strong negative correlation with Productive Self-Doubt and a medium positive correlation with Unknown. The large negative correlation shown with All TAs remains with *B_Group* only. The correlation to Unknown disappears, but positive correlations with Confirmatory Assessment and Shallow Dismissal emerge. These are internally consistent, as a teacher-focus would not be likely to encourage the reflection required for PSD, but it would be reasonable to expect in connection to confirmation and/or dismissal.

Conceptual-change/student-focus (CCSF) has negative correlations with all three interaction types amongst All TAs, and a large positive correlation with Unknown. Compare this to *B_Group* only, where student-focus has a small negative correlation to PSD and no other relationships to the interaction types. These outcomes conform to the findings from the prior section (Table 7.16) that showed a relationship between high student-focus and intervention dropout.

Table 7.17: Correlation matrices between pre-test belief/attitude measures and count of interaction type markers following transformation (refer to 7.3.2 for explanation of rubric and transformation process). This table compares the entire set of 18 TAs who reported pre- and post-tests (left), and the subset of 12 TAs in *B_Group* (right). Self-efficacy (TSES) does not seem to predict interaction types. Teacher-focus (ITTF) shows a negative correlation to productive self-doubt. Student-focus (CCSF) may show a relationship to intervention dropout. (Unknown is high for All TAs because it includes participants who did not provide sufficient data.)

Interaction types	All TAs			*B_Group* only		
	TSES-Pre	ITTF-Pre	CCSF-Pre	TSES-Pre	ITTF-Pre	CCSF-Pre
PSD-trans	-0.24	-0.71	-0.43	-0.11	-0.72	-0.20
CoA-trans	0.06	0.14	-0.28	0.23	0.34	-0.02
ShD-trans	-0.09	-0.03	-0.21	0.06	0.34	0.17
Unknown	0.33	0.47	0.81	-0.04	-0.12	0.09

7.4.3.2 How do self-efficacy and teacher/student-centeredness influence classroom behaviors?

According to the correlation matrix in Table 7.18, self-efficacy may have a positive relationship with increasing rates of student talk. Self-efficacy at pre-test (TSES-Pre) has modest or medium positive correlations for *B_Group's* pace of content questions, use of silence, and number of questions students answer. The largest positive correlation is to increasing rates of student talk (r = .55). The only positive correlation for all TAs is also related to student talk (r = .37).

Table 7.18: Correlation matrix comparing pretest beliefs to behavioral variables.

SLOPES | BELIEFS	All TAs			*B_Group* only		
	TSES-Pre	ITTF-Pre	CCSF-Pre	TSES-Pre	ITTF-Pre	CCSF-Pre
ST: Attendance	-0.16	-0.40	-0.51	-0.03	-0.21	-0.02
TA: CQ-Ratio	-0.04	-0.06	0.02	0.07	0.23	0.50
TA: CQ-Pace	0.10	-0.07	-0.11	0.28	0.3	0.39
TA: Silence	0.03	0.17	-0.09	0.21	0.41	0.35
Both: ST/TA Talk	0.37	0.09	0.05	0.55	0.31	0.13
ST: CQ Answers	0.01	0.13	-0.15	0.31	0.33	0.09

With respect to attendance, instructors who are higher in teacher-focus may lose more students over the semester compared to TAs with less teacher-focus. Note that teacher-focus had a moderate negative correlation to attendance for all TAs (r = -.40) and a small correlation for *B_Group* (r = -.21). To illustrate this point, Figure 7.10 shows the linear relationship between the two variables.

When considering *B_Group* only, information-transmission/teacher-focus (ITTF) at pre-test shows positive correlations to all variables except attendance. Regardless of exhibiting an orientation toward information-transmission, these TAs seem to be doing what the app asks them to do, and students seem to be responding. From this view, teacher-focus at pre-test may indicate a willingness for TAs to follow instructions.

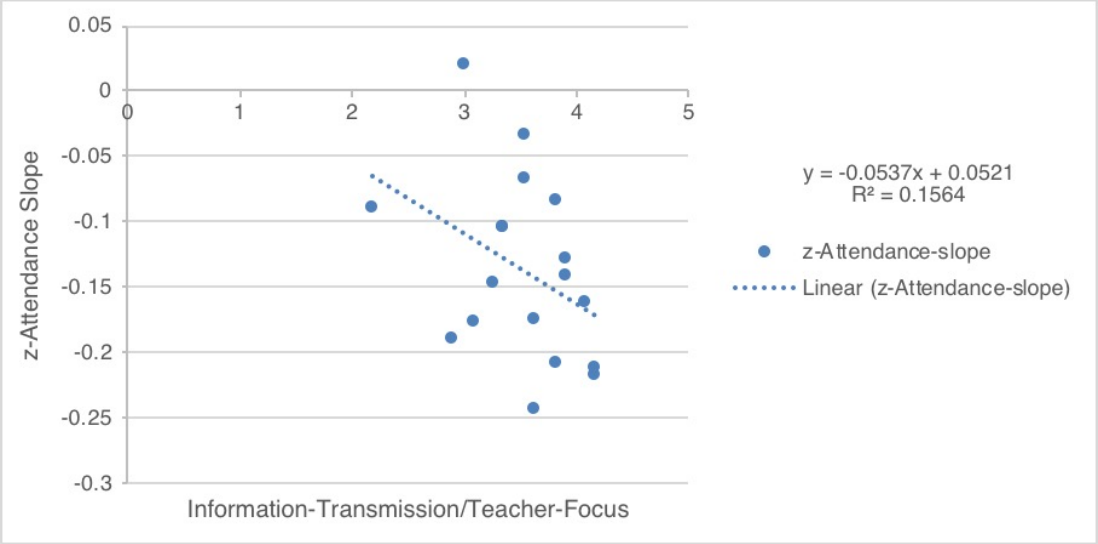


Figure 7.10: Possible relationship between teacher-focus at pre-test and declines in attendance.

7.4.3.3 How do classroom behaviors influence beliefs and attitudes?

As a high-level preview, this section will show that TAs who changed their behaviors during the semester decreased in teacher-focus. This effect was independent of what students did. Student outcome variables had their own impact on beliefs and attitudes, however, especially when student spoke up more in class. Broadly speaking, when students participated more, TAs showed an increase in student-focus. Relationships to self-efficacy were less straightforward but indicate that completing the app and seeing students change is related to increases in self-efficacy, whereas not finishing had a negative correlation.

Using the exploratory techniques that I have described so far, there are two approaches to view the impact that in-class behaviors may have had on TAs beliefs and attitudes. One way is to compare correlations between these variables and the post-test measures of TSES, ITTF, and CCSF. The other is to compare them to the change metrics described in Table 7.8, where the pre-test measure of each belief/attitude index is subtracted from its corresponding post-test measure. Table 7.19 shows both of these approaches for all TAs and for the subset of *B_Group* TAs.

Table 7.19: Correlation matrix (for all TAs) comparing classroom behaviors to post-study measures of beliefs and attitudes as well as change in those variables. Refer to Table 7.8 for explanation of belief/attitude change variable.

SLOPES | BELIEFS	TSES-Post	ITTF-Post	CCSF-Post	TSES-post-pre	ITTF-post-pre	CCSF-post-pre
All 18 TAs who reported pre- and post-tests						
ST: Attendance	-0.29	-0.34	-0.49	-0.13	0.00	-0.10
TA: CQ-Ratio	-0.33	-0.27	-0.26	-0.32	-0.32	-0.41
TA: CQ-Pace	-0.33	-0.32	-0.18	-0.50	-0.38	-0.14
TA: Silence	-0.21	-0.25	-0.16	-0.28	-0.60	-0.12
Both: ST/TA Talk	0.17	0.17	0.38	-0.26	0.13	0.50
ST: CQ Answers	-0.23	-0.01	0.09	-0.28	-0.18	0.31
Subset of 12 TAs from *B_Group* only						
ST: Attendance	-0.17	-0.08	0.09	-0.12	0.17	0.10
TA: CQ-Ratio	-0.24	-0.13	-0.27	-0.31	-0.50	-0.60
TA: CQ-Pace	-0.17	-0.14	-0.11	-0.52	-0.60	-0.37
TA: Silence	0.00	-0.05	0.06	-0.26	-0.63	-0.18
Both: ST/TA Talk	0.49	0.32	0.47	-0.25	0.04	0.37
ST: CQ Answers	0.26	0.32	0.31	-0.16	0.02	0.24

Almost all of the variables show a negative correlation to self-efficacy at post-test for All TAs. No matter what TAs tried or what students did in response, TAs in this group lost self-efficacy. *TSES-post-pre* makes this relationship even more clear. Attendance has no relationship to changes in self-efficacy,

but all the other variables do. This is particularly strong for the pace of content questions (r = -.50). The implication is that when TAs take fewer app suggestions, their self-efficacy *increases*. This may show a disconnect between what the average TA *believes* about their skills and what they actually *do*.

Comparing these self-efficacy scores for all TAs to *B_Group* only reveals some stark differences, however. People who completed the app and got their students talking/ answering questions showed a positive correlation to post-test self-efficacy. As far as *changes* in self-efficacy (TSES-post-pre), however, only app completion seemed to boost a TA's self-efficacy (r = .57, Table 7.16). All other behaviors showed either no relationship or a negative correlation. This is once again particularly noticeable with the pace of content questions having the largest negative correlation to changes in self-efficacy (r = -.52).

Teacher-focus for All TAs at post-test shows small or medium negative correlations to most of the variables. C*hanges* in ITTF, however, only correlate to the three types of TA actions. These negative changes mean that TAs who follow the app's suggestions also show declines in teacher-focus. It is worth noting that student variables do not correlate to this change, suggesting that teacher-focus is sensitive to what TAs try, but not dependent on what students do as a result. This set of relationships is even more pronounced in the change metric for ITTF in *B_Group*, where the negative correlations are all larger.

There is a notable relationship between student-focus outcomes and changes in student talk. The largest increases in CCSF for All TAs and *B_Group* alike relate to increases in the rate of student talk.

7.4.4 RQ3. What contributions result from extending instructional units?

Section 7.2.1 described the design of the app, including its module-based curricula. The most substantive change from Study 3 was the extension of the app from 3 modules per pedagogical concept to 5. This design extended the number of opportunities for TAs to practice a unit-level concept. It also included explicit requests for TAs to describe roadblocks they were encountering; prompting them to share solutions to those roadblocks.

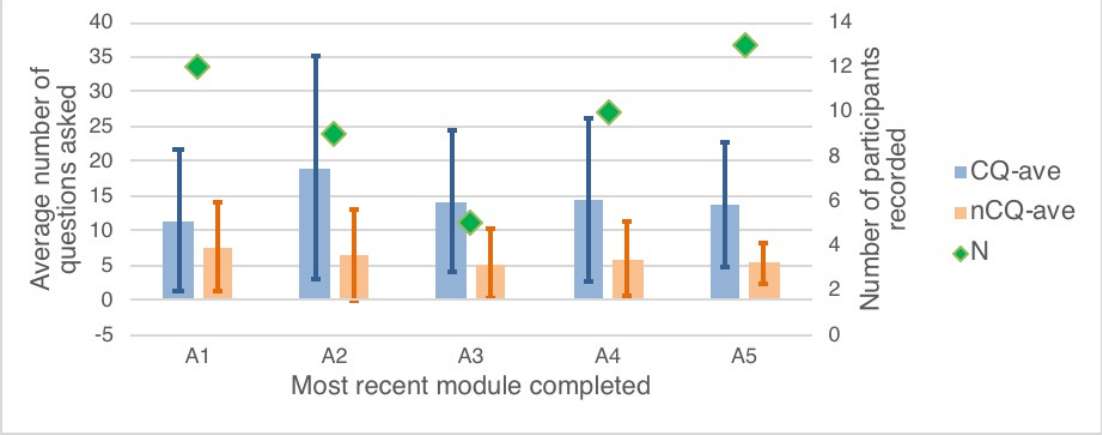


Figure 7.11: Average number of content questions asked per most recently completed Unit A module. Diamonds indicate the number of participants recorded at each stage. Blue bars represent average number of content questions all TAs at that stage asked. Orange bars represent the average number of non-content questions. Error bars are standard deviations for each module-level sample of participants.

