SCAFFOLDING REFLECTIVE PRACTICE WITH AN ECOLOGY OF DATA-DRIVEN REFLECTION SUPPORT TOOLS

by

Stephen MacNeil

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2019

Approved by:

Dr. Celine Latulipe

Dr. Mary Lou Maher

Dr. Eun Kyoung Choe

Dr. Nicholas Davis

Dr. Heather Lipford

Prof. Sybil Huskey

©2019
Stephen MacNeil
ALL RIGHTS RESERVED

ABSTRACT

STEPHEN MACNEIL. Scaffolding Reflective Practice with an Ecology of Data-Driven Reflection Support Tools. (Under the direction of DR. CELINE LATULIPE)

Reflection is a process of converting experience into understanding. Through the process of reflection, students actively engage in sense-making around an experience; situating it within their existing experiences, beliefs, and knowledge. Though many theorists have advocated for integrating more reflection into learning experiences, reflection is challenging to implement and evaluate in the classroom. Currently, there is a dearth of reflection support tools. This dissertation introduces an ecology of data-driven reflection support tools that provide scaffolding for reflection in the classroom. By automatically capturing students' behaviors and visualizing them for reflection, these tools help students obtain new insights, increase their agency, and broaden their perspective. Consistently using these tools longitudinally could also help students develop and refine their reflective practice.

Two data-driven reflection support tools, BloomMatrix and IneqDetect, were designed, implemented, and deployed in computer science classrooms to help students reflect on their behaviors and experiences. BloomMatrix crowdsources students' self-reported cognitive states and IneqDetect records and visualizes students' conversations. These tools and students' reflective writing assignments were evaluated using a mixed-methods approach to determine the effect that they had on students' reflections and reflective practices. These in-the-wild studies shed light on the opportunities and challenges presented by reflection and reflection support tools.

ACKNOWLEDGMENTS

A dissertation is the culmination of a long journey. To reflect on my own journey, there have been many pivotal people who have helped me to find my path and sustained me along the way. My interest in science was cultivated by my parents and my high school physics teacher, Marika Foreman. I developed my technical foundation working as an intern with Dr. Vladimir Veselov. I developed a passion for research, design, and visualization from Dr. Niklas Elmqvist. I built on this foundation through the guidance of my good friend, Dr. Lane Harrison. Each of these people was essential to my first steps into the Ph.D. program.

Since, then I have been fortunate enough to work with Dr. Celine Latulipe, Dr. Koji Yatani, and Dr. Mary Lou Maher. Celine introduced me to HCI and taught me to be more clear and concise. She gave me an immense amount of intellectual freedom to explore different areas and has helped me to develop significantly as an instructor. Koji taught me to approach research methodically. Mary Lou challenged me to consider alternative methods of inquiry.

My committee members have also been incredibly supportive throughout the process. Mary Lou helped me to find a unifying theme for my work. Eun Kyoung showed me the forest when I got stuck in the trees. Heather provided the right feedback at the right time. Nick entertained off-the-wall ideas and found in-roads to make them practical. Sybil's optimism and encouragement helped to bring everything together.

Thank you to my many lab mates, co-authors, collaborators, and staff. This includes Johanna, MJ, Vikash, Jinyue, Mike, Sarah, Jin, Tonya, Lina, Syeda, Mahmoud, Mingming, Sauvik, Daisuke, Takuma, Hiroki, Briana, Aileen, Erfan, Devansh, Mohsen, Manuel, Jenny, Tammy, Heather, Carol, Jodi, Sandy, and Dora.

A special thank you to Bruce, Julio, Nadia, Sarah, and Lina who volunteered their classes for this research. Erica and Miriam for helping to apply the coding schemes. My wonderful students, who are now embarking on their own adventures: Kyla Bouldin, Kyle Kiefer, Brian Thompson, Mariah Olsen, Anvesh Mekala, and Dev Takle. Finally, a special thanks to Brian Dorn who has always been a welcoming presence in the research community.

My friends and family have been extremely patient with me throughout this dissertation process. I have missed weddings, baby showers, and many other important life events. Each of you has been endlessly understanding and supportive. Thank you to the Northern Tribe: Yelena, Clara, Jascha, Max, Alex, Eteri, and Lena. Thank you to my many wonderful grand-parents: Lynn, Mitch, Connie, Tom, Dani, Christian, Boris, and Babushki Dina and Luda. Thank you to my good friends Joe, Brendan, Harsh, Dan, Ananth, Charlie, and David. To my parents, Mom, Dad, Laurent, and Marti, who raised me to work hard, but now remind me to take it easy too - thank you. Finally, I'd like to thank my brother, Jeff, who even after years apart knows me better than I know myself.

Most importantly, I would like to thank my wife, Dahlia. You have been the most critical part of this dissertation. Throughout our relationship, you have motivated, inspired, and supported me. Your intelligence and creativity have helped to shape my work and my way of thinking. You are the definition of unconditional love, and I am forever grateful to share our life together.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Learning is an essential part of what it means to be human. The ability to transform experiences into insight is what has helped humans to survive and adapt to a changing world. With the advent of computing, the world is changing faster than ever. As a result, value, which was once associated with facts and information, is now associated with how information generates actionable insights through sense-making. Information alone is no longer enough; employees are expected to adapt to apply knowledge to dynamic problems across many domains. Consequently, learning how to learn has become the most important thing to teach modern students, or more elegantly put, “Education is life itself rather than a mere preparation for life” [119].

Reflection is a tool for lifelong learning that transforms experiences into knowledge. In the context of learning, reflection is “a generic term for those intellectual and affective activities in which individuals engage to explore their experiences in order to lead to new understandings and appreciation. It may take place in isolation or in association with others” [24]. Ideally, reflection provides students with an increased **awareness** and it helps them to challenge their assumptions and biases. The ability to reflect on any experience at any time also provides people with **agency** and the ability to guide their own learning. Despite these benefits, few people are able to successfully sustain their reflective practice over time [44, 122]. According to Christopher Day, the primary reasons for abandoning reflection included a lack of time, a lack of structure, and negative perceptions about the way that reflection was presented. In addition, all reflection is not equal. Shallow reflection seldom results in new insights or transformation for the person doing the reflection. Deep, critical reflection, on the other hand, challenges assumptions and biases. Critical reflection can also lead to behavioral change and new perspectives for the person doing the reflection. Given these challenges, it may be necessary to scaffold reflection with tools to make the process effortless, informative, and sustainable.

Reflection Support Tools (RSTs) are scaffolds for reflection that provide additional information about experiences or help to structure the reflective process. In this dissertation, I use the term RSTs to refer to any tool that supports reflection. I introduce the term **data-driven RSTs** to describe tools that capture and represent data as a means to support reflection. Data-driven RSTs are similar to *Personal Informatics systems*, which help people to collect personally relevant data for the purpose of self-monitoring and reflection. The main difference between data-driven RSTs and Personal Informatics systems is that data-driven RSTs can include more data than just personally relevant data. They can include historical data about how students have behaved in a class, information about the current student cohort, or information that is only personally relevant.

Data-driven RSTs can capture and represent students' behaviors during learning activities as visualizations. Visualizations summarize information and make it easier to obtain insights and discover trends. Collecting the data from sensors or crowdsourcing it from students can also reduce subjectivity during a student's reflection by providing information that is measured externally to that student's senses. These heterogeneous data sources also provide multiple different perspectives of the same learning experience, which is essential for critical reflection. In addition to providing additional data and visualizations for students to reflect on, RSTs also provide much-needed structure by focusing on one aspect of reflection such as cognition, social interactions, or course material. In addition to focusing reflection along a single dimension, RSTs can also provide adaptive prompts that help guide reflection.

In these ways, RSTs can serve to address two of Day's three reasons [44] for abandoning reflection (a lack of time and a lack of structure). Finally, RSTs provide unique benefits compared to traditional modes of reflection. RSTs collect and present information that people may not have attended to during the experience, broadening **awareness**. Additional information, especially when aggregated temporally, can also make it easier for people to track changes and patterns over time, which allows them to form and test hypotheses. This ability to experiment, by changing their behaviors and observing the effect using data-driven RSTs, may increase students' **agency.**

In this dissertation, I introduce an ecology of data-driven Reflection Support Tools (RSTs) to help guide students' reflective practices. There are two data-driven RSTs presented in this dissertation and they focus on how students interact socially with their group members and on the cognitive processes that students engage in during learning. I evaluate these RSTs by observing the effect that the tools have on students as measured through written reflections, group interviews, individual interviews, observations, and surveys. These various heterogeneous data points help to triangulate students' experiences with reflection while accounting for some aspects of subjectivity, priming, and learning effects that are inherent aspects of reflection and learning. I also present a theoretical model for incorporating data-driven RSTs into the reflective process. Finally, a set of challenges and opportunities are outlined to continue the future exploration of data-driven RSTs and reflection.

1.1 Motivation

Over the last two decades, there has been a shift from lecture-based pedagogy to student-centered learning [139]. This shift is rooted in a constructivist view of learning in which students actively construct their own individual understanding of course concepts through experimentation and problem solving [126]. Flipped classrooms and active learning techniques are examples of pedagogy that embody a constructivist perspective. These pedagogical techniques have been widely adopted in STEM programs in part because students are expected to graduate with not only a declarative knowledge of the material but also with practical experience applying the concepts to real problems. In student-centered learning, the instructor designs and maintains an environment in which students take an active role to construct their own knowledge. Reviews of flipped-classroom and active learning report numerous successes [139, 67], but these environments also put more responsibility for constructing knowledge on the student without support or specific instructions for how to construct this knowledge.

An experiential learning perspective posits that learning happens not only during the experience itself, but also during the process of reflecting on the experience [87, 141]. This view highlights the additional importance of reflection as a learning tool. Training students to reflect on their learning can mitigate the aforementioned challenges faced by students in active learning classrooms. Based on a review of active learning in computer science (CS) by Sanders et al., there is a lack of attention to these reflection activities in current active learning research in CS pedagogy [139]. Although there has been an "upward trend" in reflection in engineering education [145], this trend does not appear to have translated to mainstream usage of reflection in engineering classrooms, especially in CS active learning classrooms [139]. In these classes, the emphasis is often on doing rather than reflecting.

Some of the possible reasons that reflection is not used more commonly in CS education have been discussed earlier. Another challenge is the limited ways in which reflection is currently supported. In engineering and CS classrooms, reflection is often scaffolded using prompts for reflective writing [53, 66, 155, 127] or through e-portfolios [2, 75]. Reflective writing can be challenging for students with limited writing backgrounds. Students are expected to develop both their ability to write and to reflect with minimal instruction about how to do either. E-portfolios consist of design artifacts that students have generated through the course and can serve as a focus of reflection. E-portfolios can effectively supporting reflection, but they place more of the emphasis on the outcome rather than on the process. Finally, and most importantly, these scaffolds do not provide students with additional information that might help them reflect more deeply. With additional information, students have increased **agency** to ask and answer questions about their behavior. They also have a broader **awareness** when these tools can capture and present more information than students are typically able to attend to themselves.

To address these issues, I have designed and developed two data-driven RSTs that help students reflect along two aspects of their learning: social interactions and cognition. By scaffolding reflection and providing students with additional information, I expect that these RSTs will increase students' awareness and agency in the classroom.

1.2 Thesis Statement

Reflection support tools can be designed to frame students' reflection along specific learning dimensions, such as their cognition, their metacognition, and how they collaborate with their peers. By capturing and visualizing data about students' learning experiences, data-driven reflection support tools can challenge students to reconsider their subjective, incomplete perception of the learning experience. By reflecting on this additional data, students will become more ***self-aware*** *of how they think and behave during learning activities. This leads to more reflective insights, more* ***agency*** *to experiment with their behaviors, and* ***higher-quality reflection*** *as measured by Fleck and Fitzpatrick's stage-based model of reflection [58].*

1.3 Research Questions

To evaluate this thesis statement, I explore the following research questions:

R1. **Focus:** Does the type of scaffolding provided (social, cognitive, conceptual) frame (or prime) the focus of a student's reflection?

R2. **Reflection Quality:** Does scaffolding reflection with data-driven RSTs lead to higher quality reflection over time compared with only reflective writing?

R3. **Awareness:** How aware are students of their behaviors during learning without RSTs? Do data-driven RSTs improve students' awareness and help them better assess their behaviors when learning?

R4. **Agency:** Do data-driven RSTs increase agency? Agency in this context is measured by intentions to make changes, actual behavioral changes, or instances where students generate and test their own hypotheses.

1.4 Methodology

In this dissertation, I apply a mixed methods approach to triangulate the reflective behaviors of students. The data that is collected from and about students includes written reflections, surveys, individual interviews, group interviews, and classroom observations. Each of these data sources can prime or affect students' reflections. For instance, as students are interviewed about their reflective experiences, they are given a second opportunity to reflect. In this case, it is not always clear whether the insights that they describe in the interview were gleaned during the initial reflection or during the interview itself. By incorporating multiple heterogeneous data sources, such as participant observations and surveys, it might be possible to isolate the insights garnered as a result of using the RSTs.

I have outlined four research questions to identify whether RSTs frame the topic of students' reflections, improve their reflection quality, increase students' awareness, and improve students' agency. To test these hypotheses, I use surveys and pre-intervention reflective writing to establish a baseline. The reflective writing assignments were analyzed using a multidimensional coding scheme presented in Chapter 6. The coding scheme includes the focus of the reflection, a stage-based model for evaluating quality, and a measure of the amount of awareness and agency demonstrated in the writing. These quantitative measures of reflection make it possible to compare changes over time, across tools, and across classrooms. I also use periodic surveys, a summative survey, and interviews to triangulate these quantitative measures. An overview of the study design for the four classes in which the study was conducted longitudinally is shown in Figure 1.