The primary question I wanted to answer with this design choice was if it deepened the learning experience for TAs by providing more opportunities to express behavior changes and more opportunities to reflect. Unfortunately, there was no clear answer. From the raw data, it became apparent that there were too many interacting variables. This make a summative evaluation of RQ3 at this stage impossible. For example, Figure 7.11 indicates that there was very little change across modules for the number of questions that TAs asked. This graph shows the average number of questions that TAs asked following a specific module (horizontal axis). In other words, if a TA taught after taking Unit A Module 1, but not before Unit A Module 2, then the data from that class are included in the graph above A1. The problem here was that not all TAs would have completed the module on the same week of the semester. Furthermore, the app was designed such that TAs *should* have always done a planning module (A2 and A4) before teaching. There should not have been any observation corresponding to A1 or A3. Clearly that was not what happened, likely due to the self-pace nature of *ClassInsight*.

A "correct" version of Figure 7.11 would show the same N on each record, and it would only have data corresponding to A2, A4, and A5. Given how the app was actually used, as well as the small number of participants, it is impossible to make a quantitative statement about the value of including roadblock reflection as part of SmartPD.

Qualitatively there is a bit more that we can say. TAs showed a broad range of reflections on roadblocks (Table 7.20). Some of their comments were astute, such as TA15 who reported, "It was mostly the same students who kept on answering my questions." A self-guided solution to this problem was, "I could call on people who don't volunteer to answer questions." However, when asked in A5 (the module that followed the next teaching opportunity) if they had tried the strategy, TA15 reported that they had not, and that it was too hard.

TA17, on the other hand, did not produce a particularly insightful reflection on roadblocks. They said the cause for not getting students to talk that week was from, "It being early in the morning." When asked for a plausible strategy, they said, "No idea."

In general, the roadblocks TAs produced were sincere and provided qualitative insight into their individual teaching experience. As it is currently designed, SmartPD did not prove to be a strong follow-up resource. This may indicate a need for stronger designs in meta-level reflection in future design iterations.

Table 7.20: Each TAs exact input for two questions from Module A4.

	What roadblocks did you encounter this week with trying to get students to speak up?	**What strategy do you think might be an effect way to address the roadblock?**
TA03	Since I was lecturing in new material, it was harder to ask question and get answers.	Go back to doing problems.
TA05	Hmm, I am not sure how to make everyone more comfortable to speak	Not sure
TA06	Getting students to speak up. Not many people were there this week (midterms), so there were fewer people to answer.	Wait for another week. Maybe also spacing questions out more.
TA07	*Did not take this module*	
TA08	The students were unusually quiet.	I waited for answers longer, so will continue to try this.
TA09	*Did not take this module*	
TA12	*Did not take this module*	
TA13	They didn't know where I was going.	Give more useful hints.
TA14	They already knew the answer to some of the questions.	Ask questions that weren't explicitly covered in lecture.
TA15	It was mostly the same students who kept on answering my questions.	I could call on people who don't volunteer to answer questions.
TA17	It being early in the morning	No idea
TA18	As usual, when I ask them how to approach the problem, nobody would speak up and suggest an idea.	I'm not sure. It seems to me that they have a pretty good understanding of the concepts in general. Maybe I can let them talk to each other briefly, and then I'll ask the questions.
TA19	Disengaged students not feeling like participating today, especially in my latest recitation, students who didn't know quite what I was asking / were unsure and so didn't want to speak up.	Not sure. This is a pretty fundamental hurdle in teaching.
TA20	not many, people were pretty vocal	break up questions into smaller parts
TA21	Most left after taking the quiz; the ones that remained were quiet	Other than trying to make the recitation as relevant as possible, I am unsure
TA22	Hard questions, they didn't know where to begin.	Scaffold the problem solving more.

7.5 Discussion

This study of SmartPD included a much larger sample of TAs than Study 3. Through a deep, qualitative, and exploratory review of the data I showed that those who used the app *ClassInsight* were more likely to implement discursive teaching strategies than those who never used it. This gives a strong indication that the PAR-TS framework and the current iteration of SmartPD are successful; providing learners with positive outcomes.

Surprisingly, those in the *Dropout* group were almost as likely to produce positive results as those in *B_Group*. This may be partially explained by the fact that high student-focus at pre-test correlated to high rates of TAs leaving the app. It is possible that participants who began the study with high learner-centeredness saw what the app was promoting, and they decided to make the corresponding changes on their own. Unfortunately, leaving the app early also correlated to a *loss* of that same student-centeredness. It may be that in their eagerness to implement strategies without any external support, these TAs ended up losing some faith in their students.

For those who did try the app, and especially those who completed it, there was a strong relationship between trying the suggested tactics and losing teaching-centeredness (Table 7.19). That is, when TAs tried what the app suggested, their teacher-focus went down at post-test. We might have expected that changes in attitude would have required that TAs first try changing their behavior and then seeing their students change in response. However, intrinsic motivation to change seemed to be sufficient, because changes in student behavior did not correlate to changes in teacher-focus. This was likely a benefit to these TAs and their students, as the TAs seemed to change their self-conception by noticing their own efforts rather than by needing students to produce external motivation.

Another surprise was that TAs who were high in self-efficacy did not show blanket improvements in teaching performance. This final study showed that the story is more complicated than it seemed to be in Study 3. *B_Group* showed a relationship between self-efficacy and confirmatory assessment (Table 7.17), which was not ideal, but does make intuitive sense given that both rely on some measure of self-confidence. From another perspective, high self-efficacy seemed to translate to more student talk, but *not* to the use of discursive strategies (

Table 7.18). This study did not measure every discursive tactic that exists, and perhaps these TAs were able to implement some of these. Or perhaps there is a more complex pathway between beliefs and actions than what this study could analyze. Either way, there seems to be a clear relationship to pre-test self-efficacy and the chances that students will talk more over time. And it seems that using SmartPD was able to enhance this phenomenon.

Changes in self-efficacy (TSES-post-pre) were even more unexpected. Negative correlations to changes in teaching behavior (Table 7.19) emerged for *B_Group* and All TAs. Similar to declines in teacher-focus, TAs who followed the app's suggestions reported *less* self-efficacy at post-test. This relationship was especially salient when TAs increased the pace of content questions. Perhaps trying to ask meaningful questions more frequently may leave TAs feeling less confident over time. This indicates a more complex story than what the raw data alone reveal.

TA13, for example, was a surprising case. By the standards of the intervention, this TA did everything right (Table 7.15). They asked more content questions, asked them more frequently, and waited longer. Their students answered more often and spoke more as the semester progressed. Despite all of this, TA13's self-efficacy decreased from 3.6 at pre-test to 3.3 at post-test. In the interview, they said that their confidence may have declined because, "I had forgotten how hard it can be to teach." Additionally, they said that the app was somewhat "irritating," because it was telling them to do things they already knew they should be doing. The TA said, "I know I *should* be doing that," but it took *ClassInsight* pointing it out to motivate them to change. This indicates that high self-efficacy may drop if a TA's sense of confidence is coming more in line with their lived experience, even when they're doing everything "right" from a metrics standpoint.

In another case, metrics and algorithms may have been the source of frustration. TA17, for example, was a *B_Start* who, like TA13, also improved in almost every measure (Table 7.15). They also decreased in self-efficacy from pre-test to post-test (3.3 to 2.7). However, where TA13 showed a reluctant agreement with *ClassInsight*, TA17 became strongly opposed to the intervention. In an in-app reflection, they wrote, "There is not a specific strategy that will work for every single topic that is covered in this class." They continued:

> ... Also, our recitation is heavily dependent on how well a lesson resonates in lectures during the week. **It is incredibly frustrating to be asked the same set of 'progress' questions** [*note: the reflection questions in modules 3 and 5*] **when there is no context being applied to them** (by context, I mean there is little mention of whether or not the material I will be teaching is conducive to whatever 'planning' [*modules 2 and 4*] i am being made to do here). I found that **I became more self-conscious about asking questions to my students**. However, the feedback that I've received from my students that come to my conceptual OH later in the week was that sometimes, waiting for questions, asking questions, etc (everything I am being 'measured' on here) is **less useful than going over the conceptual topics in detail, and THEN asking whether anyone has clarifying questions** etc. Again, all of this is HEAVILY dependent on the material being taught. And that is just not being reflected at all.
>
> — TA13's final reflection before abandoning the training after Unit B1. Emphasis added.

This TA clearly did not appreciate what they interpreted *ClassInsight* as trying to accomplish. They were frustrated and became self-conscious while teaching. Some clarification emerges when the TA mentions that it can be better to go over conceptual topics in detail "...and THEN asking whether

anyone has any clarifying questions etc." This one comment indicates that the TA was possibly missing the point of the experience. The app asked TAs to use more open questions, not to stop in the middle and ask if students had any clarifying questions. If the TA mis-interpreted the suggestion, or if the app was difficult to understand, then it seems obvious that this would be an unnatural and potentially frustrating experience for anyone.

All of this suggests that designers of SmartPD must be careful in how they present tactics and goals to their learners. They should anticipate possible points of confusion and make clear what is being asked of them. Additionally, it might be helpful to allow users to "skip" modules or units that they find unhelpful. Supporting self-regulated learning could benefit the users. One way to implement this would be to give learners a list of topics and allow them to choose which they would prefer to address rather than force them to follow a specific curriculum, as the intervention currently does.

There was also evidence that TA17 was participating in the app while not fully engaging with useful critical self-reflection. As mentioned in the findings, this TA said that students did not participate because class was early in the morning. The TA could not or would not produce a single idea of how to address this problem. As indicated previously, SmartPD did not have the resources TAs needed in order to address their roadblocks.

These outcomes suggest that there is space for collaboration between designers of SmartPD systems and existing consultation centers. Instances such as those produced by TA17 are good candidates for human intervention, helping users succeed in ways that the technology cannot. Indeed, the tone of TA17's voluntary criticism of the system indicated that algorithmic approaches to teacher PD may be patently objectionable to some users. A "human-in-the-loop" design, however, may help ameliorate these concerns without abandoning the advantages of self-regulated PD.

Finally, for the majority of participants that were not overly frustrated with SmartPD, there were still many mixed outcomes. The variance in the use of silence, for example, was a surprising outcome. Perhaps the app did not deliver a clear explanation of the value of wait time. Or perhaps this variable is of a type that needs meta-level reflection. It seems like that there will be times when TAs need to think more deeply about the strategies they are trying, not just the tactics. Future research in this space promises to be an exciting area with many opportunities to extend the applications and conventions of socio-technical systems for providing feedback and instructional development—not just to teaching assistants, but to many other professionals as well.

7.5.1 Investigation with a Path Model

I began this chapter describing the research space as a collection of reflective and formative variables (Sanchez, 2013) that build up into the high-level domains of the Interconnected Model of Professional Growth (Clark & Hollingsworth, 2002). I drew inspiration for this model from a method used to build pathway analyses of multi-variate space. As a final analysis of face validity, and to support future researchers in confirming their findings, this section explores one way to test a model of this type.