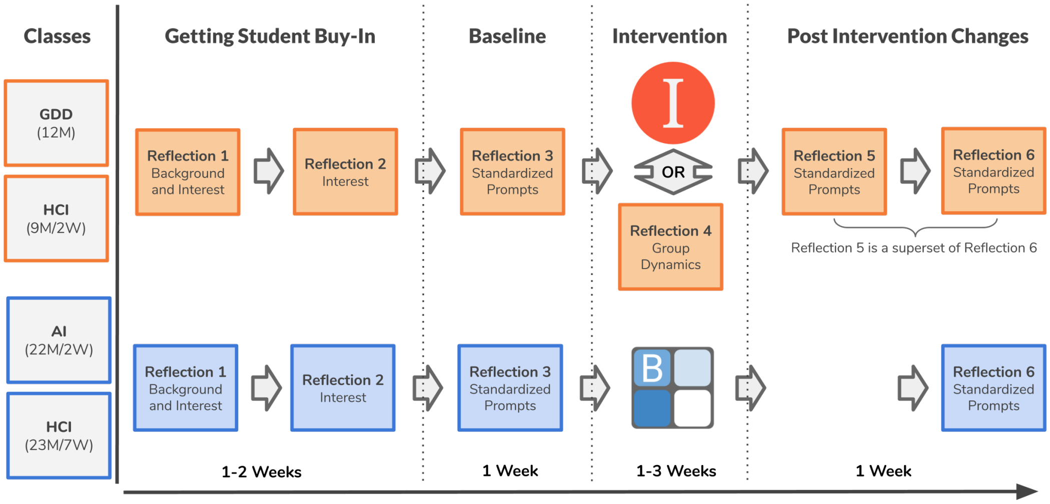
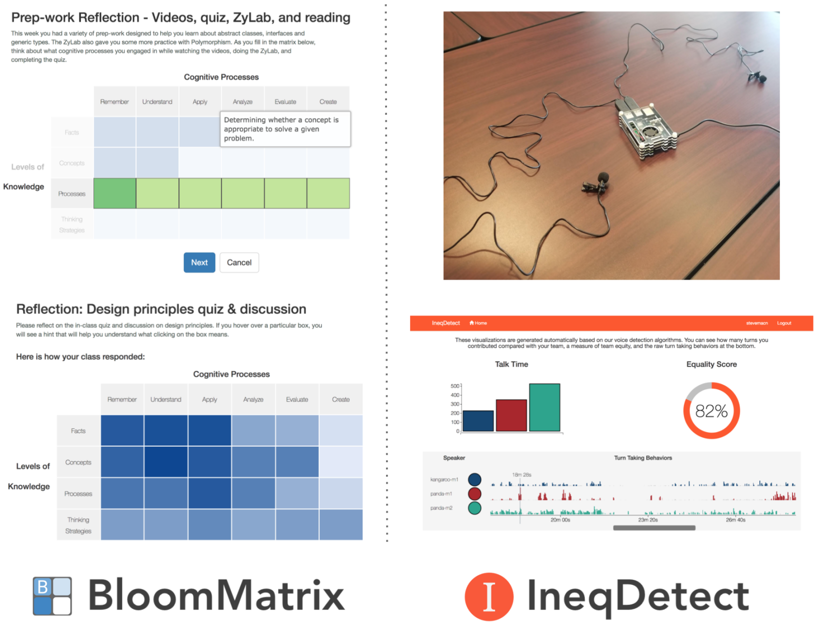

Classes
Getting Student Buy-In
Baseline
Intervention
Post Intervention Changes
GDD (12M)
HCI (9M/2W)
AI (22M/2W)
HCI (23M/7W)
Reflection 1 Background and Interest
Reflection 2 Interest
Reflection 3 Standardized Prompts
I
OR
Reflection 4 Group Dynamics
Reflection 5 Standardized Prompts
Reflection 6 Standardized Prompts
Reflection 5 is a superset of Reflection 6
Reflection 1 Background and Interest
Reflection 2 Interest
Reflection 3 Standardized Prompts
B
Reflection 6 Standardized Prompts
1-2 Weeks
1 Week
1-3 Weeks
1 Week

Figure 1: An overview of the study design and schedule. It shows when students completed reflective writing activities and when they used the RSTs as interventions.

Finally, this research was conducted in multiple classroom environments. To accomplish this goal, each instructor and each class needed to be considered individually and the study needed to be adapted to meet the individual goals of the instructor, class, and students. This makes direct comparisons between classes and interventions challenging. This is further complicated by the fact that classrooms are messy, dynamic, and unpredictable. Conducting in-the-wild studies in these environments introduces numerous confounds which cannot be easily controlled. To this end, I have adapted the design-based research (DBR) methodology [130, 77, 12] to account for these aspects. DBR is a a method of inquiry that attempts to "develop the design of artifacts, technological tools, and curriculum and to further an existing theory or develop new theories in naturalistic settings that can support and lead to a deepened understanding of learning” [83]. The goal of DBR is not to validate theories, but instead it seeks to understand existing theories and generate new theories in authentic educational contexts. DBR embraces the complexity that exists in these contexts rather than attempting to control for it.



Prep-work Reflection - Videos, quiz, ZyLab, and reading

This week you had a variety of prep-work designed to help you learn about abstract classes, interfaces and generic types. The ZyLab also gave you some more practice with Polymorphism. As you fill in the matrix below, think about what cognitive processes you engaged in while watching the videos, doing the ZyLab, and completing the quiz.

Determining whether a concept is appropriate to solve a given problem.

Reflection: Design principles quiz & discussion

Please reflect on the in-class quiz and discussion on design principles. If you hover over a particular box, you will see a hint that will help you understand what clicking on the box means.

Here is how your class responded:

These visualizations are generated automatically based on our voice detection algorithms. You can see how many turns you contributed compared with your team, a measure of team equity, and the raw turn taking behaviors at the bottom.

Figure 2: An overview of the two probes presented in this dissertation, IneqDetect and BloomMatrix. The top images show how data is collected, the bottom images show how it is presented.

1.5 Design Probes

I have created two design probes to explore the research questions posed in this dissertation. Each of these probes supports students as they reflect on different aspects of their educational experiences such as social interactions and cognition. In HCI, a design probe is a way of gathering information about people and phenomena that can inform future design. Gaver et al. define probes as “an approach that values uncertainty, play, exploration, and subjective interpretation” [65]. In this way, the goal of a design probe is not necessarily to solve a problem but instead to shed light on the problem itself. The data-driven reflection support tools (RSTs) presented in this dissertation attempt to scaffold reflection but also serve as an exploration of the nature of reflection in education. They can be seen as design probes to better understand how reflection can be supported and evaluated.

1.5.1 BloomMatrix

BloomMatrix is an interactive web application that allows students to self-report their perception of the cognitive processes that they experienced during learning activities. This information is crowd-sourced from the entire class, aggregated, and visualized as a heatmap. This encourages students to compare their own perceptions with other students' perceptions of the same learning activity. I hypothesize that using the BloomMatrix will frame students' reflections around their cognition and provide alternative perspectives of the learning activity based on their peers. The alternate perspective may challenge students to reflect on the purpose of each activity.

1.5.2 IneqDetect

IneqDetect records students' group conversations that occur during peer learning activities and then visualizes these conversations temporally. In addition to showing the detected speech by group member, IneqDetect also provides summaries that include total talk time per speaker and an overall measure of conversational equity. This work builds on prior research that has identified sociocultural inequities between students during peer learning activities [100, 153]. I hypothesize that presenting data to students about their social interactions will help students to communicate more equitably and lead to better team performance and cohesion. I tested this hypotheses with studies in four classrooms that employ a peer instruction technique [104].

1.5.3 Comparing Probes

BloomMatrix and IneqDetect differ both in the focus of reflection, but also in two important ways. IneqDetect supports group reflection where students view the data together and make sense of their data as a group. BloomMatrix does not support group reflection. Instead, students reflect individually on their own perception and then on the perceptions of others. BloomMatrix is unique because it provides information about how other students in the class reflected on the same learning activity. Finally, these two probes differ in terms of how data is collected. IneqDetect is an example of an RST that automatically captures data for reflection. BloomMatrix, on the other hand, requires students to manually enter data about themselves.

1.6 Contributions

This work contributes to understanding reflection in the context of learning. There is a gap in our theoretical understanding of how multi-dimensional reflection can be supported and evaluated in a classroom environment. There is also a practical gap in how to data-driven reflection support tools into the classroom. This work attempts to bridge these gaps in research and practice by making the following contributions:

- A survey of methods for evaluating reflection.
- A survey of reflection and reflection support tools.
- A model for integrating multiple data-driven RSTs into the classroom.
- The design, implementation, and evaluation of two RSTs in CS classes.

- Insights about the potential framing effects of reflection prompts.
- Design implications for scaffolding and supporting reflection in education.

1.7 Dissertation Overview

This dissertation introduces two design probes which serve as vehicles for exploring whether and how data-driven RSTs can aide students' reflective practices in the classroom. These probes were studied in real classroom environments to understand students' existing reflective practices and to understand the effect that these probes had on students' reflective practices and learning. To account for the complexity of this authentic learning environment in which the probes were deployed, I adopted a mixed methods research approach which draws heavily on DBR.

The goal of this dissertation is to integrate systems that support reflection into existing classrooms. The theoretical foundations for learning and reflection which underpin this goal are presented in Chapters 2 and 3. These theories are explored through the two design probes presented in Chapters 4 and 5. The analysis of the reflective writing assignments is presented in Chapter 6. Finally, insights gleaned from the RSTs and reflective writing assignments are discussed in Chapter 7.

Chapter 2 presents an overview of the research that situates this work. This includes the relevant learning theories from which reflection has been justified and applied. A definition of reflection and related work that explores the many application areas and contexts of use for reflection as well as methods for evaluating reflection. Finally, the design-based research methodology is presented along with a discussion about why non-positivist methods of inquiry are necessary in this context.

Chapter 3 introduces a model for integrating data-driven RSTs into existing reflective practices. This model, which was adapted from existing models of reflection, provides guidance for instructors and researchers that would like to develop and deploy additional RSTs into classrooms in the future.

Chapter 4 describes BloomMatrix, a system which supports students as they reflect on their cognition. Students are able to see how other students in their class reflected on the same activity. The data is crowd-sourced from the entire class and represented as a heatmap visualization. Considering other students' perceptions challenges students to integrate multiple perspectives and reflect on their own subjectivity. This chapter includes a study that evaluates BloomMatrix in three classes.

Chapter 5 describes the second design probe, IneqDetect, which records students conversations and visualizes them to support reflection about group dynamics. After each learning activity, students reflect on their team's conversations during the learning activity. I evaluate IneqDetect across four different classes.

Chapter 6 presents an evaluation of the reflective writing assignments that students completed in the classes that used BloomMatrix or IneqDetect. The reflections were coded and analyzed to better understand students' experiences with reflection and the impacts from using either RST on their reflective practices.

Chapter 7 summarizes the contributions made in this dissertation. I discuss the study results and distill them into a set of design guidelines and lessons learned to help inform the design and implementation of future data-driven RSTs. I also talk about the importance of developing reflective practice in addition to reflective insights, and conceptualize reflection as a new digital literacy.

CHAPTER 2: BACKGROUND AND RELATED WORK

This dissertation contributes insights about reflection in education through the investigation of two reflection support tools. Accomplishing this goal requires an understanding of reflection and the learning theories in which reflection is situated. This chapter presents a review of these educational theories, reflection, and related concepts that underpin the core concepts presented in this dissertation. Subsequent chapters contain brief additional theory sections that are relevant to those chapters only. For instance, Chapter 4 presents BloomMatrix, which scaffolds reflection about cognition and course concepts using the cognitive and knowledge domains of Bloom's taxonomy. Therefore, Chapter 4 contains a brief review of learning taxonomies and theories of cognition. Similarly, in Chapter 5, IneqDetect is a system that helps students reflect on their social interactions, and so that chapter contains an overview of social theory and related work.

2.1 Relevant Educational Theories

To understand how to support reflection in the classroom, it is necessary to understand how it is related to learning. Reflection builds on constructivist theories of pedagogy, including constructivism, social constructivism, and experiential learning. Each of these theories emphasizes the active role that the learner takes during the learning process. Experiential learning additionally highlights the iterative nature of learning and the importance of authentic learning experiences. In addition to these theories, the theories of metacognition and self-regulated learning make a case for the necessity of reflection within learning environments.

2.1.1 Constructivism and Social Constructivism

Constructivism is the idea that “Knowledge is constructed in the mind of the learner” [22]. It rejects the idea that knowledge is transmitted from person to person as an objective chunk of information. Instead, learners reconcile their existing knowledge with new information through the process of accommodation, assimilation, or both [126]. This describes ‘knowing’ as a subjective interpretation of the world as it has been experienced by the learner. This opposes the assumption of traditional behaviorist and cognitivist theories that the world is real, objective, and external to the learner [50].

Piaget’s Constructivist theory is based on the concept of *schema*. A schema is a conceptual structure which stores information as an abstracted representation of the world. Schemata are shaped through the processes of assimilation and accommodation. Assimilation is the process of fitting new experiences and information within our existing schema [126]. Through this process of assimilation, the schema may also be slightly altered or the new information may be subjected to biases which make it fit into an existing schema. Accommodation is the process of updating a schema to account for new information which conflicts with existing schemata [126]. In both of these cases, the individual’s subjective interpretation of new information, as well as the existing schemata available to the learner, affect the way that knowledge is constructed resulting in slightly different understandings across individual students. This view highlights the importance of the learner's previous experiences, existing knowledge, and subjective interpretation of current experiences.

Constructivism has laid the theoretical groundwork for both social constructivism and situated cognition. Social constructivism accounts for the social aspects that are integral to learning. Vygotsky describes 'Zones of Proximal Development' (ZPD) a model that describes how collaboration between a 'more knowledgeable other' and learner scaffolds the learning process [160]. The ZPD model begins with the observation that "Any learning a child encounters in school always has a previous history" [160]. In this statement, Vygotsky highlights existing mental models possessed by students and previous learning scaffolds which have been provided by parents, friends, and relatives. This view echoes the subjective aspects of constructivism and highlights the importance of social aspects, scaffolding, and subjectivity which are endemic to the process of learning. Ultimately, constructivism and social constructivism both highlight the importance of experiences and interpretations of those experiences. Experiential learning considers these experiences and the interpretations of experiences in the form of reflection.

2.1.2 Experiential Learning

Experiential learning relies on the constructivist view that knowledge is constructed by the learner [161]. Experiential learning provides models that help to operationalize the theory of constructivism. Experiential learning stresses the importance of the direct experiences and reflective observations about those experiences [98, 88, 87].

Kolb's 'Experiential Learning Cycle' is one of the most widely used models for experiential learning. According to Kolb's model, which is based on Lewin's canonical model of experiential learning [98], there are four stages which are repeated with each new experience. Kolb's model begins with a 'Concrete Experience' about which subjective or objective data is collected. In the second stage, this data is reflected on by the learner. In the third stage, 'Abstract Conceptualization', the learner makes sense of their reflection in the context of their existing understanding to form hypotheses or conclusions. Finally, the learner experiments with these hypotheses and conclusions in new settings which reset the cycle.

This experiential learning model is sequential and while students can enter at any point in the cycle, Kolb suggests that learning occurs when multiple stages of the cycle are progressed through. This sequential model is typically applied for structuring long-term learning experiences, rather than individual activities, due to the sequential nature of the model. Schön would describe this type of reflection as reflection-on-action [141]. This is contrasted by reflection-in-action where reflection happens throughout the experience. It is likely the case that reflection-on-action is more prevalent in educational settings because it is a more formal style of reflection and typically takes the written form of diaries, blogs, or writing assignments which are externalized from the student and therefore assessable by TAs and instructors.

There are a variety of criticisms about the experiential cycle presented by Lewin and Kolb. Two of these have already been voiced above; reflection does not only happen after the experience has completed and it is not always a structured, sequential process. In addition, Kolb's model does not account for broader social [79, 16] and cultural aspects that are important aspects of critical reflection [62].

Paulo Freire highlights these components in his book ‘Pedagogy of the Oppressed’ by saying that “the pedagogy of the oppressed [is] a pedagogy which must be forged with, not for, the oppressed... this pedagogy makes oppression and its causes objects of reflection by the oppressed, and from that reflection will come their necessary engagement in the struggle for their liberation” [62]. In this way, Freire believes that reflection, participatory design, and collective action are necessary to avoid the oppression which is currently embedded in education and continually reinforced by unexamined pedagogies. This view that reflection can be a tool for emancipation and cultural change is not highlighted in Kolb’s model.