I programmed a partial-least squares model in R in order to draw the direct and indirect relationships between latent variables (Sanchez, 2013) in a simplified version of the current research space (see Appendix E). The model in Figure 7.12 proposes that beliefs and attitudes at pre-test have a direct impact on student-focus at post-test, and an indirect impact via one in-class outcome, i.e., Student/TA talk ratio. The path analysis for a model that includes *B_Group* participants only demonstrates that "Beliefs_pre" (TSES, ITTF, and CCSF at pre-test) has a positive correlation to CCSF at post-test, but not as much as it does to changes in ST-TA ratio and the resulting positive correlation to attitude

outcomes. (I use 'ST//TA Ratio' and 'CCSF-Post' here only as examples of variables worth considering.)

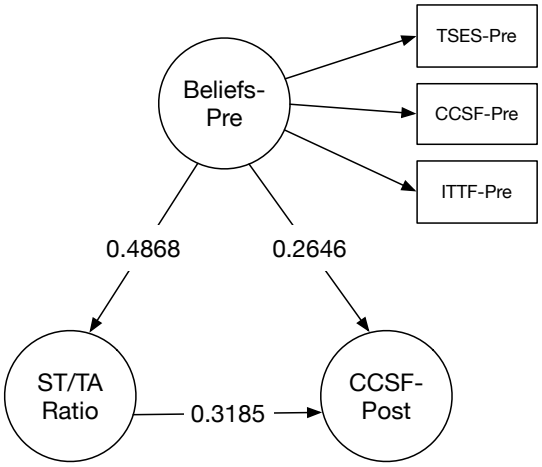


Figure 7.12: Simple Latent variable model of pretest measures of belief, Student talk, and student-focus post-test with the relevant path coefficients (*B_Group* only).

Using this same model while including all of the participants, rather than just *B_Group* (Figure 7.13), shows that in general prior beliefs have a much larger impact on CCSF, and that changes in student talk (i.e., the ST-TA pathway) does not contribute to how TAs see their students at the end of the semester. Comparing these two models, while not statistically meaningful, makes the pathway model for *B_Group* more compelling, and suggests that with more data and a larger sample, researchers could show many more details about how TAs learn and change.

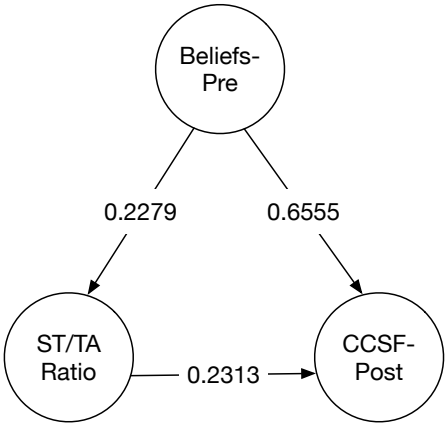


Figure 7.13: Simple Latent variable model of pretest measures of belief, Student talk, and student-focus post-test with the relevant path coefficients (*all TAs*).

7.5.1.1 Proposed PLS model

These example models of partial least square regressions suggest that the research space I outline in Figure 7.1 is prototype of a testable research space. In Figure 7.14 I lay out a more explicit confirmatory model that would assess the correlation coefficients between these different levels of phenomena. The model applies the domains of the Interconnected Model of Professional Growth as a testable structure, which accounts for a wide array of variables. For example, because high levels of

self-efficacy did not correlate to TAs following the advice of *ClassInsight*, it is possible that this variable has an influence on the likelihood that TAs express orientations toward Confirmatory Assessment. A pathways model of this type, with a larger dataset, could help uncover this relationship, if it exists. It could certainly uncover many more useful insights about how people learn to teach.

Adapting the learning experience such that it follows the evidence of a complex, interdependent model would be a promising direction for this research. Perhaps the curriculum should be open-ended and allow users to select their own pathway. However, it is not yet clear if self-efficacy, student-centeredness, or pedagogical skill should come first in the experience. Is change more likely when a TA is *able* to enact a goal, simply *thinks* that they are able of enacting it, or that they simply *believe* the goal matters? The model that I propose here, or something similar, would allow researchers to adapt this work to their particular situations. This kind of insight is an important contribution from rigorous DBR work (McKenney & Reeves, 2012). While DBR findings are not typically "generalizable," the findings from one educational domain should be "transferable" to many others. I hope that this tool helps make that possible.

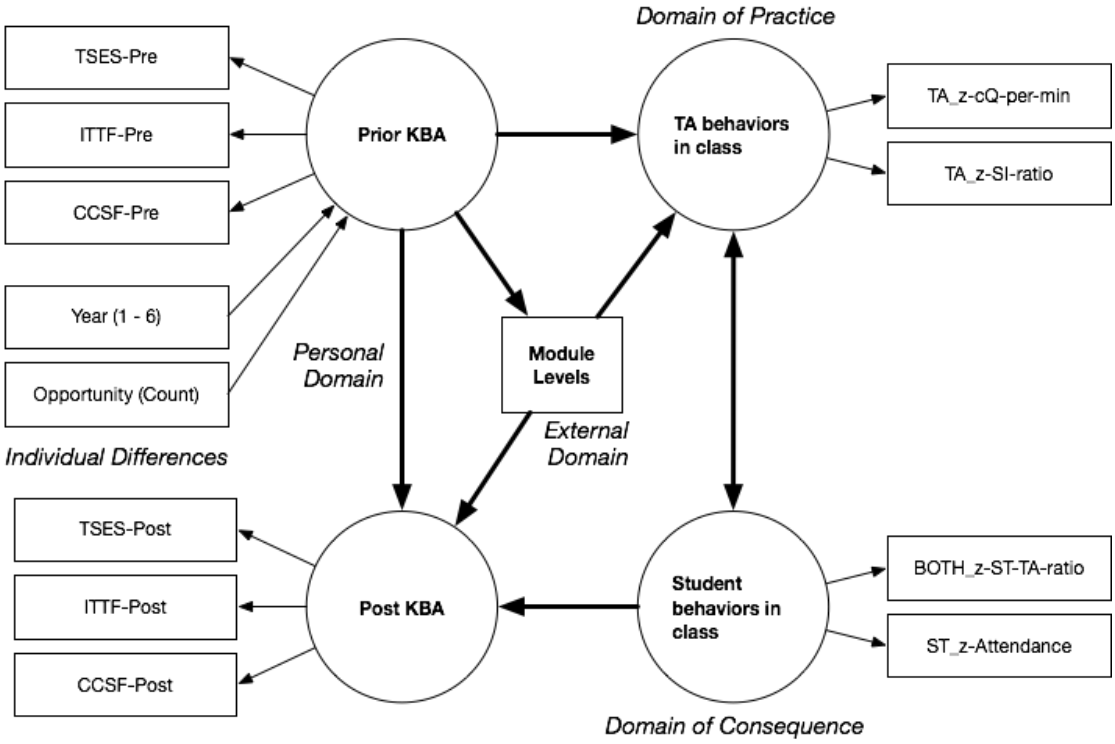


Figure 7.14: Confirmatory model of the PAR-TS Framework based on Study 4 data.

These findings offer an initial roadmap for the types of things researchers need to address as they pull together point sensing solutions into comprehensive systems for supporting teacher training. Additionally, this research offers a plausible framework in which to design similar systems. Future research may explore this framework by:

- Instituting more pedagogical concepts
- Expanding it by adding more approaches to conceptual change and learning, such as including collaborative engagement or simulated practice opportunities, for example
- Expanding it to account for more individual differences in instructors

Attending to attitudes toward how student learn and beliefs in self-efficacy may begood places to start. However, there are many other instruments and other psychological and social phenomena that should inform what we continue to learn about how people learn to teach within a complex, iterative, feedback and experience-oriented socio-technical learning system.

Chapter 8: Conclusion

The work I describe in this thesis draws from research on college instruction, professional development for teachers, personal informatics, and smart classrooms. I combined the unique strengths of each of these fields to generate a novel solution to some of the pressing limitations of college instruction.

8.1 Overview of the research

I began this work with a review of relevant literature and a field study on TAs. I wanted to understand the TA experience and explore opportunities for technology to help improve it. I read about low-cost, high-impact interventions from the field of professional development for teachers and college instructors. From a wide range of teaching tactics, I found that small steps such as waiting longer after asking questions or changing the types of questions they ask helps instructors improve their students' learning. I compared these discursive strategies to research in live detection of classroom behaviors and postulated that these types of activities are amenable to current trends in classroom sensor research. These trends include cutting-edge research in classroom detection of spoken turn-taking, pauses between speech events, as well as the presence and types of questions that instructors ask.

After pinpointing this connection between classroom behaviors and plausible detection technologies, I launched a field study where I manually logged classroom behaviors to simulate those sensors. I used the collected data to provide feedback to TAs on their teaching performance in terms of discursive teaching practices, such as how much their students were talking and how many questions the TAs were asking. I found that although TAs had few external incentives to improve their teaching methods on their own, encountering their own visualized data increased their awareness of student disengagement and motivated them to ask about concrete steps they might take to improve.

The studies which followed this initial investigation continued to explore the experience of TAs and how technology could support their use of discursive teaching. The population of interest included TAs who lead regular recitation sections with fewer than 35 students. I chose courses in STEM, where classes are most likely to issue instruction via information transmission rather than active participation. I observed small classrooms because they have a higher potential for meaningful student interactions.

I provided instructors with regular, targeted feedback on their in-class behaviors in order to help them improve their teaching. I chose behaviors that are theoretically detectable and pedagogically meaningful. I chose teaching moves that occur frequently—candidate variables for producing large datasets of what happens in the classroom and how it changes over time. I generated concrete, measurable user interactions, designed to be easy for instructors to understand and address.

To support data collection for this research, I have built observational protocols that make it simple for observers to quickly capture and catalog live interactions for fast processing. This allowed me to simulate advanced classroom sensors in order to isolate the design and delivery of feedback. Before spending years and dollars on building advanced detection systems, researchers and developers should understand the impact such a detection system might have. This dissertation measured that impact in terms of changes in instructor beliefs, attitudes, and behaviors.

8.2 Major findings

In the preliminary field study, I asked if there were meaningful TA behaviors that would be theoretically detectable, and how TAs might react to seeing visualizations of those data. I discovered that students in classes led by TAs were doing very little speaking, and TAs were not asking deep questions. Following the principles of Personal Informatics, I gave TAs visualizations of their raw data without giving them explicit direction on how to leverage that information. TAs were able to recognize that their students did not seem incentivized to speak. Unfortunately, the TAs were not able to generate solutions to this problem.

This was a potentially surprising result for researchers of Personal Informatics in other domains. Typically, the availability of data is treated as sufficient on its own under the assumption that it causes users to reflect on their past and set goals for their future. TAs, however, are selected as *domain* experts, not *instructional* experts. They do not necessarily know what behaviors they should implement, nor how to reflect on feedback targeting those behaviors. Furthermore, they are pressed for time and reluctant to change habits when they are simply doing what they have always seen done. These conditions limit the applicable theory of change outlined by PI. This suggested that the problem of PD and classroom sensors had additional depth worth exploring.

In Study 2, I tested whether TAs would improve their teaching behaviors through the use of a digital developmental platform. I removed the use of direct feedback and focused instead on building an algorithmic approach to delivering sophisticated professional development experiences that followed best practices from traditional education for teaching. Using iterative cycles of design-based research, the emerging platform encouraged changes in teaching behaviors, and influenced what TAs knew and believed about teaching. It did not, however, produce reliable self-awareness of teaching patterns, or clear ideas about what was working for their students.

With the development of the PAR-TS framework and the articulation of SmartPD, I explored the depths of TA beliefs, change, and resistance to change in the next two studies. The data showed that TA beliefs matter—partially because in many cases, those beliefs are out of alignment with what they do. TAs, as novice instructors, do not necessarily perceive themselves accurately. In study 3 I began to uncover their orientation toward teaching and learning. In Study 4 I found ways in which their beliefs and attitudes impacted their approach to instructional development and how those approaches affected their ongoing beliefs.