Experiential learning highlights the importance of reflection and authentic learning experiences. It shows the iterative nature of learning and describes how learning and reflection are situated in concrete, authentic learning experiences. Based on the constructivist theory of learning, these process is not passive; learners learn by doing. Through this active and iterative process, learning is the process of transforming experience into knowledge. It is personal to each individual learner and slowly helps to broaden the learner’s awareness and understanding of the world. Therefore, reflection can be understood as a legitimate learning tool which helps students to go beyond the classroom material and to connect what students are learning to the real world.

2.1.3 Metacognition and Self-regulated Learning

Reflection is often lauded for its ability to scaffold the development of metacognitive skills and support self-regulated learning. As learning paradigms shift toward student-centered learning environments, more responsibility is placed on students to prepare for class, develop their own study habits, and collaborate effectively with other students. This places a lot of emphasis on the students' metacognition and their ability to engage in self-regulated learning. Students are expected to become more aware of how they learn, know how they can get help from their peers, and adapt to these new learning environments.

Metacognition has been referred to as "thinking about thinking", "cognition about cognition", and as popularized in the book Metacognition by Janet Metcalfe and Arthur Shimamura, "knowing about knowing" [19]. As initially defined by John Flavell, "Metacognitive experiences are any conscious cognitive or affective experiences that accompany and pertain to any intellectual enterprise" [56]. Flavell gives the example of a person knowing that they are better at arithmetic than at spelling. This idea of knowing oneself is the foundational concept of metacognition. Metcalfe gives an example of knowing which problem-solving strategy to employ in a given situation as another example of metacognition [113]. These skills are essential for lifelong learning. Lifelong learners are expected to decompose problems, identify what new information is needed to solve the problem, and then actively seek that information on their own. Instructors often assume that students learn these skills on their own or that they have already developed them previously, but developing one's own metacognitive abilities is challenging.

Schraw and Moshman describe three origin theories of how metacognition is developed. These three metacognitive theories are Cultural Learning, Individual Construction, and Peer Interaction [142]. Cultural learning describes formal and informal skills that are taught to learners through instruction. Individual construction highlights the "important role of private, reflective analysis of ones own cognition" [142]. Finally, peer interaction is rooted in ideas of social constructivism where students develop metacognitive skills through interaction with peers. Peer interaction allows students to rely on each other's metacognitive skills to solve a given problem. Based on these categorizations of origin theories, metacognition can be developed and supported through instruction, group and self-reflection, and through social interactions.

Self-regulated learning contains metacognition but also integrates "cognitive, behavioral, motivational, and emotional/affective aspects of learning" [123]. Self-regulated learning is a more holistic framework for understanding learning. This holistic focus ensures that students are able to apply the correct learning strategy at the right time, motivate themselves to continue learning, and seek help from peers when necessary. Zimmerman's canonical model of self-regulated learning has three components and integrates an aspect of reflection [168]. Schunk and Zimmerman also describe a social aspect of self-regulated learning with four levels which begin with observation and imitation and move to self-control and finally, self-regulation [144]. These initial models have served as foundations for other models such as the model by Schmitz and Weise [140]. Pandero reviews six models for self-regulated learning and suggests that different models support different populations of learners in different contexts [123]. His review highlights the contextual aspects of self-regulated learning and also the social and reflective aspects which are integral to developing as a self-regulated learner. In these ways, reflection can be an important tool for scaffolding student development.

Like experiential learning, self-regulated learning and metacognition show the importance of reflection as a learning tool. While experiential learning shows how reflection helps students make connections to real-world contexts and make sense of their experiences, self-regulated learning and metacognition show that reflection also helps students manage their learning process itself. Across these different aspects of learning, reflection serves an important purpose for student learning.

2.2 Reflection

As shown, reflection plays an important role in learning and in managing the learning process. Reflection provides students with the ability to transform any experience into knowledge, which provides them with agency in their own learning process. Agency and ownership of one's learning is an essential part of being a life-long learner. Reflection is also supported by multiple prevalent learning theories as a legitimate form of learning. In this dissertation, reflection is used to help students identify important insights and increase their awareness. This is done using reflection support tools (RSTs) which help students to develop their own reflective practice. In addition to scaffolding learning, these tools also act as probes to understand how students reflect on their learning experiences. To this end, it is important to understand how reflection has been defined, to choose an operational definition that can help guide the design and implementation of RSTs, and to choose a method to evaluate reflection in learning. This section addresses these aspects and serves as a theoretical basis for the work presented in this dissertation.

Reflection can be challenging to define because it has been used in so many diverse domains. The emphasis of reflection also changes based on the application area. In health care, the emphasis is often placed on the outcome of reflection, such as how reflection leads to healthier life decisions. In education, the process of reflection and the ability to engage in authentic reflection are often valued over the outcomes of reflection. To understand reflection, it is necessary to understand these many application areas and use cases. For instance, in health care, reflection has been shown to promote healthy behavioral change and improve awareness for self-managing chronic health conditions [58]. The emphasis is placed on self-monitoring to improve health outcomes. In the workplace, reflection encourages employees to consider the perspectives of others which promotes questioning, honest feedback, and participation [28, 159]. Finally, in design, reflection has been used as a way to supplement feedback or make sense of a problem and solution. For example, reflection can help designers to “recall their goals, question their choices, and prioritize revisions” [167]. Reflection can also help designers to reflect on their designs [112, 147] and on the conceptual spaces in which they design [107]. These many diverse examples of reflection show that reflection is not limited to learning and educational contexts. Reflection can also be a tool for behavioral change, relationship building, and sense-making.

2.2.0.1 Reflection in Engineering and Computer Science

Reflection is used in education contexts to support learning, increase student agency in the learning process, and help students go beyond the classroom material. In a systematic review of reflection in engineering education, Sepp et al. identified an upward trend of papers referencing reflection [145]. Despite this upward trend, the use of reflection is still not widespread in engineering education and this is especially true in active learning computer science classrooms [139]. Examples of reflection in computer science classrooms include diaries [53, 66], blogs [155], and reflective discussion forums [127]. Each of these examples uses written reflection exclusively to structure students' reflections. Another common way to structure reflection in CS education is through the use of e-portfolios where students can track their progress over time [18]. In the context of programming, test-driven development has also been described as a form of reflection-in action for students. Students reflect as they iteratively plan, test, and write code [49]. These various techniques help to structure reflection but they do not provide additional information to reflect on. They also do not provide students with motivation to revisit their progress retrospectively.

The *Quantified Self movement* is a recent trend in which people collect personal data about themselves to gain insights. In many cases, technology in the form of *Personal Informatics* is developed to make data collection and representation easier. Recently, the concept of Quantified Self has made its way into the education classrooms [96]. Students can collect data about themselves and their learning to improve performance. This is an example of how technology can go beyond simply structuring reflection but also provide additional data about students' learning, behaviors, or interactions. This development may improve on the lack of reflection that has been identified in CS [139] and in education more broadly.

2.2.1 Personal Informatics and Self-Monitoring

Despite the prevalence of reflection across so many different application areas, it can be quite challenging for people to develop and maintain their reflective practice [44, 122]. As a result, technology is being used to make it easier. Personal informatics and self-monitoring applications are making it easier to collect and represent behavioral information to generate insights about oneself. *Personal informatics* systems allow people to collect information about themselves for self-reflection and to gain self-knowledge [101]. These systems have been used to address diverse problems related to "physical activity, food intake, sleeping behaviors, productivity, mental wellness, menstrual cycles, sleep progressions, and care-giving" [10].

While initially limited to diagnostic devices and clinical settings, advances in wearable computing and an increase in consumer health technologies has lead to an increase in the number and accessibility of self-monitoring systems. Products such as Fitbit [1], Apple Watch [2], and the Withings Smart Scale [3] show how seamlessly integrated these applications have become in our daily lives. In addition, many health-related research probes exist, including Healthii [5] and Sleeptight [33]. Healthii allows users to reflect on their self-reported well-being in the context of others in their community [5]. SleepTight induces positive behavioral change as users reflect on their sleep patterns [33]. These probes rely on the concept of the reactive effect. Reactivity is the idea that the process of recording information about behaviors can alter those behaviors [89]. In most cases, self-monitoring and personal informatics systems allow users to record and analyze their own data themselves.

[1] fitbit.com

[2] apple.com/watch/

[3] withings.com/us/en/scales

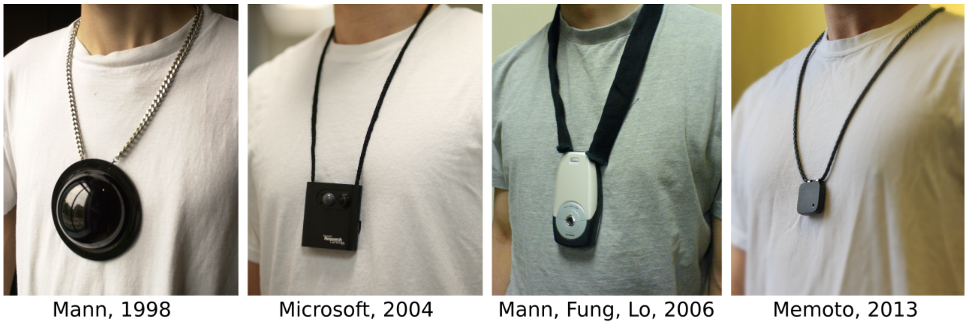
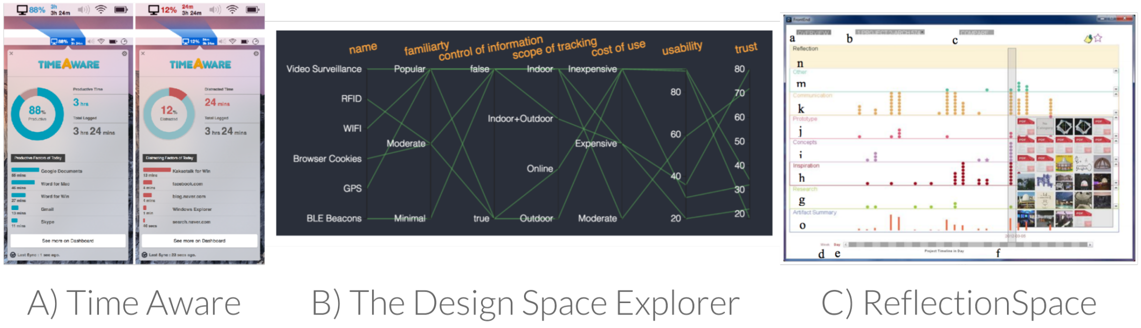


Figure 3: The Microsoft SenseCam is a wearable camera, which was used by Fleck et al. to help teachers reflect on teaching [57]. Images by Steve Mann [108].

2.2.2 Reflection Support Tools (RSTs)

Reflection support tools (RSTs) is a generic term, which encompasses personal informatics, but can also include scaffolding to guide the reflection process. Like personal informatics, some RSTs capture information about an experience and provide external representations, or *visuals* [54]. However, unlike personal informatics, RSTs can also go beyond self-monitoring to include data about other people and experiences. RSTs also includes systems that structure the reflective process without data. For example, the Design Space Explorer helps students to reflect on their research in a structured way so that users can identify gaps in their research area [107]. The Design Space Explorer provides a structured way of representing knowledge and guides researchers through the process of considering their research as a dimensional space. Another example is digital mind mapping software, which provides a structured way to "reflect on one's thought processes" [52]. These examples show how RSTs are not limited to self-monitoring but can also focus on structuring and scaffolding the reflective process.

In education, reflection and RSTs have been used to support both students and



A) Time Aware B) The Design Space Explorer C) ReflectionSpace

Figure 4: Three RSTs from the domains of health, research, and design. A) SleepTight helps users reflect on their sleeping habits [33], B) the Design Space Explorer helps users to reflect dimensionally on their research [107], C) ReflectionSpace helps users to reflect on their design projects [147].

teachers. The SenseCam system, shown in Figure 3, allows instructors and tutors to reflect on their practice through digital photographs [57]. These photographs help to ground group-reflection in a shared context. In the Co-located Collaborative Writing (CCW) system, students use a tabletop application to collaborate, but the data generated from this collaboration is also able to be leveraged by the instructor to support the instructor's reflective practice [76]. Collaid is a tabletop application that helps students monitor their physical and verbal interactions during collaboration [110]. The Subtle Stone is "a tangible technology designed to support students' active emotional communication in the classroom" [11]. The Subtle Stone allows students to reflect on their affective states during the class period based on a color that is emitted from the 'stone'. This allows students to reflect on their affective state and allows instructors to track the students' affective states during the class. The student activity monitor (SAM) was created to help support both students and instructors by visualizing learners' actions [71]. SAM serves a similar purpose to the self-tracking tools outlined earlier. Just like in those domains, self-monitoring and self-tracking can be used in education to increase student's self-awareness which can improve self-reflection [134]. Through self-monitoring and self-awareness, students are also able to develop their metacognition and self-regulated learning skills [142, 168]. Though many other examples in education exist, those described above begin to show how reflection can be scaffolded for both students and instructors along a variety of aspects such as communication, affect, and course content.

2.2.3 The Difficulty in Defining Reflection

Given the widespread use of reflection, many different definitions exist and the way that reflection is used varies widely in the literature. An early definition of reflection from Dewey specifies four necessary criteria: 1) it is a meaning-making process, 2) that involves systematic, rigorous, disciplined thinking, 3) within a community, and 4) it requires attitudes that value the personal and intellectual growth of oneself and of others [47, 29]. Said another way, reflection can be seen as a voluntary tool used to transform raw experience into meaning within a society [29]. Reflective thought is differentiated from thinking because it focuses on extrapolating insight from past experience rather than considering things that are already well known. Mezirow describes thinking as both habitual action and understanding, but reflection serves to critique our assumptions and determine whether our beliefs remain functional [114, 115]. Another way of thinking about this difference is that thinking happens within existing frames, but as Fleck describes, reflection is a process of reframing the situation to understand an experience [58].

The definitions presented to this point help to differentiate reflection from normal thought. They attempt to describe what it is but not necessarily how it happens. Schön describes reflection as the act of considering an experience either retrospectively or in real-time as it is happening. Schön refers to these two types of reflection as on-action and in-action respectively [141]. This helps to understand *how* reflection can occur in practice. Schön and Argyris's *Single and Double Loop Learning Model* describes two ways in which learning occurs [8, 7]. The single loop represents active experimentation based on assumptions that are derived from the experimenter's existing mental model. Double loop learning is the process of using insights from this experimentation to update the experimenter's assumptions and mental model.

This process-oriented view describes reflection as a problem-solving and sense-making process. Like problem solving [39], the reflection can oscillate between periods of problem solving and problem framing. Unlike problem solving, the problem itself may not be defined a priori and the person that is reflecting may instead be trying to make sense of an experience in the context of prior experiences. This sense-making process is similar to Piaget's concepts of accommodation and assimilation, in which learners construct knowledge by integrating new insights into existing schemata.

Finally, reflection can also be considered in terms of the outcomes in which it results. Some of the ways that outcomes can be defined are as new knowledge that has been generated, new ways to frame experiences, updated mental models, or behavioral change in the person doing the reflecting. Insight generation and behavioral change are often cited as the goals of reflection for RSTs [84, 107, 147]. For instance, positive reactivity is often cited as a reason for self-monitoring and self-reflection in the context of personal informatics [102, 32]. Reactivity is process by which recording a behavior results in that behavior changing [89, 40, 129]. In these cases, reflection serves as a catalyst for improvement in the person doing the reflecting. Reflection in this context includes improving sleeping practices (i.e. sleep hygiene) [33], reducing depression symptoms [82], or leading to cultural revolutions through transformative action [62].

2.2.4 A Working Definition of Reflection

This dissertation will revolve around supporting reflection and so defining reflection is an important first step. Baumer's review of reflection in personal informatics systems shows that few papers clearly define reflection or what is meant by reflection in their context [15]. To this end, I have shown how reflection can be defined in terms of what it is, how it happens, and the outcomes in which it results. The review has highlighted multiple aspects of reflection and multiple ways that reflection can be defined. To build the RST tools presented in this dissertation I offer the following operationalized definition of reflection:

> *Reflection is a form of sense-making in which a person or people transform experiences into new knowledge either during the experience, after the experience has ended, or both.*

This definition attempts to adopt most of the aspects described in the previous literature presented above. It accounts for what reflection is, when it can happen, and what it can result in. It does not account for aspects of behavioral and cultural change which are often associated with reflection because the RSTs designed, implemented, and studied in this dissertation do not attempt to address these aspects.

2.2.5 Evaluating Reflection

Reflection is challenging to evaluate because, as shown in previous sections, it is subjective to each individual learner. When students are aware that their reflections are being evaluated, it changes the way that they engage in reflection. For instance, when students are forced to engage in reflection, especially for assessment, they may reflect inauthentically or they may employ strategies to obtain a better grade, eschewing the actual purpose of reflection. Prior work has demonstrated some of these problems associated with evaluating student reflections for a grade [78, 125]. These problems are related to the concept of demand characteristics, where participants form their own interpretation of the study and behave accordingly [120, 121]. Considering demand characteristics in the context of evaluating reflection, students may read the prompt and try to understand the underlying goals that the instructor had in mind for the reflection activity and write a response that meets those goals.

Despite these challenges, there are a number of ways that reflection has been evaluated previously in education and other domains such as health-care and design. Typically these methods belong in one of four categories: 1) outcomes of reflection, 2) quality of reflection, 3) type of reflection, and 4) time-on-reflection. Outcomes of reflection include behavioral change, ability to remember information better, or transformation in the person doing the reflecting. The quality of reflection is typically evaluated using a stage-based model of reflection which presents an ordinal classification of reflection quality. Finally, time-on-reflection is commonly used to evaluate reflection (e.g.: Gnome Surfer [146]), but concerns have been voiced about quantifying reflection using time spent reflecting or the total quantity of reflection [14].

Many of these evaluation methods use written reflections or transcriptions of interviews as a way to evaluate reflection. This can be problematic because it is unclear whether the reported reflection occurs during the writing and interview stage or whether it occurred as a result of the intended cause of reflection. It can be hard to identify the specific cause of reflection. Written and transcribed reflections are often evaluated using a stage-based model. Stage-based models of reflection order aspects of reflection in terms of complexity or temporally. King and Kitchener's model considers not only reflective thinking, but categorizes the types of thinking that build to reflective thought [85]. Fleck's Five Stage Model describes multiple levels of reflective thinking and includes critical reflection as the highest level [58]. Finally, Jenkin's Five I's of Organizational Learning model considers the process by which reflection can occur and considers the actions associated at each level of reflection [80].

On the other hand, Time-on-reflection does not face this problem because it only measures the intended cause of reflection; however, it has been argued that time is not a reliable indicator of reflection [14]. Time-on-reflection is considered unreliable in part because it is not clear that reflection is occurring during the time that the reflection is being measured. It is possible that the person being evaluated is doing something else during the time they are assumed to be reflecting. Time-on-reflection also does not provide any information about the quality of the reflection that is occurring.

There have been many attempts to evaluate nebulous aspects of learning such as critical thinking, design thinking, and computational thinking. Like reflection, each of these thought processes can be challenging to define and consequently difficult to evaluate. Appropriating techniques for evaluation from these other domains is another possible approach for evaluating reflection. Bourner explored the possibility of adapting criteria for evaluating critical thinking to evaluate reflection [25]. In this work, Bourner adapts questions designed to assess critical thinking to assess reflective thinking instead. Questions include "What happened that most surprised you?", "What patterns can you recognise in your experience?", or "What happened that contradicted your prior beliefs? What happened that confirmed your prior beliefs?". Reflection can be assessed when there is evidence that these questions are being explored. This approach does not distinguish the quality of reflection, like stage-based models, but it does provide a broad lens through which reflection can be assessed.

2.3 Choosing a Method of Inquiry

A gap exists between research and practice in many fields such as design and education. Traditionally, knowledge was generated through research in the context of controlled randomized studies and practitioners would apply this knowledge to address problems in their domain. This results in a disconnect between the theory, the implementation of theory as practice, and the outcomes of practice which can inform theory. By removing all known confounds and creating a sterile environment in which to conduct studies, it is unclear that the knowledge which is measured in-vitro is actually what happens in-vivo. In some domains, such as education, the context in which theories are applied can be so highly varied that it is not clear the extent to which knowledge can even be generalized.

Table 1: An overview of the differences between traditional research and design according to the *Encyclopedia of Human-Computer Interaction* [154]

	Research	**Design**
Purpose	general knowledge	specific solution
Result	abstracted	situated
Orientation	long-term	short-term
Outcome	theory	realization

In addition to these disconnects between traditional empirical research and practice, there are also cases where it is impractical to do randomized controlled studies due to logistic or ethical concerns. For instance, in an educational context, it would be unethical to only provide selected students with an intervention that researchers believe provides those students with an advantage. Doing so would disadvantage the students who are not selected to receive the intervention. The goal of educators is to improve educational outcomes for all of their students and promote learning. Randomized controlled studies may conflict with this goal. For instance, a learning activity may take longer than expected and consequently there is not enough time to administer a post-test to assess its impact. One solution is for the instructor to cut the activity short to make time for the research, but this happens at the expense of their students' learning. Based on these challenges there has been some interest in new methods of inquiry which are capable of generating knowledge in these contexts where traditional methods of inquiry are less practical.

The rise of the research-practitioner begins to address this gap by closing the loop between the researcher and the practitioner. It re-introduces the confounds of real-world research environments but also attempts to contribute knowledge that can move research forward. An initial step in this direction was Schön's 'reflective practitioner' [141]. Schön reconsiders the traditional view of experts as objective and removed from the client's problem context. Applied to science, scientists have historically been valued for their objectivity and their ability to create controlled, repeatable studies that validate theories within a context. But this also removes them from the real context in which their theories are eventually applied. In many cases, especially in the natural sciences, this is beneficial because scientists can begin to understand the mechanisms that govern individual phenomena. As these theories are applied in more complex environments interactions between phenomena and emergent properties can limit the applicability of those theories. In these ways, research and practice are complementary. Traditional research is needed to understand individual phenomena and research-practice is needed to understand interactions and emergent phenomena. An overview of these two views of research is presented in Table 1.

2.3.1 Design-based Research (DBR)

In education, addressing the gap between research and practice has taken the form of design-based research (DBR). Multiple scholars, such as Labaree, Kuhn, and Cronback, have described the challenges associated with conducting empirical controlled experiments to study education [130]. For example, based on a review of computer-based instruction, Reeves shows that many studies published did not meet the criteria to qualify as science [131]. These critiques do not eschew the need for empirical studies; instead, traditional research methods need to be supplemented by new ways of generating knowledge that are better suited to deal with the messy, confounded, real-world contexts in which education occurs.

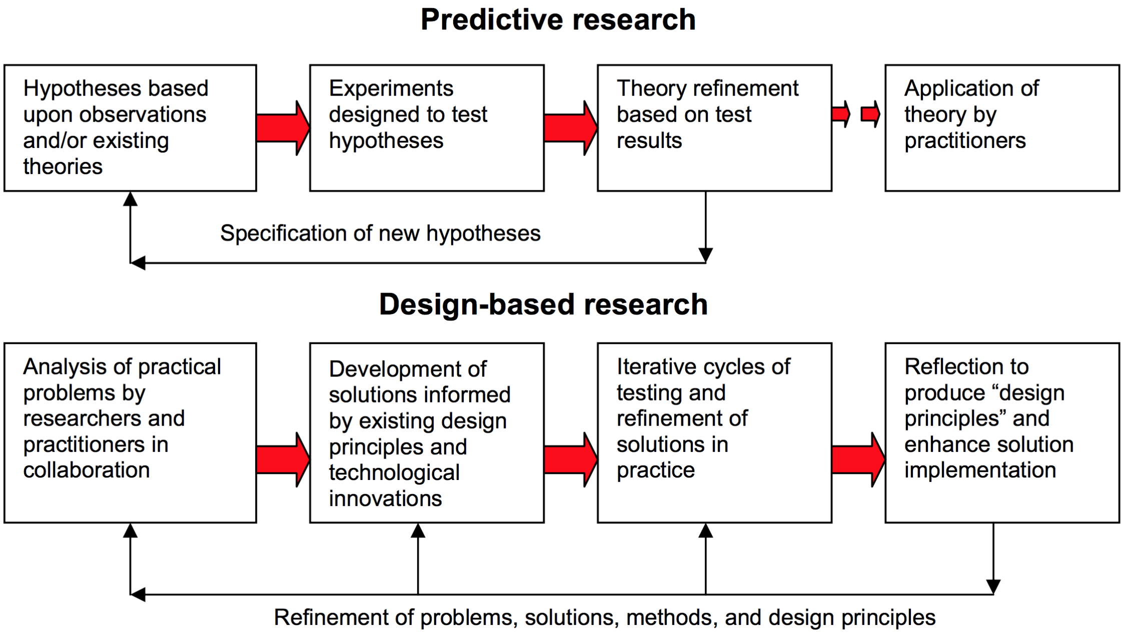

Predictive research
Hypotheses based upon observations and/or existing theories
Experiments designed to test hypotheses
Theory refinement based on test results
Application of theory by practitioners
Specification of new hypotheses
Design-based research
Analysis of practical problems by researchers and practitioners in collaboration
Development of solutions informed by existing design principles and technological innovations
Iterative cycles of testing and refinement of solutions in practice
Reflection to produce "design principles" and enhance solution implementation
Refinement of problems, solutions, methods, and design principles

Figure 5: Differences between empirical research and design-based research as presented by Amiel and Reeves [3]

Based on these critiques about the current state of education research, some researchers have argued that DBR may serve the role of generating knowledge which is not necessarily intended to generalize to every context [130, 77, 12]. For example, Reeves claims that in DBR "the researcher focuses on trying to understand, interpret, and portray the human experience and discourse that occurs in educational settings. In this way, the goal of appreciating complexity is given precedence over the goal of achieving generality" [130]. Each of these methods are subject to their own critiques and for this reason, the best picture of what is actually happening in classrooms requires both. One way of thinking about this is that DBR can be used to generate hypotheses and connect theory to practice, while empirical control studies can be used to validate these hypotheses. A comparison of design-based research and traditional empirical research is shown in Figure 5.

Table 2: “Guidelines for Carrying out Design Research”, from Collins et al. [38].

Implementing a design	
	Identify the critical elements of the design and how they interact
	Characterize how each was addressed in the implementation
Modifying a design	
	If elements of a design are not working, modify the design
	Each modification starts a new phase
	Characterize the critical elements for each phase
	Describe the reasons for making the modifications
Multiple ways of analyzing the design	
	Cognitive
	Resources
	Interpersonal
	Group or classroom
	School or institution
Measuring dependent variables	
	Climate variables (e.g., engagement, cooperation, and risk taking)
	Learning variables (e.g., dispositions, metacognitive, and learning strategies)
	System variables (e.g., ease of adoption, sustainability, spread)
Measuring independent variables	
	Setting
	Nature of learners
	Technical support
	Financial support
	Professional development
	Implementation path
Reporting on design research	
	Goals and elements of the design
	Settings where implemented
	Description of each phase
	Outcomes found
	Lessons learned
	Multimedia documentation

There are a number of things that are reported in randomized control studies. Although there are many ways to conduct these types of studies, the following is a common approach. The study setting should be described along with the control and treatment conditions. The recruitment method is outlined which often consists of inclusion and exclusion criteria as well the sampling strategy. The factors that could be controlled are explained and the factors that could not be controlled are listed in a limitations section. Finally, based on the data collected during the experiment a statistical test of significance is performed to determine whether there was any statistical differences between the conditions.

The procedure for reporting on DBR studies can also vary widely. Collins et al.'s model for reporting on DBR studies [38], shown in Table 2, is an early but often cited model for guiding the report of DBR studies. DBR studies use *phases* to compare differences between classroom environments, student populations, or intervention designs. Each change initiates a new phase. Insights may occur at each phase but these insights may or may not generalize across phases. Challenges that are faced in each phase are also important to report. Finally, because of differences that exist across phases, it is important to report each setting accurately and with details.

2.3.2 Research through Design (RtD)

Within the field of HCI, Daalsgaard [43] argues that a similar "*practice turn* in the study of science" [151] is beginning to occur. Like education, HCI also experiences challenges integrating research and practice. First, technology, communities, and society are constantly evolving and changing. Furthermore, each of these aspects affect the others as they change. Secondly, the presence of *Wicked Problems* [133] in HCI and design, which are challenging to model due to multiple conflicting criteria, limits the applicability of traditional methods of inquiry [43, 169]. These problems, which entail aspects of design, require similar design research methods to address them. Finally, traditional metrics of evaluation, such as accuracy, time to complete a task, and user satisfaction, only capture a small aspect of what HCI may hope to measure and improve. As HCI continues to evolve, the scope expands and what is being measured becomes more complex. These challenges each point to an important aspect of design which is missing from traditional methods of inquiry. Where HCI once positioned design as enabling the study of phenomena, design in the context of HCI is also capable of generating knowledge in its own right.

To address these problems and empower designers to contribute to HCI research Zimmerman, Forlizzi, and Everson [169] formalized a new method of inquiry in HCI which is rooted in Frayling's concept of *Research through Design* (RtD) [61]. Building on Frayling's ideas, the resulting RtD method of inquiry for HCI broadens the types of research that can be conducted in the field of HCI. While working out the details of operationalizing the RtD framework is still an active endeavor [43], it provides a lens through which knowledge can be generated.

RtD provides a new way of thinking about knowledge generation. It is a flexible method of inquiry which help to address many new problems faced by HCI which traditional methods of inquiry were not capable of addressing. Despite its flexibility and wide applicability, there are still many challenges to face in adopting RtD. There is not a clear operationalized definition, it is challenging to distinguish which parts

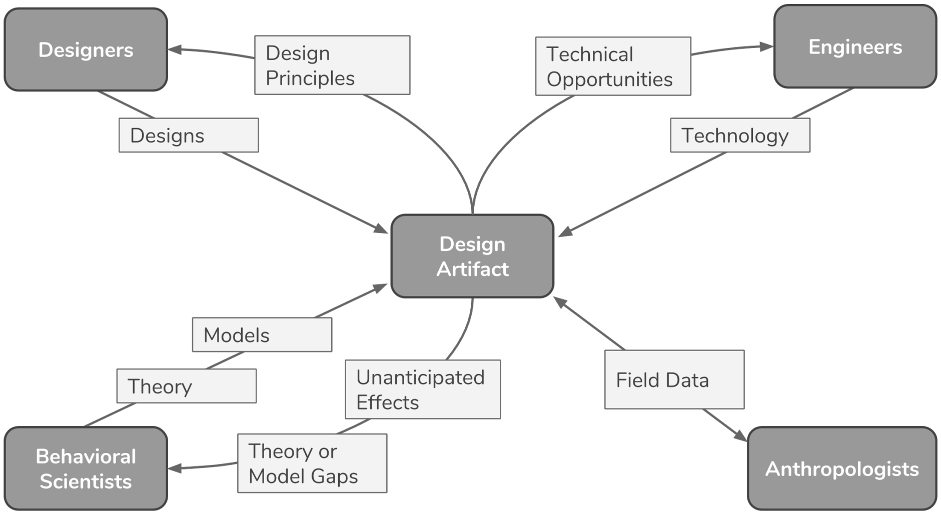


Figure 6: An adapted version of Zimmerman et al.'s RtD model [169]. It shows how aspects of my research can be situated within broader epistemological traditions.

apply to design and research, and there is still debate about the nature of the knowledge generated or how to evaluate it [43]. These frameworks are still relatively new when compared with DBR and they are still evolving; however, they provide a useful alternative to positivist research.