It turns out that when TAs are open to the experience, SmartPD can be a low-cost approach to reducing teacher-centeredness. For those TAs who follow the suggestions, finish the program, and see their students change, increases in student-focus may follow. Surprisingly, self-efficacy did not predict the adoption of new practices as measured in this work, although it did predict increasing rates of student talk. This shows some differences from what self-efficacy for teaching typically signifies in the larger field of teacher education. That field of work shows strong correlations, for example, between self-efficacy and self-motivation—or at least the ability to avoid burnout (Schwarzer & Hallum, 2008). But the current population, one which has a major impact on the success of thousands of students every year, is not one that contends with problems that face teachers late in their careers. They represent a different population that has been poorly researched and given sparse support.

8.3 Putting these findings to work

The solution I propose is a new genre of socio-technical training system which I call SmartPD. The following list shows the high-level objectives of SmartPD and the research fields from which each objective inherits inspiration and guidance. (Professional development for teachers = PD, "Smart classrooms"/technology-enhanced learning = SC, and Personal Informatics = PI.)

SmartPD has many potential goals:

- Gather behavioral data from real-world classrooms (PD and SC)
- Discriminate and classify pedagogically meaningful behaviors (PI and PD)
- Support grounded reflection on users' implementation of behaviors (PI)
- Provide direct support in learning new pedagogical ideas (PD)
- Motivate changes in beliefs and actions (PD and PI)

Going deeper into the overlapping contributions of each field, the unsupervised and scalable aspects of personal informatics and classroom sensors inform professional development such that it scales to a large set of users. In return, professional development offers personal informatics a more sophisticated view of users, doing more than delivering data visualizations. It also supports self-regulated learning and belief change. Finally, the combination of personal informatics and professional development for teachers can help designers of smart classrooms focus on developing detection technologies which show pedagogical merit.

8.3.1 Who can use these findings

The research I describe in this document combines and extends the PI and PD aspects of SmartPD. Those who develop classroom instrumentation in the future should be able to focus their work based in part on my findings. If they decide to design sensors that have a theoretical basis, this research will help guide their work. For the purposes of my research, I simulated the SC through the use of human observers operating live classification tools, but at each step these tools were designed to simulate realistic output of sensors, and should extend to fully instrumented smart classrooms.

My work may also outline pathways for data scientists to use as they investigate learning spaces. The learning sciences have made strides in online learning spaces, and educational psychology reveals a great deal about how people learn. Each group of researchers may find that the current work reveals research possibilities that emerge when gathering large amounts of data from a real-world learning space. Technology-enhanced learning spaces can do more than just support or train instructors. They may also help develop theories about the underlying conditions which support changes in teachers and students as they operate within their lived context.

This work shows promise in developing new, scalable solutions to the training gap in higher education. It also has implications for advancing theoretical explanations of how people learn to teach. By gathering and organizing large datasets from real classrooms across many different contexts, learning scientists will be able to (a) uncover previously invisible trends in how teaching and learning interact, and (b) test new hypotheses about learning in very large samples of the population.

My design-based research uncovers theoretically detectable variables of classroom interaction that can support TAs in self-reflection, goal-setting, and changes in belief and behavior. I try to create interactions that elicit the knowledge of the user and open an additional channel of communication

among his or her community of practitioners. I use this approach to honor the expertise of the individual instructor, and to avoid the need for a robust model of each individual classroom or course topic. My thesis reveals some of the roadblocks that future designers of SmartPD will need to overcome; hopefully it also provides grounded insights for how to approach these limitations.

This dissertation focuses on (a) the practicality of this vision, e.g., the viability of current and near-term sensors to provide meaningful data and the scalability of heuristic training principles, and (b) how to design for enhancing the value that TAs might take from such an approach. The work provides some insights to these questions while it also generates potential ongoing research avenues for the learning sciences.

8.3.2 Where this work goes next

My broad vision of SmartPD is that of a socio-technical system that combines features of traditional PD and personal informatics. I laid the groundwork for personalized learning that future researchers can advance. SmartPD helps users discover new concepts and reflect on the practicality of those ideas. It provides feedback to support reflection, goal-setting, and behavior change. It automates substantive parts of the educational development process. It prompts users with text to read, questions to answer, and data visualizations to explore. For now, the use of broad heuristics about the population has made it possible for me to produce a framework for how to build such systems, as well as perform qualitative assessments of many design principles.

Future researchers will be able to refine the framework and design principles as they learn more about how to adapt instructional development to account for individual differences, or how to deliver real-time prompts in live classrooms. To deliver on these promises, SmartPD will come to implement live sensors to build a database of real-world actions. It will use these data to build a hypothetical model of the skills and knowledge of the instructor. And it will provide developmentally appropriate feedback and support for goal-setting.

In the context of this research, I investigate PD as it relates to formal learning spaces in a single American university. There is no reason to limit SmartPD this way in the future. The work could grow into collaborations with existing centers of teaching, learning, and instructional development for TAs and faculty. It may expand to support aspiring professional teachers in schools of education. It could certainly support practicing K-12 teachers.

SmartPD could also add value to other learning spaces, be they formal, informal, scholastic, or industrial. Learning to teach may not be all that different from learning other complex social tasks. It might be able to support, say, new management team members learning to lead their staff, or budding entrepreneurs learning to share their big ideas.

The methods of SmartPD could grow as well. Researchers of adaptive, intelligent tutoring systems are likely to become interested in it as a learning space. Implementing robust models of what users know could enhance self-regulated learning beyond more than just independent, “outer loop” unit selection. It could prompt learners to discover hints when they are reaching roadblocks in their current learning opportunities.

So far, the practical conditions I describe in this document are all post-class reflections. But there is no reason that the scaffolding of SmartPD could not be introduced to the live teaching environment. In fact, given the trajectory of technology it is likely that this will happen without much push from researchers. One aspect of this work that I have not explored is the set of questions around what it

means to put data collection systems in classrooms. For example, who owns the data? Are the instructors in charge, or their hiring bodies? Do the data ultimately become a way of simply weeding out sub-par performers, or will designers choose to continue to frame the user as the learner?

My hope on behalf of the teachers, TAs, professors, cab drivers, or widget spinners that end up learning new skills through this kind of technology is that they feel supported rather than suppressed by the socio-technical systems they must use. They should see themselves as co-authors of their learning experience—collaborators in their experience development. This not only reinforces their position as the human subject in the enterprise but makes good practical sense in terms of maintaining self-worth. Our machines should support our humanity and extend our autonomy. I hope that SmartPD moves in that direction.

References

Akiha, K., Brigham, E., Couch, B. A., Lewin, J., Stains, M., Stetzer, M. R., … Smith, M. K. (2018). What Types of Instructional Shifts Do Students Experience? Investigating Active Learning in Science, Technology, Engineering, and Math Classes across Key Transition Points from Middle School to the University Level. *Frontiers in Education*, *2*(January).

Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How Learning Works: Seven Research-Based Principles for Smart Teaching*. Jossey-Bass.

Andrew, A., Borriello, G., & Fogarty, J. (2011). Understanding Self-Efficacy and the Design of Personal Informatics Tools. In *CHI '11—Workshop on Personal Informatics in Practice: Improving Quality of Life Through Data*. Vancouver, BC, Canada: ACM.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(2), 356–370.

Austin, A. E. (2002). Preparing the Next Generation of Faculty: Graduate School as Socialization to the Academic Career. *The Journal of Higher Education*, *73*(1), 94–122.

Bain, K. (2004). *What the Best College Teachers Do*. Harvard University Press.

Baker, P. J., & Zey-Ferrell, M. (1984). Local and Cosmopolitan Orientations of Faculty: Implications for Teaching. *Teaching Sociology*, *12*(1), 82–106.

Baker, R. S. J. d., & Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), *Cambridge Handbook of the Learning Sciences*. Cambridge, UK: Cambridge University Press.

Bakker, S., van den Hoven, E., & Eggen, B. (2014). Evaluating Peripheral Interaction Design. *Human–Computer Interaction*, *30*(6), 473–506.

Ball, D. L. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389–407.

Bandura, A. (1977). Self-efficacy: Toward a Unifying Theory of Behavioral Change. *Psychological Review*, *84*(2), 191–215.

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. New York: W.H. Freeman & Company.

Barab, S. (2014). Design-Based Research: A Methodological Toolkit for Engineering Change. In *Handbook of the Learning Sciences* (pp. 151–170).

Bautista, G., & Borges, F. (2013). Smart classrooms: Innovation in formal learning spaces to transform learning experiences. *Bulletin of the Technical Committee on Learning Technology*, *15*(3), 18–21.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*(2), 131–160.

Bell, A., & Mladenovic, R. (2008). The benefits of peer observation of teaching for tutor development. *Higher Education*, *55*(6), 735–752.

Berman, J., & Skeff, K. M. (1988). Developing the motivation for improving university teaching. *Innovative Higher Education*, *12*(2), 114–125.

Beyer, H., & Holtzblatt, K. (1998). *Contextual Design: Defining Customer-Centered Systems*. San Francisco: Morgan Kaufmann Publishers.

Blanchard, N., Donnelly, P., Olney, A. M., Borhan, S., Ward, B., Sun, X., … D'Mello, S. K. (2016). Identifying Teacher Questions Using Automatic Speech Recognition in Classrooms. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 191–201).

Blau, I. (2011). Teachers for "Smart Classrooms": The Extent of Implementation of an Interactive Whiteboard-based Professional Development Program on Elementary Teachers' Instructional Practices. *Interdisciplinary Journal of E-Learning and Learning Objects*, *7*, 275–288.

Bolen, J. (2009). *An Investigation of Limited Professional Development on Teacher Questioning and Learner Responses*. Walden University.

Borg, M. (2004). The apprenticeship of observation. *ELT Journal*, *58*(3), 274–276.

Boud, D. (2001). Using journal writing to enhance reflective practice. *New Directions for Adult and Continuing Education*, *2001*(90), 9.

Boyd, A., Gorham, J. J., Justice, J. E., & Anderson, J. L. (2013). Examining the apprenticeship of observation with preservice teachers: The practice of blogging to facilitate autobiographical reflection and critique. *Teacher Education Quarterly*, *Summer*, 27–49.

Brinko, K. T. (1993). The Practice of Giving Feedback to Improve Teaching: What Is Effective? *The Journal of Higher Education*, *64*(5), 574.

Brown, A. L. (1992). Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *The Journal of the Learning Sciences*, *2*(2), 141–178.

Brownell, S. E., & Tanner, K. D. (2012). Barriers to faculty pedagogical change: Lack of training, time, incentives, and...tensions with professional identity? *CBE—Life Sciences Education*, *11*(Winter), 339–346.

Bureau of Labor Statistics U.S. Department of Labor. (2016). Occupational Employment Statistics. https://www.bls.gov/oes/current/oes251191.htm

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The Effect of Type and Timing of Feedback on Learning From Multiple-Choice Tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273–281.

Caldwell, J. E. (2007). Clickers in the large classroom: Current research and best-practice tips. *CBE—Life Sciences Education*, *6*(Spring), 9–20.

Chen, G., Clarke, S. N., & Resnick, L. B. (2014). An Analytic Tool for Supporting Teachers' Reflection on Classroom Talk. In *Proceeding of the International Conference of the Learning Sciences* (pp. 583–590). Boulder, CO: International Society of the Learning Sciences.

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, *49*(4), 219–243.

Chism, N. V. N., Holley, M., & Harris, C. J. (2012). Researching the impact of educational development: Basis for informed practice, *31*(1), 129–145.

Choe, E. K., Lee, N. B., Lee, B., Pratt, W., & Kientz, J. A. (2014). Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the 32nd annual ACM Conference on Human factors in Computing Systems - CHI '14* (pp. 1143–1152). New York, New York, USA: ACM Press.