2.3.3 Adopting DBR as a Method of Inquiry

DBR and RtD each provide valuable lens through which multidisciplinary research can be conducted and understood. They both attempt to bridge the gap between research and practice by providing a conduit through which practice can inform research. I chose to use an adapted version of DBR to guide the way that I structure and report on my studies.

The studies were conducted iteratively with changes made to the RSTs and study as needed. To use DBR in the this research: 1) I partnered with educational practitioners, 2) integrated learning theory into the design and evaluation of the interventions,

3) developed, integrated, and revised RSTs and the studies to fit in each unique learning context, and 4) partially report on the DBR using Collins et al.'s guidelines [38]. DBR is traditionally carried out by research teams due to the amount of work involved. To adapt DBR for this dissertation, where the bulk of the work was done by one individual, results are only partially reported according to DBR guidelines.

Though I chose to use DBR, both methods of inquiry inspired my work. For example, I employ a mixed-methods approach by collecting multiple kinds of data and combining quantitative and qualitative analyses techniques. This helps to triangulate the observations made in these classes and to provide a clearer picture of the classroom and the students' experiences. I use a combination of interviews, observations, reflections, and surveys to understand students' experiences. Finally, the RtD framework, shown in Figure 6, has helped me to consider how multiple disciplines affect my work and how my results can inform those fields.

CHAPTER 3: A MODEL FOR SCAFFOLDING MULTIDIMENSIONAL REFLECTIVE LEARNING WITH DATA-DRIVEN RSTS

3.1 Motivation

Developing and maintaining a reflective practice is challenging for most people. Knowing what to reflect on and when to reflect on it is the first challenge. The second challenge is related to cognitive biases and assumptions that people apply when reflecting on their experiences. These cognitive limitations affect the way that we interpret what we see. Finally, our attention is limited and our memories are imperfect reconstructions of real events which can change over time. These perceptual limitations make it difficult to see interesting patterns and trends over time. These practical, cognitive, perceptual challenges limit our ability to reflect successfully.

Supporting reflection with technology begins to address many of these challenges. For instance data-driven RSTs can provide an empirical, falsifiable, static representation of an experience. When people reflect on empirical representations of an experience that conflict with their own subjective interpretation, it may result in cognitive dissonance, challenging them to consider their biases and accept or ignore them. Unlike our memory, representations based on data are static and do not change over time, making it easier to observe temporal trends. This allows students to reflect more deeply and asked more complex questions about their own learning. Data-driven RSTs can also capture many different aspects of an experience through the use of multiple sensors and sensor types. This augments our attentional and perceptual limitations. Finally, data-driven RSTs can structure reflection by focusing on a specific reflective frame, such as cognition or collaboration. In these many ways, data-driven RSTs may improve the reflection process.

This dissertation presents an ecology of data-driven RSTs. Each RST frames reflection around a specific aspect of students' learning, such as collaboration and cognition. By providing multiple reflective frames, students can consider their learning more holistically. They can consider how different aspects of their learning are related.

In this chapter, I review existing models for reflection to consider how data-driven RSTs might be fit into the reflection process. I present a design space based on these existing models for reflection. I consider how the unique aspects of data-driven RSTs might lead to new forms of reflection, such as by providing empirical, static representations along multiple dimensions of learning. Based on this review, I make adaptations to the DEAL model to account for reflection along multiple dimensions using multiple different reflective scaffolds, such as data-driven RSTs. This adaptation results in the first model of reflection which explains how multiple data-driven RSTs might be integrated into the reflection process. The resulting model, which **F**rames **D**imensional **R**eflection, is named the *New DEAL model.*

3.2 Existing Models for Reflection

As mentioned in previous sections, there is still some concern that the term *reflection* is not well defined or used consistently [60]. The ambiguity of the definition, the many diverse contexts in which reflection is used, and the many different research

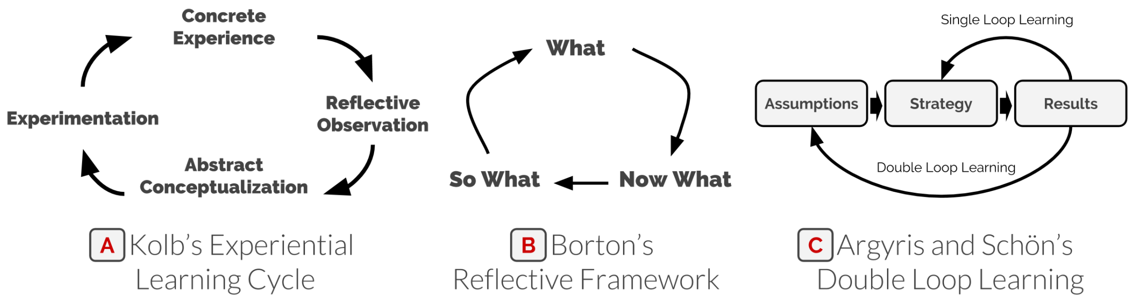


Figure 7: Three examples of reflective learning models. A) Kolb's experiential learning cycle [88], B) Borton's three stage model for reflection [23], C) Argyris and Schön's double loop learning cycle [7]

communities that are interested in reflection has resulted in many different models for reflection. When reviewing these models, I noticed that none of the models considered data-driven reflection support tools and only one model focused on integrating multiple reflective frames. Often, reflection was described as a step in a process, without a clear description of how reflection could be carried out or supported.

To integrate data-driven RSTs into the reflective process, I searched for and compared existing reflective learning models in the research literature. Many reflective learning models exist for different purposes. For example, Kolb's Experiential Learning Cycle features reflection as a distinct step in the process of experiential learning [88, 98]. Kolb's model is helpful for understanding how reflection can be integrated into a course, but it is not prescriptive about how to scaffold the reflection activity itself. This is contrasted by Borton's model, 'What, So What, Now What', which describes a more prescriptive process for conducting reflection [23]. Borton's model is more specific about how reflection occurs, but it does not relate reflection to a specific pedagogical technique or learning theory. Both are necessary and each provides its own emphasis on how reflection can happen or when it should happen.

My search resulted in seven models that are often associated with reflection. And as shown by Kolb's and Borton's models, some reflection models include reflection as a component and others describe the reflective process itself. The seven models include Kolb's Experiential Learning Cycle [88, 98], Argyris and Schön's Double Loop Learning [8, 7], Ash & Clayton's DEAL model [9], Gibb's Reflective Cycle [68], Borton's 'What, So What, Now What' model [23, 135], Boud et al.'s model of reflection [24], Krogstie et al.'s CSRL model [91, 92], and Crossan et al.'s 4I framework [41].

3.2.1 Design Space of Reflection Models

To understand the differences and similarities between these models I have created a design space of reflective learning models, shown in Table 3. This design space contains four main dimensions along which I saw important differences and similarities that helped me to understand how appropriate each model might be for supporting different types of reflection. For example, I saw that some models were better suited for 'in-action' reflection and others for 'on-action' reflection. Considering both in-action and on-action reflection is important because this determines which types of scaffolding (e.g.: RSTs) can be employed. To support reflection-in-action, any visual representations should be simple so that it does not distract from the current task. When supporting reflection-on-action, more elaborate and interactive visualizations can be used because people have more time to explore the data.

Another consideration was the depth of reflection. Some models support narrow reflection resulting in task-level insights. In the case of reflection-in-action, these insights are contextualized by the current context and task. Other models support broader reflection and challenge the person reflecting to consider alternate explanations. Understanding these differences was essential for choosing how and when to integrate RSTs into these existing models.

The four dimensions in the design space are the number of cycles, the scope of the reflection, when the reflection occurs, and the focus of the reflection. Each of the dimensions are outlined below with a few examples to illustrate what they mean:

Cycles: All of the reviewed models were cyclical or iterative, but the number of cycles varied from model to model. For instance, Argyris and Schön's Double Loop Learning model [8, 7] contains two cycles. The first cycle represents active experimentation which is based on a current mental model and the second is the process of updating the mental model. These cycles could be seen as reflection-in-action and critical reflection-on-action, respectively. This is contrasted by Borton's model, which contains only a single cycle [23, 135]. This model focuses on what happened, why it matters, and what actions can be taken in the future as a result.

Scope: Some models encourage broad reflection that goes beyond the immediate context and experience. For example, the DEAL model [9], the CSRL model [91, 92], and the model by Crossan et al. [41] each consider broader contexts such as communities or organizations. These models extrapolate beyond the immediate context and are therefore considered broad. While it is possible to reflect on an experience in a broader context regardless of the model, these models explicitly support broad reflection. All other models are considered narrow by default.

When: Depending on how complex the model was, some were more appropriate for reflection-in-action and others were more appropriate for reflection-on-action. For instance, Ash and Clayton's DEAL model calls for describing an experience and then reflecting on a described experience along three different categories: 'Civic Engagement, Personal Growth, and Academic Enhancement' [9]. This process takes time and is not possible to do in-action. On the other hand, Borton's 'What, so what, now what?' model only has three steps and can be done as the experience, task, or action is in progress. All of the reflection models which supported reflection-in-action were labeled as supporting both because reflection-in-action can be easily adapted to reflection-on-action by extending the duration of reflection beyond the immediate task. There were no models that explicitly supported reflection-in-action only.

Purpose: Some models had a unique purpose for reflection. For instance, both Gibb's Reflective Cycle and Boud et al.'s model each integrated affect into the reflective process [24, 68]. These models cue users to consider how they feel about an experience in addition to what happened objectively. Others, like Crossan et al., have multiple cycles with each cycle encompassing a broader context [41], moving the focus of reflection from individual to group to organizations. Ash and Clayton's model focuses on creating a description of an experience and then considering that description and the experience along a diverse set of dimensions, such as personal growth or academic learning [9]. Krogstie et al.'s CSRL model focuses on revisiting the frame of the reflection to update the "objective, participants, approach and resources" that were considered within the current reflection session [91, 92].

These dimensions can be used to design reflective experiences, build technology to support reflection, or to understand the reflection process. I have used these dimensions to inform the design of the RSTs that I present in this dissertation.

Table 3: A Design Space of Reflective Learning Models

Model	**Cycles**	**Scope**	**When**	**Purpose**
Argyris & Schön	multiple	both	both	transfer
Ash & Clayton	single	broad	post hoc (on-action)	diverse
Boud et al.	multiple	narrow	post hoc (on-action)	affect
Crossan et al.	multiple	broad	post hoc (on-action)	context
Gibbs	single	narrow	post hoc	affect
Kolb	single	narrow	both	future action
Borton	single	narrow	both	future action
Krogstie et al.	multiple	broad	both (on-action)	framing reflection

For example, this dissertation presents IneqDetect, a data-driven RST which helps students reflect on their group conversations. Theoretically, students can reflect on this feedback in real-time during the group discussion (reflection-in-action) or after the discussion has ended (reflection-on-action). Choosing between these two modes of reflection informed the design of IneqDetect. First, considering reflection-in-action, receiving feedback during a group conversation might be distracting. Therefore, IneqDetect might support reflection-in-action with a light that turns on and starts blinking when someone is interrupted. This is analogous to ‘call waiting’ and reminds those who interrupted to transition back to the person that was interrupted. When supporting reflection-in-action, system status should be simple and represent only a narrow aspect of the conversation. On the other hand, IneqDetect might support reflection-on-action with more complex summative visualizations that support exploration and may even span multiple conversations.

This example shows the importance of considering the dimensions from this design space and choosing the most appropriate reflection model when designing reflective experiences for students.

3.3 A Model for Multidimensional Reflective Learning

In the previous section, I defined four dimensions along which I was able to differentiate some of the existing models for reflection. Through this process I was able to see which models were best suited to include data-driven RSTs. I was also able to use these insights to inform the design of the RSTs presented in this dissertation.

By including multiple data-driven RSTs, reflection can be scaffolded and guided toward a specific goal or purpose. For instance, IneqDetect is designed to help students reflect on their group conversations. The other tool presented in this dissertation, BloomMatrix, is designed to support reflection about cognition. Including both might result in a more holistic reflective learning experience which encompasses multiple reflective frames. For example, by using both RSTs, students can reflect on their collaboration and their cognition. They can also consider how the interactions between these two aspects affect their learning.

RSTs can be designed to support reflection-in-action, as mentioned earlier, but currently, most examples from education and HCI literature are designed to support summative reflection-on-action. One possible reason for this trend is that RSTs feature data and visualizations to support reflection, and exploring this data is cognitively demanding. Another possible reason is that capturing, processing, and visualizing data can be difficult to do in real-time. Presenting enough data, in a way that allows students to iteratively construct meaning, is necessary to support critical reflection in which students form a hypothesis, gather data to inform the hypothesis, and then re-evaluate the hypothesis while incorporating new perspectives. This process is often more amenable to reflection-on-action when supported by technology.

3.3.1 Adopting and Adapting the DEAL Model

Given these constraints, that multiple RSTs might be used together, that the data can vary by scale and focus, and that reflecting on data can distract from the task at hand, three models models appeared to be most relevant.

First, Crossan's model is relevant because it extrapolates reflection from the individual, to a group, to an organization [41]. This helps to scaffold reflection to a broader context along a single dimension, but it is unclear how multiple dimensions or RSTs could be integrated. Second, the CSRL model focuses on how the reflection frame can be updated throughout the reflection process [91, 92]. Re-framing reflection along multiple dimensions is one way to consider supporting multi-dimensional reflection. Each dimension could be scaffolded by a different RST which focuses on a specific learning aspect. However, it is not clear at what stage in the cycle RSTs might be integrated. Finally, Ash and Clayton's DEAL model [9] encourages students to describe what happened based on their point of view and then reflect on that description along a few different dimensions, 'Personal Growth', 'Academic Learning', and 'Civic Engagement'. This process leads to broad reflections, but it does not force students to challenge their own narrative of the experience. By including multiple data-driven RSTs, students are presented with information that may corroborate or refute their own account of the experience. This challenges students to be critical of their own assumptions and beliefs which could lead to better insights.