Clark, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teaching Education*, *18*, 947–967.

Clawson, J., Pater, J. A., Miller, A. D., Mynatt, E. D., & Mamykina, L. (2015). No longer wearing: Investigating the abandonment of personal health-tracking technologies on craigslist. *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 647–658.

Clow, D. (2012). The Learning Analytics Cycle: Closing the loop effectively. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, 134–138.

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *The Journal of the Learning Science*, *13*(1), 15–42.

Comber, R., Thieme, A., & Rafiev, A. (2013). BinCam: Designing for engagement with Facebook for behavior change. *Human-Computer Interaction-INTERACT 2013*, 99–115.

Consolvo, S., Klasnja, P., McDonald, D. W., & Landay, J. A. (2009). Goal-setting considerations for persuasive technologies that encourage physical activity. In *Proceedings of the 4th International Conference on Persuasive Technology*. Claremont, CA.

Consolvo, S., McDonald, D. W., Toscos, T., Chen, M. Y., Froehlich, J., Harrison, B., … Landay, J. A. (2008). Activity sensing in the wild: A field trial of UbiFit Garden. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1797–1806). Florence, Italy: ACM.

Cox, B. E., McIntosh, K. L., Reason, R. D., & Terenzini, P. T. (2011). A Culture of Teaching: Policy, Perception, and Practice in Higher Education. *Research in Higher Education*, *52*(8), 808–829.

Cox, M. D. (2004). Introduction to faculty learning communities. *New Directions for Teaching and Learning*, *97*(97), 5–23.

D'Mello, S., Picard, R. W., & Graesser, A. (2007). Toward an Affect-Sensitive AutoTutor. *IEEE Intelligent Systems*, *22*(4), 53–61.

Dillenbourg, P., & Jermann, P. (2010). Technology for classroom orchestration. *New Science of Learning: Cognition, Computers and Collaboration in Education*, 525–552.

Duval, E. (2011). Attention please! Learning Analytics for Visualization and Recommendation. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge - LAK '11*, (November), 9–17.

Dyke, G., Adamson, D., Howley, I., & Rosé, C. P. (2013). Enhancing scientific reasoning and explanation skills with conversational agents. *IEEE Transactions on Learning Technologies*, *6*(3), 240–247.

Easterday, M. W., Lewis, D. R., & Gerber, E. (2014). Design-based research process: Problems, phases, and applications problems arising from the ill-definition of DBR. *Proceedings of the International Conference of the Learning Sciences*, 317–324.

Ellis, J., Deshler, J., & Speer, N. (2016). Supporting instructional change: A two-pronged approach related to

graduate teaching assistant professional development. In *Proceedings of the 19th Annual Conference on Research in Undergraduate Mathematics Education*.

Ellis, K. (1993). Teacher Questioning Behavior and Student Learning: What Research Says to Teachers. *Convention of The Western States Communication Association*.

Epstein, D. A., Caraway, M., Johnston, C., Ping, A., Fogarty, J., & Munson, S. A. (2016). Beyond Abandonment to Next Steps: Understanding and Designing for Life after Personal Informatics Tool Use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 1109–1113). San Jose.

Epstein, D. A., Ping, A., Fogarty, J., & Munson, S. A. (2015). A lived informatics model of personal informatics. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*, 731–742.

Fairweather, J. S. (1993). Faculty reward structures: Toward institutional and professional homogenization. *Research in Higher Education*, *34*(5), 603–623.

Finelli, C. J., Ott, M., Gottfried, A. C., Hershock, C., O'neal, C., & Kaplan, M. (2008). Utilizing instructional consultations to enhance the teaching performance of engineering faculty. *Journal of Engineering Education*, *97*(4), 397–411.

Freeland, R. (2007). *Collected Wisdom: Strategies & Resources for TAs*. Pittsburgh, PA: Eberly Center for Teaching Excellence.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8410–8415.

Friedman, J. (2017). 10 Universities Where TAs Teach the Most Classes. https://www.usnews.com/education/best-colleges/the-short-list-college/articles/2017-02-21/10-universities-where-tas-teach-the-most-classes

Fritschner, L. M. (2000). Inside the undergraduate college classroom: Faculty and students differ on the meaning of student participation. *The Journal of Higher Education*, *71*(3), 342–362.

Fritz, T., Huang, E. M., Murphy, G. C., & Zimmermann, T. (2014). Persuasive Technology in the Real World: A Study of Long-Term Use of Activity Sensing Devices for Fitness. *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, 487–496.

Froehlich, J. E. (2011). *Sensing and Feedback of Everyday Activities to Promote Environmental Behaviors*. University of Washington.

Gall, M. D. (1970). The Use of Questions in Teaching. *Review of Educational Research*, *40*(5), 707–721.

Galloway, C. G., & Mickelson, N. I. (1973). Improving Teachers' Questions. *The Elementary School Journal*, *74*(3), 145–148.

Goodyear, P., & Retalis, S. (Eds.). (2010). *Technology-Enhanced Learning: Design Patterns and Pattern Languages* (Vol. 2). Rotterdam/Boston/Taipei: Sense Publishers.

Gormally, C., Evans, M., & Brickman, P. (2014). Feedback about Teaching in Higher Ed: Neglected Opportunities to Promote Change. *Cell Biology Education*, *13*(2), 187–199.

Govaerts, S., Verbert, K., Duval, E., & Pardo, A. (2012). The student activity meter for awareness and self-reflection. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (pp. 869–884).

Graesser, A. C., & Person, N. K. (1994). Question Asking During Tutoring. *American Educational Research Journal*, *31*(1), 104–137.

Greller, W., & Drachsler, H. (2012). Translating Learning into Numbers: A Generic Framework for Learning Analytics. *Educational Technology & Society*, *15*(3), 42 – 57.

Gullatt, D. E., & Weaver, S. W. (1997). Use of faculty development activities to improve the effectiveness of U.S institutions of higher education. In *22nd Annual Conference of the Professional and Organizational Development Network in Higher Eduation* (p. 33). Hines CIty, Florida.

Gustafson, K. L., & Branch, R. M. (1997). *Survey of Instructional Development Models* (Third). ERIC Clearinghouse on Information & Technology.

Guzzetti, B. J., Snyder, T. E., Glass, G. V., & Gamas, W. S. (1993). Promoting Conceptual Change in Science: A Comparative Meta-Analysis of Instructional Interventions from Reading Education and Science Education. *Reading Research Quarterly*, *28*(2), 117–159.

Hardré, P. L. (2005). Instructional design as a professional development tool-of-choice for graduate teaching assistants. *Innovative Higher Education*, *30*(3), 163–175.

Hardré, P. L., & Burris, A. O. (2012). What contributes to teaching assistant development: Differential

responses to key design features. *Instructional Science*, *40*(1), 93–118.
Hartman, H. J. (2001). Teaching metacognitively. In *Metacognition in Learning and Instruction: Theory, Research,and Practice* (pp. 149–169). Dordrecht, The Netherlands: Kluwer Academic Publishers.
Haskell, R. E., & Theall, M. (1997). Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century. *Education Policy Analysis Archives*, *5*(6).
Hatton, N., & Smith, D. (1995). Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education*, *11*(1), 33–49.
He, H. A., Greenberg, S., & Huang, E. M. (2010). One size does not fit all: Applying the transtheoretical model to energy feedback technology design. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 927–936). Atlanta, Georgia, USA: ACM.
Hellermann, J. (2002). The interactive work of prosody in the IRF exchange: Teacher repetition in feedback moves. *Language in Society*, *32*(01), 79–104.
Henderson, C., Beach, A., & Finkelstein, N. (2011). Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *Journal of Research in Science Teaching*, *48*(8), 952–984.
Hill, L., Kim, S. La, & Lagueux, R. (2008). Faculty Collaboration as Faculty Development. *Peer Review*, *9*(4), 17–20.
Hmelo-Silver, C. E. (2004). Problem-Based Learning: What and How Do Students Learn? *Educational Psychology Review*, *16*(3), 235–266.
Ho, A., Watkins, D., & Kelly, M. (2001). The Conceptual Change Approach to Improving Teaching and Learning: An Evaluation of a Hong Kong Staff Development Programme, *42*(2), 143–169.
Hoadley, C. P. (2002). Creating context: Design-based research in creating and understanding CSCL. *Engineering Education*, 453–462.
Holstein, K., McLaren, B. M., & Aleven, V. (2017). Intelligent Tutors as Teachers' Aides: Exploring Teacher Needs for Real-time Analytics in Blended Classrooms. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17*, 257–266.
Howard, J., Short, L. B., & Clark, S. M. (1996). Students' Participation in the Mixed-Age College Classroom. *Teaching Sociology*, *24*(1), 8–24.
Hubball, H., Collins, J., & Pratt, D. (2005). Enhancing Reflective Teaching Practices: Implications for Faculty Development Programs. *The Canadian Journal of Higher Education*, *35*(3), 56–81.
Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, *64*(4), 349–371.
Kamal, N., Fels, S., & Blackstock, M. (2011). Personal Health Informatics: What is the role for online social networks ? In *Proceedings of the SIGCHI conference on human factors in computing systems*. Vancouver, BC, Canada: ACM.
Kay, M., Choe, E. K., Shepherd, J., Greenstein, B., Watson, N., Consolvo, S., & Kientz, J. A. (2012). Lullaby: A Capture & Access System for Understanding the Sleep Environment. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (pp. 226–234). Pittsburgh, USA: ACM.
Kennedy-Clark, S. (2013). Research by Design: Design-Based Research and the Higher Degree Research student. *Journal of Learning Design*, *6*(2), 26–32.
Klassen, R. M., Bong, M., Usher, E. L., Chong, W. H., Huan, V. S., Wong, I. Y. F., & Georgiou, T. (2009). Exploring the validity of a teachers' self-efficacy scale in five countries. *Contemporary Educational Psychology*, *34*(1), 67–76.
Kreber, C. (2005). Reflection on teaching and the scholarship of teaching: Focus on science instructors. *Higher Education*, *50*(2), 323–359.
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134.
Land, S. M., & Hannafin, M. J. (1996). Student-Centered Learning Environments: Foundations, Assumptions, and Implications. In *Proceeding of selected Research and Development Presentation at the 1996 National Convention of the Association for Educational Communication and Technology* (pp. 395–400). Indianapolis, IN.
Larson, L. R., & Lovelace, M. D. (2013). Evaluating the Efficacy of Questioning Strategies in Lecture-Based Classroom Environments: Are We Asking the Right Questions? *Journal of Excellence in College Teaching*, *24*, 1–18.
Lazar, A., Koehler, C., Tanenbaum, J., & Nguyen, D. H. (2015). Why we use and abandon smart devices. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing -*

UbiComp '15, 635–646.
LeGros, N., & Faez, F. (2012). The Intersection Between Intercultural Competence and Teaching Behaviors: A Case of International Teaching Assistants. *Journal on Excellence in College Teaching*, *23*(3), 7–31.
Lemov, D. (2010). *Teach Like a Champion: 49 Techniques that Put Students on the Path to College (K-12)*. San Francisco, CA, USA: Jossey-Bass.
Li, I., Dey, A., & Forlizzi, J. (2010). A stage-based model of personal informatics systems. *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, 557.
Lin, J. J., Mamykina, L., Lindtner, S., Delajoux, G., & Strub, H. B. (2006). Fish'n'Steps: Encouraging Physical Activity with an Interactive Computer Game. In *International conference on ubiquitous computing* (pp. 261–278). Springer-Verlag.
Lindblom-Ylanne, S., Trigwell, K., Nevgi, A., & Ashwin, P. (2006). How approaches to teaching are affected by discipline and teaching context. *Studies in Higher Education*, *31*(03), 285–298.
Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705–717.
Lowell Bishop, J., & Verleger, M. (2013). The Flipped Classroom : A Survey of the Research. *Proccedings of the Annual Conference of the American Society for Engineering Education*, 6219.
Luft, J. A., Kurdziel, J. P., Roehrig, G. H., & Turner, J. (2004). Growing a Garden without Water: Graduate Teaching Assistants in Introductory Science Laboratories at a Doctoral/Research University. *Journal of Research in Science Teaching*, *41*(3), 211–233.
Marsh, H. W. (1984). Student evaluations of university teaching: Dimensionality, reliability, validity, potential biases, utility. *Journal of Educational Psychology*, *76*(5), 707–754.
Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52*(11), 1187–1197.
Martyn, M. (2007). Clickers in the Classroom: An Active Learning Approach. *Educause Quarterly*, (2), 71–74.
Mauksch, H. O. (1986). Teaching within institutional values and structures. *Teaching Sociology*, *14*(1), 40–49.
McConnell, J. J. (1996). Active learning and its use in computer science. *ACM SIGCSE Bulletin*, *28*, 52–54.
McDaniel, T. R. (1985). A Primer on Classroom Discipline : Principles Old and New. *The Journal of Adventist Education*, *47*(4), 22–35.
McKenney, S. E. (2001). *Computer-Based Support for Science Education Developers in Africa: Exploring Potentials*. University of Twente, Enschede.
McKenney, S. E., & Reeves, T. (2012). *Conducting educational design research*. Routledge.
McShannon, J., Hynes, P., Nirmalakhandan, N., Venkataramana, G., Ricketts, C., Ulery, A., & Steiner, R. (2006). Gaining retention and achievement for students program: A faculty development program. *Journal of Professional Issues in Engineering Education and Practice*, *132*(3), 204–208.
Michael, J. (2006). Where's the evidence that active learning works? *Advances in Physiology Education*, *30*(4), 159–167.
Nunn, C. E. (1996). Discussion in the College Classroom: Triangulating Observational and Survey Results. *The Journal of Higher Education*, *67*(3), 243–266.
Nyquist, J. D., & Wulff, D. H. (1996). *Working effectively with graduate assistants*. Sage Publications.
O'Neal, C., Wright, M., Cook, C., Perorazio, T., & Purkiss, J. (2007). The impact of teaching assistants on student retention in the sciences: Lessons for TA Training. *Journal of College Science Teaching*, *36*(5), 24–29.
Oliveira, A. W. (2010). Improving teacher questioning in science inquiry discussions through professional development. *Journal of Research in Science Teaching*, *47*(4), 422–453.
Osborne, J., Simon, S., Christodoulou, A., Howell-Richardson, C., & Richardson, K. (2013). Learning to argue: A study of four schools and their attempt to develop the use of argumentation as a common instructional practice and its impact on students. *Journal of Research in Science Teaching*, *50*(3), 315–347.
Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning (NCER 2007-2004)*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.
Pearson, L. C., & Moomaw, W. (2005). The relationship between teacher autonomy and stress, work satisfaction, empowerment, and professionalism. *Educational Research Quarterly*, *29*(1), 38–54.
Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research*, *74*(2), 215–253.
Penuel, W. R., Fishman, B. J., Haugan Cheng, B., & Sabelli, N. (2011). Organizing Research and Development at the Intersection of Learning, Implementation, and Design. *Educational Researcher*, *40*(7), 331–337.

Postareff, L., Lindblom-Ylänne, S., & Nevgi, A. (2007). The effect of pedagogical training on teaching in higher education. *Higher Education*, *23*, 557–571.
Prieto, L. R., & Altmaier, E. M. (1994). The Relationship of Prior Training and Previous Teaching Experience to Self-Efficacy among Graduate Teaching Assistants. *Research in Higher Education*, *35*(4), 481–497.
Prochaska, J., & Velicer, W. (1997). The Transtheoretical Model of Health Behavior Change. *American Journal of Health Promotion*, *12*(1), 38–48.
Prosser, M., & Trigwell, K. (2006). Confirmatory factor analysis of the Approaches to Teaching Inventory. *British Journal of Educational Psychology*, *76*, 405–419.
Prytula, M. P. (2012). Teacher metacognition within the professional learning community. *International Education Studies*, *5*(4), 112–121.
Rapp, A., & Cena, F. (2014). Self-monitoring and Technology: Challenges and Open Issues in Personal Informatics. In C. Stephanidis & M. Antona (Eds.), *Universal Access in Human-Computer Interaction* (pp. 613–622).
Rapp, A., & Cena, F. (2016). Personal informatics for everyday life: How users without prior self-tracking experience engage with personal data. *International Journal of Human Computer Studies*, *94*(August 2018), 1–17.
Redfield, D. L., & Rousseau, E. W. (1981). A Meta-Analysis of Experimental Research on Teacher Questioning Behavior. *Review of Educational Research*, *51*(2), 237–245.
Reymen, I. M. M. J., Hammer, D. K., Kroes, P. A., Van Aken, J. E., Dorst, C. H., Bax, M. F. T., & Basten, T. (2006). A domain-independent descriptive design model and its application to structured reflection on design processes. *Research in Engineering Design*, *16*(4), 147–173.
Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education*, *30*(4), 387–415.
Rivera-Pelayo, V., Munk, J., Zacharias, V., & Braun, S. (2013). Live interest meter: Learning from quantified feedback in mass lectures. In D. Suthers, K. Verbert, E. Duval, & X. Ochoa (Eds.), *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 23–27). Leuven, Belgium: ACM.
Rivera-Pelayo, V., Zacharias, V., Müller, L., & Braun, S. (2012). A framework for applying quantified self approaches to support reflective learning. *IADIS International Conference Mobile Learning 2012*, 123–131.
Rocca, K. A. (2010). Student Participation in the College Classroom: An Extended Multidisciplinary Literature Review. *Communication Education*, *59*(2), 185–213.
Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.
Rounds, P. L. (1994). Student questions: When, where, why and how many. In C. G. Madden & C. L. Myers (Eds.), *Discourse and Performance of International Teaching Assistants* (pp. 103–115).
Rowe, M. B. (1972). Wait-Time and Rewards as Instructional Variables: Their Influence on Language, Logic, and Fate Control. Part II-Rewards. *Journal of Research in Science Teaching*, *11*(4), 291–308.
Rowe, M. B. (1980). Pausing principles and their effects on reasoning in science. *New Directions for Community Colleges*, *1980*(31), 27–34.
Sahin, A. (2007). Teachers' Classroom Questions. *School Science & Mathematics*, *107*(1), 369–370.
Sanchez, G. (2013). PLS Path Modeling with R. *R Package Notes*, 235.
Schön, D. A. (1983). *The Reflective Practitioner: How Professionals Think in Action*. Basic Books.
Schwarzer, R., & Hallum, S. (2008). Perceived teacher self-efficacy as a predictor of job stress and burnout: Mediation analyses. *Applied Psychology*, *57*, 152–171.
Settlage, J., Southerland, S. A., Smith, L. K., & Ceglie, R. (2009). Constructing a doubt-free teaching self: Self-efficacy, teacher identity, and science instruction within diverse settings. *Journal of Research in Science Teaching*, *46*(1), 102–125.
Shavelson, R. J., Phillips, D. C., Towne, L., & Feuer, M. J. (2003). On the Science of Education Design Studies. *Educational Researcher*, *32*(1), 25–28.
Sherin, M. G., & Elizabeth, V. E. (2005). Using Video to Support Teachers' Ability to Notice Classroom Interactions. *Journal of Technology and Teacher Education*, *13*, 475–491.
Shi, Y., Xie, W., Xu, G., Shi, R., Chen, E., Mao, Y., & Liu, F. (2003). The smart classroom: Merging technologies for seamless tele-education. *IEEE Pervasive Computing*, *2*(2), 47–55.
Sorcinelli, M. D. (2007). Faculty development: The challenge going forward. *Peer Review*, *9*(4), 4–8.
Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., … Young, A. M.

(2018). Anatomy of STEM teaching in North American universities. *Science*, *359*(6383), 1468–1470.
Stes, A., Min-Leliveld, M., Gijbels, D., & Van Petegem, P. (2010). The impact of instructional development in higher education: The state-of-the-art of the research. *Educational Research Review*, *5*(1), 25–49.
Swift, J. N., & Gooding, C. T. (1983). Interaction of wait time feedback and questioning instruction on middle school science teaching. *Journal of Research in Science Teaching*, *20*(8), 721–730.
Tillema, H. . (2000). Belief change towards self-directed learning in student teachers: Immersion in practice or reflection on action. *Teaching and Teacher Education*, *16*, 575–591.
Tobin, K. (1987). The Role of Wait Time in Higher Cognitive Level Learning. *Review of Educational Research*, *57*(1), 69–95.
Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783–805.
Tschannen-Moran, M., & Hoy, A. W. (2007). The differential antecedents of self-efficacy beliefs of novice and experienced teachers. *Teaching and Teacher Education*, *23*(6), 944–956.
van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (Eds.). (2006). *Educational Design research*. London and New York: Routledge.
Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning Analytics Dashboard Applications. *American Behavioral Scientist*, *57*(10), 1500–1509.
Visscher-Voerman, I., & Gustafson, K. L. (2004). Paradigms in the theory and practice of education and training design. *Educational Technology Research and Development*, *52*(2), 69–89.
Wademan, M. R. (2005). *Utilizing Development Research to Guide People Capability Maturity Model Adoption Considerations*. Syracuse University.
Wang, F., & Hannafin, M. J. (2005). Design-Based Research and Technology-Enhanced Learning Environments. *Educational Technology Research and Development*, *53*(4), 5–23.
Weaver, R. R., & Qi, J. (2005). Classroom Organization and Participation: College Students' Perceptions. *The Journal of Higher Education*, *76*(5), 570–601.
Wergin, J. F., Mason, E. J., & Munson, P. J. (1976). The practice of faculty development: An experience-derived model. *The Journal of Higher Education*, *47*(3), 289–308.
Wiggins, G., & McTighe, J. (2005). *Understanding by Design, Expanded 2nd Edition*. Pearson.
Wilen, W. W., & Clegg, A. A. (1986). Effective Questions and Questioning : A Research Review, *XIV*(2), 153–161.
Young, S. L., & Bippus, A. M. (2008). Assessment of Graduate Teaching Assistant (GTA) Training: A Case Study of a Training Program and Its Impact on GTAs. *Communication Teacher*, *22*(4), 116–129.
Zhu, X., Barras, C., Lamel, L., & Gauvain, J. L. (2008). Multi-stage speaker diarization for conference and lecture meetings. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4625 LNCS*, 533–542.
Zimmerman, B. J. (1995). Self-efficacy and educational development. In *Self-efficacy in Changing Societies* (pp. 202–231).

Appendix A: Self-Efficacy instrument

The following 10 items represent a single dimension of self-efficacy for teaching. They are drawn directly from the repeatedly tested and validated Teachers' Sense of Efficacy Scale (Duffin et al., 2012; Klassen et al., 2009; Scherer et al., 2016).

Response options are (1) Not at all true, (2) Barely true, (3) Moderately true, and (4) Exactly true.

1. I am convinced that I am able to teach successfully all relevant subject content to even the most difficult students.
2. I know that I can maintain a positive relationship with students, even when tensions arise.
3. When I try really hard, I am able to reach even the most difficult students.
4. I am convinced that, as time goes by, I will continue to become more and more capable of helping to address my students' needs.
5. Even if I am disrupted while teaching, I am confident that I can maintain my composure and continue to teach well.
6. I am confident in my ability to be responsive to my students' needs, even if I am having a bad day.
7. If I try hard enough, I know that I can exert a positive influence on both the personal and academic development of my students.
8. I am convinced that I can develop creative ways to cope with system constraints (such as budget cuts and other administrative problems) and continue to teach well.
9. I know that I can motivate my students to participate in innovative projects.
10. I know that I can carry out innovative projects, even when I am opposed by skeptical colleagues.