I have adapted the DEAL model to include data-driven RSTs which frame reflection

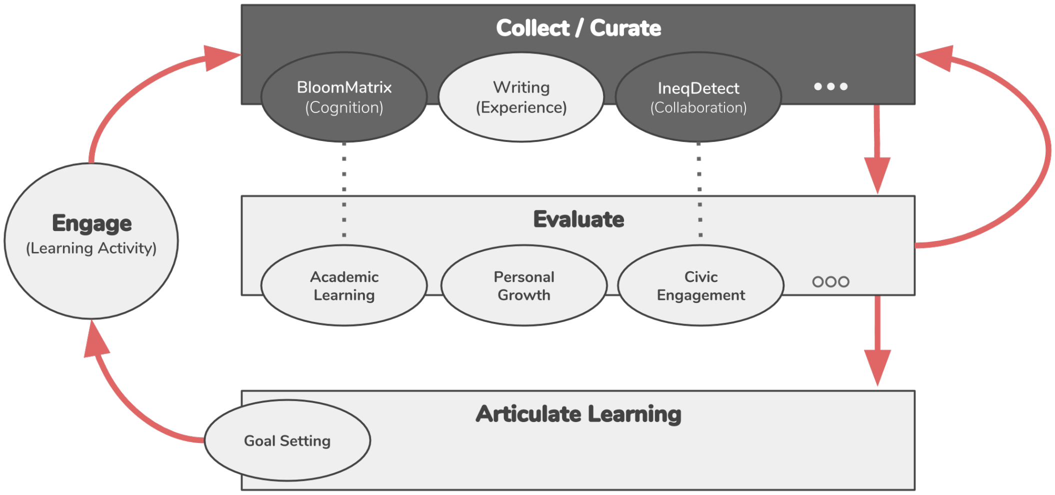


Figure 8: The DEAL model for reflective learning adapted to include personal informatics. Additions are shown in darker gray along with a loop back to curate step. The three dots in the 'Curate' and 'Evaluate' stages indicate that more RSTs or dimensions could be included in the model.

along multiple dimensions of learning. The adaptations to the DEAL model are shown in dark gray in Figure 8. To adapt the DEAL model, a 'Collect/Curate' stage was added to show how data-driven RSTs can be integrated to encourage broader reflection that is framed along multiple aspects of learning. I also added a loop from the 'Evaluate' stage back to the newly created Collect/Curate stage because RSTs can be explored iteratively to obtain evidence to answer questions that arise during the evaluation phase. This cycle can be seen as a sense-making process that exists within the broader cycle of reflection. Like in the CSRL model, this is an opportunity to re-frame the reflection by changing which dimensions and RSTs are considered.

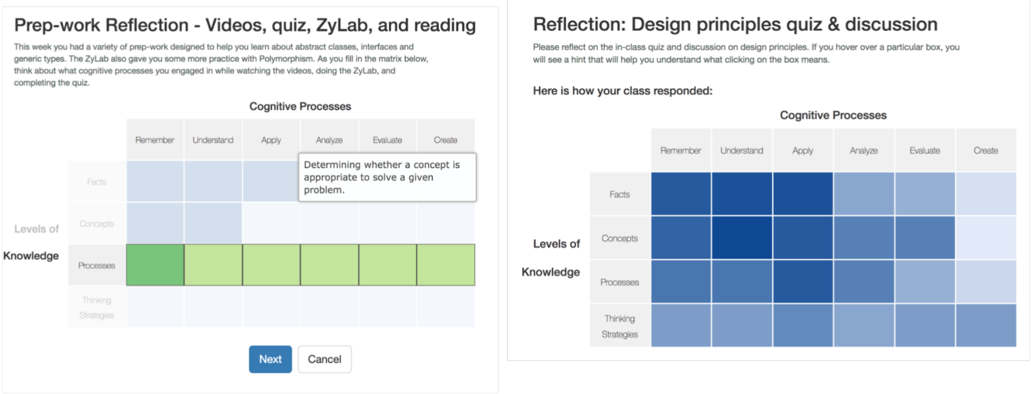
By **F**raming **D**imensional **R**eflection, my *New DEAL Model* integrates data-driven RSTs to support reflection. This model can help to conceptualize how multiple RSTs might be used together to support a holistic reflective practice that considers multiple aspects of learning, such as cognition, collaboration, or personal growth.

This is the first model for reflective learning that integrates multiple ways of scaffolding reflection through the use of RSTs. The intention of this model is that reflection can oscillate between establishing questions across the dimensions and searching for answers by interacting with the RSTs. This sense-making process may lead to unexpected and emergent meaning for the person that is reflecting. It may also challenge them to reflect critically on their experiences. Data-driven RSTs provide evidence that may help to refute or confirm a student's perceptions of a learning experience. In this dissertation, I intend to determine whether and how RSTs can scaffold reflection along a single intended dimension of learning. For future work, I would like to integrate multiple RSTs into the same intervention to see whether and how these help support broader, critical reflections that span multiple dimensions.

CHAPTER 4: BLOOMMATRIX: REFLECTING ON COGNITION

In this chapter, students reflect on the cognitive aspects of learning. BloomMatrix provides opportunities for students to reflect both on their own cognition and on the cognition of the other students in class. After each learning activity, students reflect, individually, on the cognitive processes that they engaged in during a learning activity. They do this by filling in a matrix-based representation of Bloom's Taxonomy [20, 21]. After the individual reflection, students can review a heatmap that aggregates the cognitive processes of the rest of the students in the class. Combining reflection on individual and class levels may prompt students to challenge their own perceptions and biases of learning, leading to cognitive dissonance and critical reflection. The questions that this chapter attempts to answer are whether reflection support tools (RSTs) can support reflection about cognition, how students interpret differences between their own responses and other students' responses, and whether seeing the responses of other students is helpful for reflection.

To investigate these questions, I designed, implemented, and deployed BloomMatrix in two different classes. BloomMatrix is a web-based, mobile-responsive RST that presents students with a two dimensional matrix to support students as they reflect on their cognition and metacognition. The first dimension contains the cognitive processes in which they might engage and the second dimension contains the types of knowledge that they might employ while engaged in a learning activity. These di-



Prep-work Reflection - Videos, quiz, ZyLab, and reading

This week you had a variety of prep-work designed to help you learn about abstract classes, interfaces and generic types. The ZyLab also gave you some more practice with Polymorphism. As you fill in the matrix below, think about what cognitive processes you engaged in while watching the videos, doing the ZyLab, and completing the quiz.

Determining whether a concept is appropriate to solve a given problem.

Reflection: Design principles quiz & discussion

Please reflect on the in-class quiz and discussion on design principles. If you hover over a particular box, you will see a hint that will help you understand what clicking on the box means.

Here is how your class responded:

Selecting Boxes in the Matrix

Reviewing the Heatmap

Figure 9: Students fill out the matrix on the left by clicking on the cells that they perceive were employed during a learning activity. Students are guided by responsive enabling (the green row). This focuses their reflection on the the intersection of each kind of knowing with each set of cognitive processes. After completing the matrix, students reflect on their peers' responses in the aggregated heatmap.

mensions provide a vocabulary for students to use when reflecting on their cognition and metacognition. The system is shown in Figure 9.

4.1 Motivation

Active learning and flipped classroom pedagogies place more responsibility for learning on the student. The resulting learning environments can empower students to develop into life-long learners, but they also make assumptions about students' abilities to self-regulate their own learning. One assumption is that students already possess the metacogntive awareness to consider the learning environment holistically. For example, in a flipped classroom, the in-class activities rely on students having completed prep work at home before coming to class. If students do not connect the importance of prep work to the in-class learning activities they may not buy-in to the flipped classroom learning environment and skip the prep work. Students need to have a holistic understanding of the importance of the many aspects of their learning to be successful in these environments. For this reason, instructors often explain to students the reasoning that led to the design of the learning environment. Justifying these design decisions can improve student buy-in. However, some students may not possess the necessary metacognitive skills to internalize these concepts. These students may continue to ignore the prep work activities and show up to class unprepared. Students coming from lecture-based pedagogies may expect that the prep work materials will be repeated or covered during the lecture. This gives students the impression that the prep work is optional. Students can benefit from guided reflection which helps them to understand how they learn and compare their experiences from in-class and out-of-class learning experiences.

As reviewed in Chapter 2, reflection is a common way to help students develop metacognition and self-regulated learning (SRL) skills. Early SRL models focused on the individual and how they develop SRL skills. More recently socio-cognitive perspectives have been incorporated into SRL [72]. In the same way that Vygotsky's 'Zones of Proximal Development' (ZPD) describes how social interactions with others can scaffold learning [160], co-regulated learning describes how SRL skills can be scaffolded by social interactions. This social scaffolding occurs through "observational learning (modeling, verbal description, social guidance, and feedback) and later by self-imitation and self-regulation" [72]. Based on this research, it is important to not only consider how students reflect on their individual cognition and metacognition, but also help them reflect on other students' perceptions of the same activity.

BloomMatrix is an RST that is designed to support students as they reflect on their cognition and metacognition. This reflective process occurs in two stages. First, students select the cognitive processes and types of knowledge that they perceive they used during the learning activity. Second, students view a visualizations which aggregates and represents how other students in the class reflected on the same activity. The first stage aligns with traditional models for scaffolding SRL skills based on the individual. The second stage adopts a socio-cognitive perspective which allows students to observe how other students reflected on the same activity. By using the BloomMatrix system after multiple learning activities, it might be possible to see how different aspects of the learning environment require the use of different cognitive processes and knowledge. For instance, we expect that BloomMatrix may show different cognitive processes for prep work and in-class activities. For prep work, students may focus on lower-order cognitive processes, such as remembering and understanding. For activities, students may focus on higher-order cognitive processes, such as applying, evaluating and creating. By reflecting on these different aspects of their learning, students may begin to view their learning environment more holistically and develop the necessary skills to succeed in these student-centered environments.

4.1.1 Learning Taxonomies

BloomMatrix uses Bloom's Learning Taxonomy to scaffold students' reflection. A learning taxonomy is a way to represent and classify learning. In practice, they are often used to establish course goals and objectives. Most learning taxonomies provide a hierarchical representation of the students' cognitive processes, the phases of their learning, or the complexity of the knowledge that students use to solve problems. Learning taxonomies are beneficial for instructors and curriculum designers because they serve as a common language that can be used to describe what is happening in a course in a standard way. This is similar to the way that educational design patterns can be used to formalize and share pedagogical techniques [70, 45]. Describing learning in a standard way helps instructors develop common ground when talking about learning. It can also help instructors evaluate and reflect on their course designs more systematically. This can help instructors identify gaps in their course design, where topics are not covered in enough detail to move students beyond remembering the facts and concepts. By applying learning taxonomies, instructors can also more effectively sequence material and cognitive processes so that students are consistently challenged to learn, but not overwhelmed by work that is too challenging. Visual representations of these sequences [103] may also be helpful for students and instructors.

A variety of taxonomies have been introduced to categorize different aspects of knowledge, cognition, learning, and ability. Bloom's Taxonomy [20] initially focused on capturing and categorizing the cognitive aspects of learning. It was later extended to clarify the cognitive aspects and categorize the types of knowledge that students use as they learn [4, 90]. Similarly, 'Webb's Depths of Knowledge' considers complexity and cognitive aspects of learning. The 'SOLO Taxonomy' categorizes only the complexity of a student's understanding. 'Fink's Taxonomy of Significant Learning' focuses on how the student changes as a learner and incorporates high-level aspects such as "[the] human dimension, caring, and learning how to learn" [55]. Wiggins and McTighe present 'Six Facets of Understanding' which categorizes understanding on a spectrum from specific and straightforward to abstract and complex [163]. 'Gardner's

Multiple Intelligences' shows that learning is not limited to single modality and that aspects of learning, such as 'kinesthetic', 'interpersonal', and 'naturalistic', should be considered when designing learning environments for students [64]. Finally, some taxonomies such as 'Niemierko's Taxonomy; [118] and the 'CS Specific Taxonomy' [63] build on other taxonomies, either to incorporate new aspects or contextualize the taxonomy for a specific use case.

This short review of learning taxonomies shows that each taxonomy highlights specific aspects of learning that could be considered when designing learning environments, creating learning outcomes, sequencing course material and skill-building activities, or designing assessments. Taxonomy are often adapted to a specific context or to achieve a specific goal. As an example, the CS Specific Taxonomy extends Bloom's Taxonomy to be more relevant for a CS context. It restructures the original matrix with the new meta-categorizations 'producing' and 'interpreting' [63]. Existing cognitive aspects are mapped within these two meta-categorizations and pathways that CS students might take through the matrix are described to illustrate the utility of the taxonomy for addressing the needs of computer science.

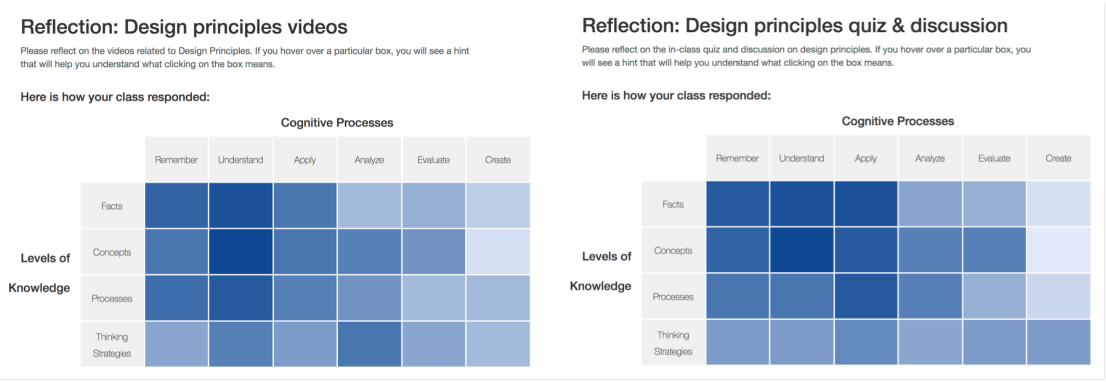
Like the CS Specific Taxonomy, I have chosen to adopt Bloom's Taxonomy as a way to scaffold learning. Although Bloom's Taxonomy is the most widely used learning taxonomy, some researchers have identified challenges with using it for assessing learning in CS [63, 157, 81]. Given that I am using the taxonomy for reflection, rather than assessment, the concerns raised about its effectiveness as a tool for assessment do not apply in this context. For this reason, I chose to use Bloom's Taxonomy rather than the CS Specific Taxonomy which was adapted for assessment. Furthermore, because

Bloom's Taxonomy is less specific, it can be used in classes that are not only focused on CS topics, such as HCI. Finally, Alaoutinen and Smolander have shown that CS students can place their knowledge within Bloom's Taxonomy [1]. This is compelling for using Bloom's matrix as a way to scaffold reflection for non-assessment purposes such as development of metacognitive and SRL skills. Consequently, I chose to use Bloom's Taxonomy due to the widespread use of Bloom's Taxonomy in CS [63, 157, 81, 1], the ability students have to place their knowledge within the taxonomy [1], and the absence of assessment as a goal.

BloomMatrix uses the revised version [4, 90] of Bloom's original taxonomy [20]. In the original taxonomy levels of knowledge and cognitive processes were represented together as categorizations. In the revised version of the taxonomy, these categorizations were converted from nouns to verbs to reflect the idea that cognition is enacted. The levels of knowledge were separated into a new dimension and the resulting taxonomy can be represented as a two dimensional matrix. This representation allows students to reflect on the learning as it pertains to their cognition and metacogntion.