Appendix B: Teaching Perspective Instrument

The following 21 items represent two dimension that make up the Approaches to Teaching Inventory (Prosser & Trigwell, 2006; Trigwell, Prosser, & Ginns, 2005). This repeatedly validated instrument measures Information-Transmission/Teacher-Focus (ITTF) and Conceptual-Change/Student-Focus (CCSF). ITTF represents a traditional view of teaching as an act of providing necessary facts to students. CCSF represents a relatively more modern view of students as active participants in co-creating knowledge structures. Research in active (and interactive) learning promotes higher student achievement in terms of grades and retention (Chi & Wylie, 2014; Freeman et al., 2014). I have adapted the text to suit the context of the participants.

Response options are (1) Is only rarely or never true for me, (2) Is sometimes true for me, (3) Is true for me about half the time, (4) Is frequently true for me, and (5) Is always or almost always true for me.

1. In my class/recitation, students should focus their study on what I provide them. (ITTF)
2. It is important that I describe the subjects I teach completely in terms of specific objectives that relate to the formal assessment items. (ITTF)
3. In the classes/recitations that I lead for this course I try to develop a conversation with students about the topics we are studying. (CCSF)
4. It is important to present a lot of facts so that students know what they have to learn for this course. (ITTF)
5. I feel that my class/recitation has to provide an opportunity for students to reveal their changing thoughts on the subject matter. (CCSF)
6. I take time out in class so that the students can discuss, among themselves, the key concepts and ideas that they encounter. (CCSF)
7. For this course I concentrate on covering the information that might be available from key texts and readings. (ITTF)
8. I encourage students to restructure their existing knowledge in terms of the new way of thinking about the subject that they are developing. (CCSF)
9. During the classes/recitations that I lead for this course I try to provoke debate. (CCSF)
10. I structure my teaching for this class/recitation to help students pass the formal assessments. (ITTF)
11. An important objective of my classes/recitations is for students to take useful notes. (ITTF)
12. For this course I provide the students with the information they need to pass the formal assessments. (ITTF)
13. In this course I feel that I should know the answers to any questions that students might put to me about the subject. (ITTF)
14. I make time for students in this class/recitation to discuss their developing understanding of the subject with each other. (CCSF)
15. It is better for students in my class/recitation to generate their own notes, rather than copy mine (diagrams on the board, transparencies, slides, etc.). (CCSF)
16. I use time in class to question students assumptions. (CCSF)
17. My teaching focuses on the good presentation of information to students. (ITTF)
18. My focus in teaching this class/recitation is to help students develop new ways of thinking in this subject. (CCSF)
19. My teaching in this subject focuses on delivering what I know to the students. (ITTF)
20. When teaching this course/recitation I help students question their own assumptions about the subject matter. (CCSF)
21. When teaching this course/recitation, I help students find their own learning resources. (CCSF)
22. I present information to enable students to build up a knowledge base in this subject. (ITTF)

Appendix C: Rubric for assigning interaction types to participants.

	Productive self-doubt	Confirmatory assessment	Shallow dismissal
Reflection			
Look at their initial reactions to seeing evaluations. Pay particular attention to data that show sub-goal performance.	Reflections are elaborative	Reflections are terse. They may say that they agree or disagree, but do not go deep.	Reflection are short and low on information.
When presented with data about their teaching performance, did they make an effort to explain some challenge they had faced in class?	They explain some challenge they faced in the class. They try to explain what they think may have gone wrong. They may indicate that their practices were misaligned with effective teaching practices.	Unlikely to mention a challenge. They may mention a constraint (different because it cannot be solved). Unlikely to elaborate or explain underlying causes, context, or conditions. They may acknowledge accuracy of data, but without any reflection to explain or interpret it.	They may mention a challenge, but more as an excuse rather than as an explanation or insight.
Is there any self-critique?	They engage in (possibility optimistic) self-critique.	Unlikely to talk about mistakes, but may mention unexpected (external) problems.	Usually no self-critique.
Do they acknowledge when their practices did not align with the stated objectives provided by the app?	Likely.	Most likely to say that the app's objectives are not relevant or useful.	Unlikely to mention.
How much reflection is there?	Usually a fair amount of response to reflection prompts.	A small amount, but more than the bare minimum.	Not very much at all.
If they have inaccurate recall of the most recent teaching performance, do they reflect on the discrepancy?	Usually acknowledges the lack of accuracy.	May question the reliability of the app.	Does not usually notice any discrepancy.
Do they notice when they don't do as well as they thought they did?	Yes	Likely to challenge	Not usually
Accuracy of recall			
Are they able to accurately recall their selected strategies?	This TA is likely able to remember their selected strategies.	This TA may be able to remember their selected strategies.	This TA is unlikely to remember what they said they would try.
How much time passes between planning and reflection modules?	Usually only a few days. They stay on top of things. This may help their recall.	This TA can be slow to respond, but usually muddles through.	This TA can take a very long time to respond, and may forget what they had said previously.

Goal-setting			
Are their goals relevant? List their strategies and goals and check for coherence with stated and derived problems.	Their goals will usually have something to do with the problems they have outlined. They are concrete.	Their goals may relate to stated problems, and will usually be concise. They may dismiss suggestions as not relevant to their particular class.	Goals may not connect well to stated problems.
Is the TA thoughtful about the suggestions? Look at the selected strategies. Are they vague or concrete? Do they address real or perceived problems or simply follow along?	TA is likely to select strategies that address a problem that they identified.	TA may dismiss suggestions outright as being irrelevant, or else set goals with rating themselves "Highly confident."	TA may simply accept suggestions without reflection as to relevance. They may be disconnected from what students actually need. For example, addressing the number of questions by "Asking better questions?"
Enactment			
Did their in-class behavior show an attempt to enact their plans? Did they alter their teaching approach? List the goals they set for themselves and check the behavioral data to see if it matches. Read reflections in review modules.	Likely to enact changes.	Actions in class likely to change to meet stated goal.	Likely to exhibit changed behaviors immediately following relevant modules but decaying soon thereafter. Students unlikely to show responsiveness to changes. Check for self-sabotage. For example, do they increase number of questions but decrease wait time? Likely to follow the current suggestions from the system regardless of reflection or relevance.
Is their strategy selection aligned with their actual needs, or is it out of sync? Are reflections connected to the outcomes? Review their reflections, any stated needs, their goals, and their in-class performance.	Likely to show an awareness of their attempts to change, exhibit reflection in review module, and persist in new behaviors. Students may exhibit changes.	Not likely to exhibit deep reflection, but persists in new behaviors. Students may exhibit changes.	Reflection is likely to be shallow, struggling to connect low-level tactics with overall strategies. Behaviors are not likely to persist. Students unlikely to exhibit changes. Look for indications that the TA blames the students. Look for indications that the TA plans to continue to do something that isn't working
Following study			
Does TSES change?	TSES likely to increase at post-test.	TSES likely to stay high.	TSES likely to decrease.

Appendix D: Correlations from Study 4

Correlation matrix for B_group only participants in Study 4

	Opp-count	*ModuleNo*	*TSES-Pre*	*ITTF-Pre*	*CCSF-Pre*	*z-Attendance*	*z-c-ncQratio*	*z-cQ-per-min*	*z-SI-ratio*	*z-ST-TA-ratio*	*z-cQansRatio*
Opp-count	1.00										
ModuleNo	0.26	1.00									
TSES-Pre	-0.29	-0.29	1.00								
ITTF-Pre	0.09	-0.06	0.34	1.00							
CCSF-Pre	-0.25	-0.20	0.62	0.14	1.00						
z-Attendance	-0.22	-0.65	-0.03	-0.21	-0.02	1.00					
z-c-ncQratio	-0.07	0.01	0.07	0.23	0.50	-0.17	1.00				
z-cQ-per-min	-0.38	-0.50	0.28	0.30	0.39	0.17	0.67	1.00			
z-SI-ratio	-0.25	-0.17	0.21	0.41	0.35	-0.38	0.14	0.42	1.00		
z-ST-TA-ratio	0.10	-0.43	0.55	0.31	0.13	0.22	-0.04	0.48	0.13	1.00	
z-cQansRatio	0.26	-0.26	0.31	0.33	0.09	0.33	-0.19	0.05	-0.17	0.65	1.00
PSD-trans	0.07	0.41	-0.11	-0.72	-0.20	-0.10	-0.12	-0.31	-0.45	-0.14	-0.44
CoA-trans	-0.26	0.20	0.23	0.34	-0.02	-0.15	-0.42	-0.50	0.01	-0.31	0.17
ShD-trans	-0.01	-0.67	0.06	0.34	0.17	0.25	0.24	0.48	0.45	0.09	-0.10
Unkown	-0.22	0.12	-0.04	-0.12	0.09	-0.21	0.20	0.28	0.02	0.12	-0.10
TSES-Post	0.15	0.21	0.63	0.07	0.36	-0.17	-0.24	-0.17	0.00	0.49	0.26
ITTF-Post	0.49	0.03	0.19	0.74	-0.18	-0.08	-0.13	-0.14	-0.05	0.32	0.32
CCSF-Post	0.17	-0.14	0.49	0.33	0.29	0.09	-0.27	-0.11	0.06	0.47	0.31
TSES-post-pre	0.51	0.57	-0.70	-0.37	-0.46	-0.12	-0.31	-0.52	-0.26	-0.25	-0.16
ITTF-post-pre	0.59	0.11	-0.18	-0.26	-0.45	0.17	-0.50	-0.60	-0.63	0.04	0.02
CCSF-post-pre	0.33	0.00	0.06	0.23	-0.39	0.10	-0.60	-0.37	-0.18	0.37	0.24
YearNo	0.43	0.36	-0.25	-0.10	0.14	-0.51	0.23	-0.05	0.46	-0.18	-0.38

	PSD-trans	*CoA-trans*	*ShD-trans*	*Unkown*	*TSES-Post*	*ITTF-Post*	*CCSF-Post*	*TSES-post-pre*	*ITTF-post-pre*	*CCSF-post-pre*	*YearNo*
Opp-count											
ModuleNo											
TSES-Pre											
ITTF-Pre											
CCSF-Pre											
z-Attendance											
z-c-ncQratio											
z-cQ-per-min											
z-SI-ratio											
z-ST-TA-ratio											
z-cQansRatio											
PSD-trans	1.00										
CoA-trans	-0.38	1.00									
ShD-trans	-0.55	-0.23	1.00								
Unkown	0.14	-0.23	-0.37	1.00							
TSES-Post	0.33	0.05	-0.41	0.14	1.00						
ITTF-Post	-0.33	0.24	0.14	-0.16	0.30	1.00					
CCSF-Post	-0.07	0.13	-0.10	0.21	0.76	0.61	1.00				
TSES-post-pre	0.45	-0.25	-0.45	0.19	0.11	0.03	0.07	1.00			
ITTF-post-pre	0.49	-0.11	-0.24	-0.07	0.33	0.45	0.44	0.54	1.00		
CCSF-post-pre	0.07	0.14	-0.21	0.15	0.49	0.71	0.77	0.37	0.72	1.00	
YearNo	0.23	-0.41	0.11	-0.22	0.06	-0.15	-0.17	0.38	-0.09	-0.26	1.00

Correlation matrix for all participants who completed beliefs/attitudes surveys in Study 4

	Opp-count	*ModuleNo*	*TSES-Pre*	*ITTF-Pre*	*CCSF-Pre*	*z-Attendance*	*z-c-ncQratio*	*z-cQ-per-min*	*z-SI-ratio*	*z-ST-TA-ratio*	*z-cQansRatio*
Opp-count	1.00										
ModuleNo	0.37	1.00									
TSES-Pre	-0.32	-0.29	1.00								
ITTF-Pre	-0.03	-0.25	0.40	1.00							
CCSF-Pre	-0.38	-0.72	0.55	0.50	1.00						
z-Attendance	0.05	0.36	-0.16	-0.40	-0.51	1.00					
z-c-ncQratio	-0.07	0.10	-0.04	-0.06	0.02	-0.06	1.00				
z-cQ-per-min	-0.27	-0.37	0.10	-0.07	-0.11	0.18	0.59	1.00			
z-SI-ratio	-0.09	-0.10	0.03	0.17	-0.09	-0.07	-0.01	0.48	1.00		
z-ST-TA-ratio	0.00	-0.63	0.37	0.09	0.05	-0.04	-0.01	0.58	0.25	1.00	
z-cQansRatio	0.24	-0.18	0.01	0.13	-0.15	0.04	-0.13	0.05	0.04	0.48	1.00
PSD-trans	0.18	0.46	-0.24	-0.71	-0.43	0.20	-0.01	-0.12	-0.19	-0.15	-0.22
CoA-trans	-0.10	0.30	0.06	0.14	-0.28	0.15	-0.35	-0.27	0.21	-0.24	0.15
ShD-trans	0.01	-0.25	-0.09	-0.03	-0.21	0.32	0.31	0.59	0.44	0.22	-0.06
Unkown	-0.27	-0.83	0.33	0.47	0.81	-0.57	-0.18	-0.37	-0.39	-0.07	-0.15
TSES-Post	-0.07	-0.21	0.66	0.33	0.64	-0.29	-0.33	-0.33	-0.21	0.17	-0.23
ITTF-Post	0.18	-0.36	0.36	0.75	0.45	-0.34	-0.27	-0.32	-0.25	0.17	-0.01
CCSF-Post	-0.17	-0.72	0.49	0.49	0.74	-0.49	-0.26	-0.18	-0.16	0.38	0.09
TSES-post-pre	0.31	0.14	-0.48	-0.11	0.05	-0.13	-0.32	-0.50	-0.28	-0.26	-0.28
ITTF-post-pre	0.31	-0.18	0.02	-0.18	0.02	0.00	-0.32	-0.38	-0.60	0.13	-0.18
CCSF-post-pre	0.21	-0.30	0.05	0.12	-0.13	-0.10	-0.41	-0.14	-0.12	0.50	0.31
YearNo	0.39	0.34	-0.21	0.05	-0.01	-0.30	0.18	-0.17	0.27	-0.28	0.08

	PSD-trans	*CoA-trans*	*ShD-trans*	*Unkown*	*TSES-Post*	*ITTF-Post*	*CCSF-Post*	*TSES-post-pre*	*ITTF-post-pre*	*CCSF-post-pre*	*YearNo*
Opp-count											
ModuleNo											
TSES-Pre											
ITTF-Pre											
CCSF-Pre											
z-Attendance											
z-c-ncQratio											
z-cQ-per-min											
z-SI-ratio											
z-ST-TA-ratio											
z-cQansRatio											
PSD-trans	1.00										
CoA-trans	-0.14	1.00									
ShD-trans	-0.29	-0.07	1.00								
Unkown	-0.38	-0.39	-0.45	1.00							
TSES-Post	-0.01	-0.13	-0.44	0.58	1.00						
ITTF-Post	-0.48	-0.03	-0.11	0.58	0.58	1.00					
CCSF-Post	-0.37	-0.20	-0.24	0.71	0.71	0.73	1.00				
TSES-post-pre	0.28	-0.22	-0.40	0.27	0.34	0.23	0.22	1.00			
ITTF-post-pre	0.20	-0.24	-0.13	0.26	0.44	0.51	0.45	0.49	1.00		
CCSF-post-pre	-0.02	0.04	-0.10	0.07	0.27	0.53	0.58	0.25	0.64	1.00	
YearNo	0.19	-0.25	-0.13	-0.02	-0.17	-0.22	-0.20	0.06	-0.39	-0.28	1.00

Appendix E: Partial-least squares model

Data sample

TA##	TSES-Pre	ITTF-Pre	CCSF-Pre	ST-TA-ratio-slope	CCSF-Post
TA03	3.2	4.1	2.9	-0.103	2.7
TA06	2.3	2.9	2.1	-0.161	1.9
TA08	2.9	3.4	2.5	0.009	3.2
TA13	3.6	3.6	2.9	0.170	2.1
TA14	3	3.8	2.3	-0.055	2.3
TA15	2.9	3.4	2.0	0.129	2.9
TA18	3.4	3.3	2.7	-0.164	2.2
TA19	3.3	2.2	2.5	-0.066	2.4
TA20	3.5	3.9	2.0	0.060	3.0
TA05	3.3	3.5	2.2	0.098	2.1
TA17	3.9	3.5	3.2	0.167	3.6
TA22	2.5	3.0	2.3	-0.062	1.8
TA07	3.3	3.1	3.1	0.173	3.7
TA09	3.4	3.6	3.2	0.208	3.7
TA12	3.3	4.2	3.9	0.122	4.3
TA11	3.8	3.9	3.6	-0.052	3.5
TA01	3.3	3.8	3.3	-0.037	3.7
TA02	3.4	4.2	4.0	-0.161	3.6

R code sample (All TAs)

Code drawn liberally from PLS Path Modeling with R (Sanchez, 2013). Book available from http://www.gastonsanchez.com/PLS_Path_Modeling_with_R.pdf

```
#necessary library
library(plspm)
# rows of the inner model matrix
Beliefs_pre = c(0, 0, 0)
ST_TA_ratio = c(1, 0, 0)
CCSF_post = c(1, 1, 0)
# path matrix created by row binding
change_path = rbind(Beliefs_pre, ST_TA_ratio, CCSF_post)
# add column names (optional)
colnames(change_path) = rownames(change_path)
# let's see it
change_path
# plot the path matrix
innerplot(change_path)
# blocks of indicators (outer model)
change_blocks = list(2:4, 5, 6)
# vector of modes (reflective)
change_modes = c("A", "A", "A")
# run plspm analysis
```

```
change_pls = plspm(mixed_analysis_simple, change_path, change_blocks, modes = change_modes)
#look at the matrix of correlations
change_pls$path_coefs
#look at the models
change_pls$inner_model
# plotting results (inner model)
plot(change_pls)
```

Output

```
> change_pls$path_coefs
```

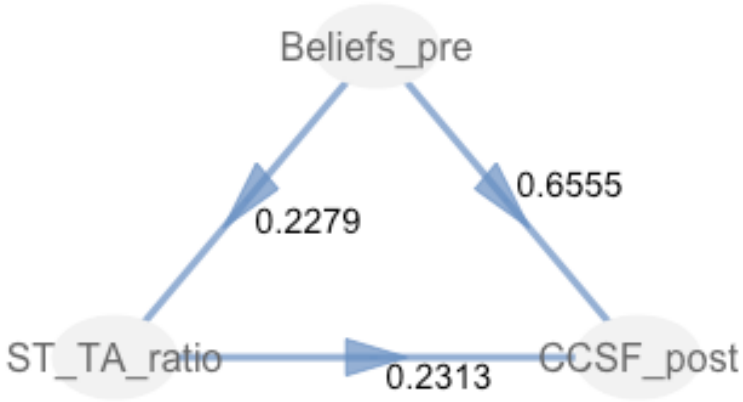
	Beliefs_pre	ST_TA_ratio	CCSF_post
Beliefs_pre	0	0	0
ST_TA_ratio	.228	0	0
CCSF_post	0.6554921	0.2313487	0

```
> change_pls$inner_model
$ST_TA_ratio
```

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	4.11E-17	0.2434221	1.69E-16	1
Beliefs_pre	2.28E-01	0.2434221	9.36E-01	0.3631031

$CCSF_post

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	2.37E-17	0.1727603	1.37E-16	1
Beliefs_pre	6.55E-01	0.1774288	3.69E+00	0.00216379
ST_TA_ratio	2.31E-01	0.1774288	1.30E+00	0.21192498



This seems to show more of the direct impact of prior beliefs on outcomes. The latent variable ## Beliefs_pre now has a significant impact on CCSF_post.

```
> summary(change_pls)
PARTIAL LEAST SQUARES PATH MODELING (PLS-PM)

------------------------------------------------------------
MODEL SPECIFICATION
 1    Number of Cases          18
 2     Latent Variables         3
 3   Manifest Variables         5
 4        Scale of Data  Standardized Data
 5     Non-Metric PLS        FALSE
 6   Weighting Scheme      centroid
 7       Tolerance Crit     0.000001
 8       Max Num Iters          100
 9   Convergence Iters            3
10        Bootstrapping        FALSE
11   Bootstrap samples         NULL

------------------------------------------------------------
BLOCKS DEFINITION
          Block        Type      Size  Mode
1  Beliefs_pre   Exogenous         3     A
2  ST_TA_ratio  Endogenous         1     A
3    CCSF_post  Endogenous         1     A

------------------------------------------------------------
BLOCKS UNIDIMENSIONALITY
             Mode  MVs  C.alpha  DG.rho  eig.1st  eig.2nd
Beliefs_pre     A    3    0.737   0.851     1.97    0.609
ST_TA_ratio     A    1        1       1        1        0
CCSF_post       A    1        1       1        1        0

------------------------------------------------------------
OUTER MODEL
                                  weight  loading  communality  redundancy
Beliefs_pre     1  TSES-Pre        0.475    0.841        0.707           0
                1  ITTF-Pre        0.321    0.724        0.524           0
                1  CCSF-Pre        0.431    0.854         0.73           0
ST_TA_ratio     2  BOTH_z-ST-TA-ratio   1        1            1      0.0519
CCSF_post       3  CCSF-Post       1        1            1      0.5523
```

CROSSLOADINGS

			TSES_pre	ST_TA_ratio	CCSF_post
TSES_pre	1	TSES-Pre	0.841	0.3723	0.492
	1	ITTF-Pre	0.724	0.0936	0.492
	1	CCSF-Pre	0.854	0.049	0.735
ST_TA_ratio	2	BOTH_z-ST-TA-ratio-slope	0.228	1	0.381
CCSF_post	3	CCSF-Post	0.708	0.3807	1

INNER MODEL

$ST_TA_ratio

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	4.11E-17	0.243	1.69E-16	1
Beliefs_pre	2.28E-01	0.243	9.36E-01	0.363

$CCSF_post

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	2.37E-17	0.173	1.37E-16	1
Beliefs_pre	6.55E-01	0.177	3.69E+00	0.00216
ST_TA_ratio	2.31E-01	0.177	1.30E+00	0.21192

CORRELATIONS BETWEEN LVs

	Beliefs_pre	ST_TA_ratio	CCSF_post
Beliefs_pre	1	0.228	0.708
ST_TA_ratio	0.228	1	0.381
CCSF_post	0.708	0.381	1

SUMMARY INNER MODEL

	Type	R2	Block_ Communality	Mean_ Redundancy	AVE
Beliefs_pre	Exogenous	0	0.654	0	0.654
ST_TA_ratio	Endogenous	0.0519	1	0.0519	1
CCSF_post	Endogenous	0.5523	1	0.5523	1

GOODNESS-OF-FIT

[1] 0.4444

TOTAL EFFECTS

	relationships	direct	indirect	total
1	Beliefs_pre -> ST_TA_ratio	0.228	0	0.228
2	Beliefs_pre -> CCSF_post	0.655	0.0527	0.708
3	ST_TA_ratio -> CCSF_post	0.231	0	0.231