4.2 The BloomMatrix System

BloomMatrix provides students with 6 categorizations of cognitive processes and 4 categorizations of the types of knowledge used. There are 24 cells that represent combinations of cognitive process with knowledge type. After completing a learning activity, students self-report which categorizations best described their learning. The categories of cognitive processes are "remember, understand, apply, evaluate, analyze, create". The categories of the knowledge dimension are "facts, concepts,



Reflection: Design principles videos

Please reflect on the videos related to Design Principles. If you hover over a particular box, you will see a hint that will help you understand what clicking on the box means.

Here is how your class responded:

Reflection: Design principles quiz & discussion

Please reflect on the in-class quiz and discussion on design principles. If you hover over a particular box, you will see a hint that will help you understand what clicking on the box means.

Here is how your class responded:

Prep Work (at home)

Clicker Quiz (in class)

Figure 10: BloomMatrix heatmap visualization. Students reflect on an aggregate representation of how their peers filled out the matrix for the same activity.

processes, and thinking strategies (metacognition)". These terms provide students with a vocabulary for talking and thinking about their cognition, metacognition, and knowledge types. These terms were derived from the revised taxonomy [4, 90]. To use the matrix, students select any number of cells from the matrix by clicking on the cell. They select the cells based on which cognitive processes they think that they engaged in and the types of knowledge they accessed. For example, students might indicate that during a learning activity they 'remembered facts', 'understood concepts' or they 'evaluated a process.' These intersections of cognition and knowledge are represented by each cell. After filling in the matrix, they reflect on a heatmap which aggregates their peers' responses in the matrix. The color in the heatmap is determined by mapping the counts to a linearly interpolated color scale between zero and the maximum count for any cell in the matrix. A comparison of two heatmaps from two different learning activities are shown in Figure 10.

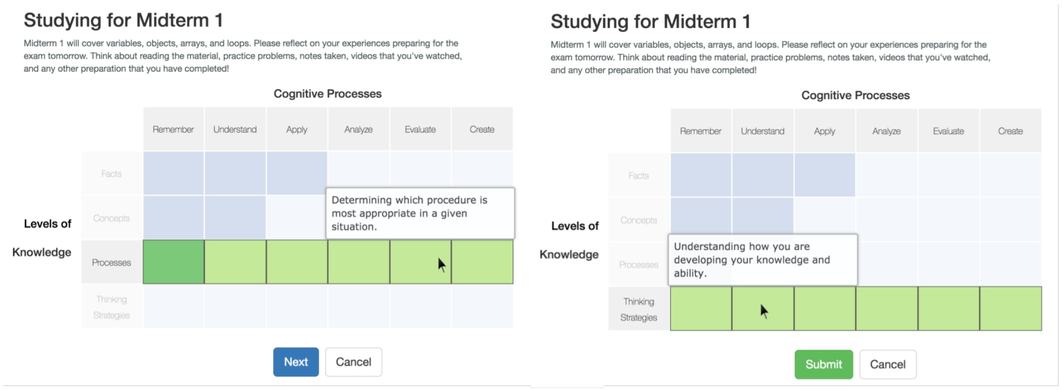
4.2.1 Formative Studies and System Design

The design and development of BloomMatrix went through a series of iterations to test the appropriateness of using Bloom's Taxonomy for reflection and the usability of the BloomMatrix system. This took place over the course of two summers and the work was assisted by two students in the Research Experiences for Undergraduates (REU) summer program. We used a combination of low-fidelity and medium-fidelity prototypes to develop the BloomMatrix probe which was used in these in-the-wild classroom studies.

In the first set of iterations, I worked with an REU student to determine whether Bloom's Taxonomy was appropriate for supporting reflection. We built a few paper prototypes and deployed and tested them in an informal usability study with students in the HCI lab. We observed through these studies that users had trouble with some of the vocabulary used in Bloom's taxonomy. Students were able to fill in the matrix which is similar to the findings by Alaoutinen and Smolander [1]. However, we did observe that students were confused by some of the terms. For instance, many users were unfamiliar with the term 'Metacognition' and they had trouble distinguishing between other terms, such as 'Analyze' and 'Evaluate' or 'Facts' and 'Concepts'. This led us to create hints that showed a definition when the user moused over a cell. Based on these prototypes we created a web application with static pre-defined hints. As a result of this pilot study, we added hints to the prototype to help students understand the vocabulary of the taxonomy.

In the second iteration, which occurred in the following summer, I worked with an REU student to further investigate the problems that we observed in the first study. We explored the use of contextual hints and responsive enabling to address these problems. Contextual hints are descriptive tooltips that pop up when students hover over a cell. Responsive enabling is an interaction technique where features are gradually enabled as they become relevant to the current context. Examples of these two features from the final version of BloomMatrix are shown in Figure 11. We expected that both features would lessen the cognitive load on students as they filled out the matrix and reduce confusion. To explore this possibility, we created a folded paper prototype which revealed each row of the matrix one at a time. In an informal user study, new users appeared to prefer the row by row prototype to the full matrix. Most participants claimed that this aspect helped them to focus on what they were doing. They also described it as being less overwhelming. Additionally, we created a set of contextual hints for each cell in the matrix which were intended for an introductory programming course. These hints were added to the web application to evaluate their usefulness. REU students used the web application to reflect on their professional development workshops. The hints appeared to be helpful even though they were not contextualized for the professional development activity. This suggests that hints are helpful, but that creating contextual hints may not always be necessary.

Following these formative studies, I re-implemented the web application prototype and added administrative features for instructors. These features allow instructors to create and configure custom contextual hints, create BloomMatrix reflection assignments linked to learning activities inside or outside of class, and upload their class roster to send activation email invites to students to join BloomMatrix. These fea-



Studying for Midterm 1

Midterm 1 will cover variables, objects, arrays, and loops. Please reflect on your experiences preparing for the exam tomorrow. Think about reading the material, practice problems, notes taken, videos that you've watched, and any other preparation that you have completed!

Determining which procedure is most appropriate in a given situation.

Studying for Midterm 1

Midterm 1 will cover variables, objects, arrays, and loops. Please reflect on your experiences preparing for the exam tomorrow. Think about reading the material, practice problems, notes taken, videos that you've watched, and any other preparation that you have completed!

Understanding how you are developing your knowledge and ability.

Figure 11: Responsive enabling guides students line by line and contextual hints provide students with additional information about what each cell means.

tures allow instructors to use BloomMatrix in their class without assistance. The final version also includes a set of standard hints in case instructors do not want to add their own contextual hints. After discussing the final version of the research probe with instructors, they requested an ability to check whether students had completed the matrix. This was the last feature added before conducting the studies outlined below. The final version of the platform that students used was presented earlier in this chapter, in Figure 9.

4.3 Research Questions

BloomMatrix is an RST designed to use cognition and metacognition as a conceptual frame for students' reflections. BloomMatrix uses Bloom's Taxonomy to structure students' reflection. Bloom's Taxonomy provides students with a vocabulary to help categorize and evaluate their own cognition. It is my hypothesis that students will use this vocabulary in their reflective writing assignments which will help them to talk about their cognition and metagcognition.

BloomMatrix is unique because it provides students with two opportunities to reflect on their cognition and metacognition. First on their own perceptions and then on their peers' perceptions of the cognitive and metacognitive aspects of a single learning activity. These two phases of reflection first acknowledge the students' own perception, and then potentially challenge this perception if it conflicts with the aggregated responses from the class. These conflicts can lead to cognitive dissonance in the individual and may challenge them to question their own responses, leading to deep critical reflection. Alternatively, students may discount their peers' responses and ignore these discrepancies in the responses.

Finally, if there is a significant amount of variance between students' responses, the heatmaps may be rendered ineffective. It is possible that when crowdsourcing the responses from students there will be too much variance and no obvious consensus about the cognitive processes in which students engaged. This would result in a heatmap with all the boxes filled equally. I expect to see different patterns emerge for different classes and activity types. I also intend to investigate the amount of variance present in students' responses and the sources of these variances.

These questions are summarized below. The first four questions are explored in this chapter. The second three questions are addressed in Chapter 7.

R1. **Capturing Cognition** Does BloomMatrix show that different cognitive processes are employed by students during different types learning activities?

R2. **Variation** How much variation exists between how students fill out the matrix? What causes these variations? Is critical mass necessary?

R3. **Impact** How helpful is it for students to reflect on cognition and metacognition?

R4. **Awareness (Holistic Learning)** Does BloomMatrix cue students to reflect on how their cognition is affected by different types of learning activities.

R5. **Awareness (Existing Practice)** How much do students currently reflect on their cognition, metacognition, or thinking styles?

R6. **Focus** Do students reflect more on their cognition and metacognition after using BloomMatrix? Do they use the vocabulary provided by Bloom's Taxonomy?

R7. **New Practice** How does BloomMatrix affect students' reflection habits? Do students increase their awareness or depth of reflection?

4.4 Hypotheses

Based on the research questions presented in the last section, I present the following hypotheses for the first four research questions:

H1. Heatmaps will be similar for the same type of learning activity, such as watching lecture videos (prep work), regardless of content and will differ otherwise.

H2. The heatmaps will contain less variance for larger classes than for smaller classes. In larger classes, the effect of outliers on the heatmap is less significant.

H3. Students will prefer using BloomMatrix to reflective writing because it is faster to fill out and easier to use.

H4. Students will talk about their learning more holistically. They will have increased awareness of the aspects that affect their learning.

4.5 Studies in the Wild

To understand how students reflect on their cognition and metacognition and to evaluate BloomMatrix as an RST, I conducted a series of studies in existing classrooms. In laboratory studies, the goal is to remove the confounds of the real-world environment to isolate a specific aspect of a participants experience. Applying inferential statistics on the data obtained in these studies can show causal relationships between the intervention and the observed phenomena. In-the-wild studies do not attempt to control for these confounds. They acknowledge that real-world environments are complex and they seek to better understand the many different aspects that affect the intervention and phenomena. As a result, the findings are often presented with detailed descriptions of the settings, problems faced, and solutions. I have been inspired by design-based research (DBR) and employ a mixed methods analysis approach so that I can collect data holistically and preserve descriptions of the existing classroom structure and experience. In education, this is important, because poorly designed interventions that significantly change the way students learn may be detrimental to students learning, their GPA, and consequently their careers. Finally, the study described in this chapter was designed as a within subjects study because a between subjects intervention may have given some students an advantage over other students within the class. I hypothesized that BloomMatrix would be beneficial for students in the class; therefore, it was made accessible to all students.

4.5.1 Setting

All three phases of this study took place at the University of North Carolina at Charlotte. All of the classes in which the study took place are offered by the College of Computing and Informatics (CCI). These classes are predominantly attended by students who have majors in the college. However, these courses can attract students from other disciplines. The Human-Computer Interaction (HCI) class in particular often attracts non-majors and majors alike. Of the classes presented in this paper, two of them took place in specialized active learning classrooms (HCI and CS2). In the HCI and AI classes, the desks were reconfigurable and could be pushed together to allow for group work or traditional rows and columns for lectures and tests. The CS2 class had tables where each group sat together.

4.5.2 Recruitment

The recruitment process for the study started with snowball sampling the instructors from my college. I started by asking instructors who were involved with the CCI Center for Education Innovation (CEI). The CEI is a center of excellence for education in my college which develops and coordinates initiatives that contribute to computer science education (CSEd) research and practice. The instructors that are active in the CEI are often practicing innovative pedagogy in their classrooms. Instructors from the CEI were recruited by email for the study. Some of the instructors that I reached out to, pointed me to other instructors who they thought might have more time or who teach more appropriate classes for reflection. A few instructors that I approached were interested, but declined to participate. They expressed an interest in developing reflective practice in their courses, but were concerned about the amount of effort it would be to integrate and the amount of additional cognitive load it would place on students in the class.

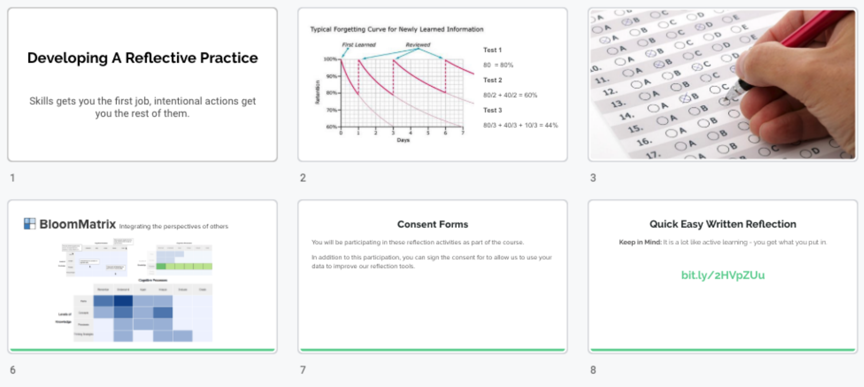
Classes did not have to implement aspects of active learning to be considered for the study, although classes with active learning were preferred because they likely contain multiple different types of learning activities such as prep work, in-class discussions, peer instruction clicker quizzes, or programming labs. The courses that were recruited for this study are summarized in Table 4. One of the instructors that I spoke with was interested in using reflection to help students connect the dots between the course concepts and the many different activities that students participated in for her class.

Table 4: An overview of the courses which were recruited for the study.

Course	**When**	**Class Size**
Introduction to Computer Science II (CS2)	Spring '18	110
Human-computer Interaction (HCI)	Summer '18	30
Introduction to Artificial Intelligence (AI)	Summer '18	26

Instructors were recruited to use either BloomMatrix or IneqDetect. In each meeting with potential instructors, I discussed the study, their class, the things they wanted students to achieve through reflection, and the two RSTs presented in this dissertation. Depending on their needs, I would suggest one or the other. For instance, I have already described the instructor who wanted students to connect the dots between concepts and activities. I suggested that BloomMatrix might be an appropriate tool because students can compare their cognition across multiple different learning activities. This may help them to see how each aspect of the classroom design is intended to target a specific aspect of their cognition.

After coordinating with the instructors, I uploaded the class roster on their behalf. Instructors were able to do this themselves, but in a pilot study, the instructor forgot to upload the roster and missed a planned reflection activity. To avoid this, I uploaded the roster (emails only) to BloomMatrix which automatically emailed each student with personalized log-in credentials. On the first day of class, I gave a presentation



1

Developing A Reflective Practice

Skills gets you the first job, intentional actions get you the rest of them.

2

Typical Forgetting Curve for Newly Learned Information

Test 1

80 = 80%

Test 2

80/2 + 40/2 = 60%

Test 3

80/3 + 40/3 + 10/3 = 44%

3

6

BloomMatrix Integrating the perspectives of others

7

Consent Forms

You will be participating in these reflection activities as part of the course.

In addition to this participation, you can sign the consent for to allow us to use your data to improve our reflection tools.

8

Quick Easy Written Reflection

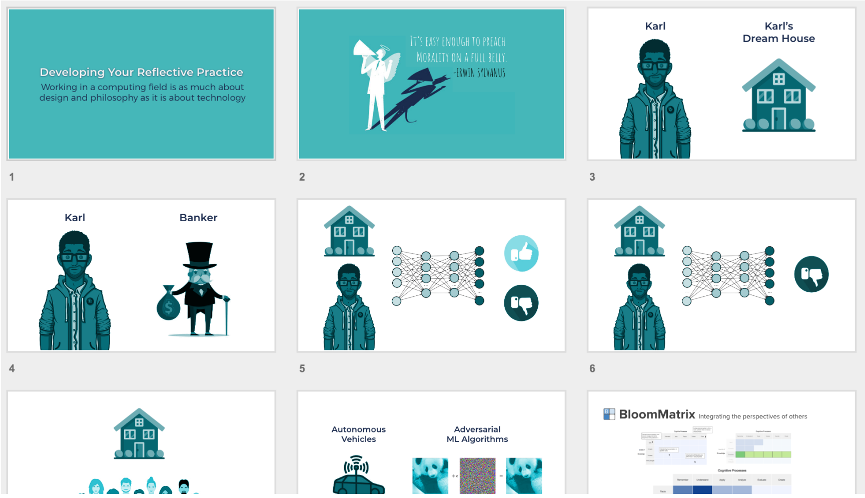
Keep in Mind: It is a lot like active learning - you get what you put in.

bit.ly/2HVpZUu

Figure 12: I presented these slides to introduce reflection and BloomMatrix in both spring classes. Using the forgetting curve, we discussed pros and cons of cramming versus incremental studying. Later in the semester, students were reintroduced to BloomMatrix with a practice reflection.

to students to help promote the idea of reflection. The slides from two of these presentations can be seen in Figures 12 and 13. Initially, I used a generic set of slides that could be presented in any class, shown in Figure 12. These slides asked students to reflect on whether cramming for exams was an effective. In the class where these slides were used, fewer students signed the consent forms. The low participation rate was exacerbated by having to obtain consent from students all at one table. Frequently, one or more students would either refuse the consent form or turn it over when they received it. This would prompt the rest of the group to do the same.

In the HCI and AI classes, I customized the slides and attended at least one class to get to know the students before presenting the RSTs and obtaining consent. This dramatically improved the consent rate, despite all other aspects remaining unchanged. The slides for the AI class are shown in Figure 13. In that class, I explained how algorithmic biases can be carelessly integrated into systems that affected real people. I explained that reflection is extremely important in fields like AI where decisions



Developing Your Reflective Practice

Working in a computing field is as much about design and philosophy as it is about technology

It's easy enough to preach morality on a full belly.
-Erwin Sylvanus

1
2
Karl
Karl's Dream House
3
Karl
Banker
4
5
6
Autonomous Vehicles
Adversarial ML Algorithms
BloomMatrix Integrating the perspectives of others

Figure 13: These slides were customize for the AI class to encourage reflection. I presented the often overlooked ethical aspects of black-box algorithms, and as consent forms were distributed, we discussed "explainable AI."

can impact so many people, even accidentally. Every student who attended this presentation consented to the study. It may have helped that students were not sitting together in groups for that class. These observations say a lot about how important student buy-in can be for fostering successful reflective practices in the classroom.

4.5.3 Procedure and Data Collection

To evaluate BloomMatrix and its ability to support reflection, I have designed the study outlined in Figure 14. The within subjects study featured both BloomMatrix and reflective writing assignments. Students were able to use both of these tools for reflection. On the first day, students were introduced to the study, provided with consent forms, and asked to complete a survey. Then, students engaged in reflective writing assignments to establish a baseline. The number of reflective writing assignments varied based on the needs of the class and the preferences of the instructor. In each class, at least one baseline reflection was about the course topics or course

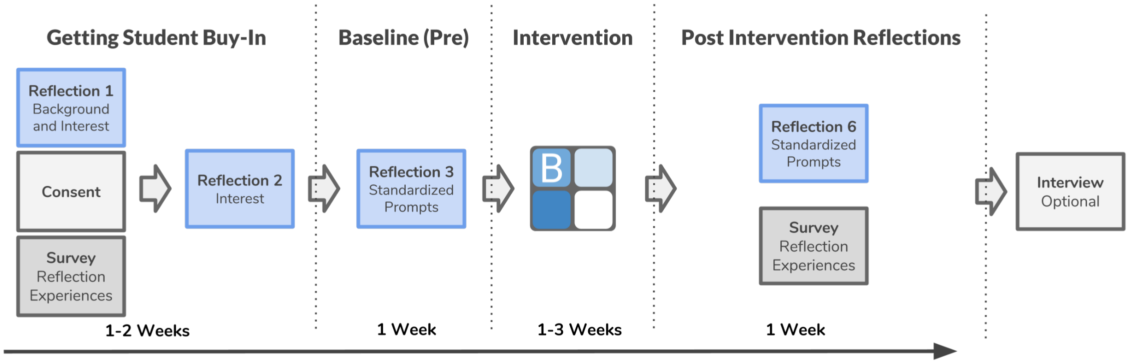

Getting Student Buy-In
Reflection 1 Background and Interest
Consent
Survey Reflection Experiences
Reflection 2 Interest
1-2 Weeks
Baseline (Pre)
Reflection 3 Standardized Prompts
1 Week
Intervention
B
1-3 Weeks
Post Intervention Reflections
Reflection 6 Standardized Prompts
Survey Reflection Experiences
1 Week
Interview Optional

Figure 14: The sequencing of the study components throughout the semester. The specific duration of each component varied by course, due to instructor's preferences for more or less reflection and the demands of the existing curriculum.

structure. These reflective writing assignments provide a baseline which represents how students reflect without using BloomMatrix. These reflective writing activities included prompts that were intended to have students reflect about their cognition, metacognition, or course topics. For the studies in the summer classes, I provide students with feedback on each of their reflective writing assignments. This feedback included encouragements and highlighted aspects of their writing that were interesting to me. After a few weeks, students used BloomMatrix to reflect on their learning activities. Before using BloomMatrix, we went through a practice session as a class. In the practice session, we discussed each column to determine whether it was appropriate given the learning activity we just completed. After using BloomMatrix for a few class sessions, students returned to reflective writing, which serves as comparison to see whether students' exposure to BloomMatrix changed their reflective behaviors. Finally, students completed a survey and have the option to participate in a follow-up interview. Students were offered a $5 gift-card from Starbucks as compensation for participating in the follow-up interview.

To understand student's experiences with reflection and BloomMatrix, I collected reflective writing assignments, survey responses, interview data, and the data that they entered into BloomMatrix. I used the BloomMatrix data to compare across multiple activities to see how these activities affected their cognition. I also looked to see whether there were differences in variance between the two classes and across the different learning activities. The survey data provides information regarding what students liked and did not like about reflection and using BloomMatrix.

4.5.4 Evaluation

To evaluate the research questions and hypotheses presented above, I analyzed the survey responses and the BloomMatrix form data that is represented by heatmaps. Students were repeatedly invited to attend a follow-up interview for a $5 gift-card from Starbucks. However, no student volunteered to be interviewed for the study. These classes are very condensed and some students leave campus immediately after the final exam to enjoy their summer. It is likely that this and the low compensation rate contributed to the lack of interview volunteers. The reflective writing assignments were also analyzed but that data will be presented in Chapter 6.

To understand students' experiences with BloomMatrix and reflection, I collected survey data at the beginning and end of the semester. The survey had a mix of free response questions and statements that students rated on a 5-point Likert scale. Before analysis, I read through the responses and de-identified them. Free-response questions were coded based on *First Cycle* and *Second Cycle* coding methods [138]. In the first cycle, both *In Vivo coding* and *Structural coding* were employed to label the data. In Vivo coding labels the data using words extracted from the responses. The resulting labels are in participant's own words rather than in the words of the qualitative coder. In the second cycle, patterns were distilled from the codes by grouping labels together thematically. This data was used to evaluate the effectiveness of the intervention. Therefore, an external evaluator was recruited to apply the coding scheme. This reduced biases that I might have as an investigator. Once the data was coded, I analyzed and wrote up the analysis based on the coded data. In addition to this qualitative analysis, the Likert responses were aggregated for the students in the class and then graphed to indicate students' preferences.

To better understand the types of insights that BloomMatrix might support, I also collected data from the system itself. This data consists of how each student filled out the matrix for each activity. This data is presented as a heatmap to students during reflection and they can review the heatmaps for previously completed activities at any time. I used this data to compare the heatmaps across classrooms and learning activities. I also computed the variance for each box in the matrix. Because the responses to the matrix are binary (selected or unselected), the data results in a non-normal distribution. The mean for each cell is equivalent to the probability of students choosing that cell. In other words, it is the number of times the cell was selected divided by the number of students who completed the matrix. The variability of the sample proportion, which is analogous to the standard deviation, is given by the formula $\sqrt{\frac{p(1-p)}{n}}$. In the formula, p represents the probability that the cell was selected and n represents the total number of students who filled out the matrix. The standard deviation in this case is a measure of agreement, thus it is lowest when many respondents selected the cell or when few respondents selected the cell. I also considered using inter-rater reliability (IRR) as a measure of student agreement in the heatmap. However, IRR is best suited for few raters with many ratings, rather than many raters with few ratings. I also considered using Principal Component Analysis (PCA) to identify parts of the heatmap with the most variation, but I save that analysis for future work.

4.5.5 Initial Deployment Study

The first phase of this study was conducted in the CS2 class. The instructor for this course had a very clear goal for how she wanted to use reflection in her class. She wanted students to connect the dots between course concepts and activities. She explained that she had trouble getting students to complete the prep work. She believed that many of her students thought they could learn all of the material during the in-class learning activities. She was excited that BloomMatrix might help students to reflect on the purpose of the many different learning activities in her class, including prep work, to view the course more holistically. She said that she put a lot of time into the course design and she wanted students to understand the purpose of the different activities.

Teaching the CS2 course at most universities can be very challenging. The preparedness levels of the students at this level in the curriculum are highly variable. I have previously published a study about these variations in preparedness of the students in our college leading up to the data structures course [95]. These classes are also very large, which makes it difficult to provide remediation for students who are less prepared than their peers. These challenges make it especially difficult to integrate new aspects, such as reflection. Not every student in these classes has experience with active learning. They may not have fully bought into the concept at this point in their academic careers. Trying to introduce another concept, such as reflection, at this point can be challenging.

These many challenges made it difficult to obtain consent from students and to conduct reflection activities reliably. I have previously discussed the slides presented on the first day of the study to get students to buy into the concept of reflection. I have also described the challenges getting students to consent in large classes where students sat together in teams. These challenges led to very sparsely collected data in this class. Only two reflection activities were conducted, both using BloomMatrix. No surveys or reflective writing assignments occurred in the course. The instructor struggled to design reflection activities and reflection hints for the BloomMatrix system. As a result, we had a few meetings to create customized hints for the two different BloomMatrix reflection activities. We also created some reflective writing activities, but these were never used in the class due to time constraints. In one class, the learning activity took longer than expected and there was only five minutes left in the class period to do reflection. The students were exhausted from the learning activity, so the instructor opted to skip the reflection. In another class period, the reflection activity was forgotten altogether.

After the experience with this class, I realized that I would have to take a more active role integrating reflection into the curriculum. It is my experience from recruiting instructors and classes that reflection can be intimidating. In the College of Computing and Informatics, we provide workshops, design patterns, and a variety of other resources for instructors as they adopt active learning pedagogy. It is my feeling that similar support is needed to successfully integrate reflection.

4.5.6 Studies in Summer Classes

In subsequent phases, I worked more closely with instructors, providing them with hints, developing reflective writing prompts and activities, and co-designing the goals that instructors wanted reflection to accomplish in their courses. I attended all of the class periods and I worked with two instructors to design and implement reflective activities in their courses for the first and second summer sessions. In these classes, I helped to create hints that were relevant for each of the two courses. Each class was five weeks long and met multiple times per week. I also helped out during learning activities, giving students some feedback. This helped with student buy-in as they knew that I cared about their learning and not just about my research. It also gave me a chance to understand how reflection might help them with legitimate week-to-week problems that they faced within the class.

The summer studies took place in two different classes. In the first half of the summer, BloomMatrix was used by students in the HCI class. I participated in the first class period, getting to know the students, before presenting the study at the end of class. The instructor also introduced me at the beginning of the class as a design expert and asked me to provide design critiques during class. At the end of the class period, I gave the same presentation slides as in the CS1 class, shown in Figure 12. This time, the response rate was much higher. In the AI class, I did the same, but created a more customized presentation that was more relevant to the course topics. Both classes took place over a period of five weeks, with one week for final exams. It was a very condensed schedule and so all of the reflection activities needed to be created in advance. Some of the reflection activities were given as take home assignments to reserve as much time as possible for class activities. These changes dramatically improved the consent rate and students' enthusiasm about reflection when compared with the initial deployment study.

4.5.7 Deployment in the HCI Class

The Human-Computer Interaction (HCI) class is focused on teaching user-centered design methods. It typically combines many different types of activities which focus on teaching students the concepts, giving them opportunities to practice, and then using those design methods on a course-long project. The summer offerings of this course typically have fewer students than either the fall or spring semesters. This specific course is structured with a lot of group work, but it also combined mini-lectures and class-wide discussions into most class periods. The discussions helped students get to know students outside of their group, and it led to inter-group collaboration, even during group work. The majority of the reflection activities happened during class-time which was preferred by the instructor. The reflection activities also happened immediately after the activity, rather than at the end of class, as was preferred by many of the instructors with whom I spoke. Students were mostly positive about reflection, but did indicate that the reflection was a lot of additional work.

4.5.8 Deployment in the AI Class

The Artificial Intelligence (AI) class combines active learning and lecture-style instruction. The instructor used a significant amount of class time to work through problems and concepts, but he involved students in the discussion and frequently asked the class to participate. Most classes had at least some portion of the class time set aside for active learning activities. These activities included working through problems on paper or programming. Students received additional coding practice at home. Based on my experiences in the HCI class and to adapt to the addition work load, I tried to minimize the amount of reflective activities that students had to complete. We also assigned most of the reflection activities to be completed at the end of class or at home. This ensured that more class time could be reserved for the course material and practice problems. However, having students complete their reflections at home led to a lower response rate for the activities and less data for the study compared to the HCI class.

4.6 Results

I conducted studies in three classes to understand the effect that reflecting on cognition and metacognition had on students and to evaluate BloomMatrix as an RST. These classes were very similar in size, but had some differences, which have been discussed earlier in this chapter. I collected a variety of data to understand students' experiences from multiple perspectives and applied a mixed-methods approach to analyze the data. In the following subsections I describe the results of these analyses and my findings.

4.6.1 Different Cognitive Processes for Different Activities

BloomMatrix crowdsources the cognitive and metacognitive processes that students experience during learning activities from the students themselves. Crowdsourcing leverages scale to reduce the impact of outliers. It relies on the wisdom of crowds [156], which states that aggregated estimates from a large group of people can often be more accurate than a single expert's estimate. I expected the same would be true when aggregating students' perceptions of a learning activity. Given a large enough class, the outliers' experiences would regress to the mean experiences of students in the class, resulting in a more reliable estimate. Therefore, I expected that the resulting heatmap, which aggregates students' responses, would be relatively uniform across students. Furthermore, I expected that it would differ based on the learning activity, with more variation for small classes. Based on these ideas, I hypothesized that the heatmaps would look distinctively different for different types of learning activities, regardless of the course material.

To answer these questions, I extracted the students' responses from BloomMatrix. I used this data to create new heatmaps which aggregated students responses for multiple activities. For example, I constructed a heatmap which represents the responses for multiple prep work assignments. I also combined all the responses for all the activities in a class to show one heatmap that represents every in-class activity from the class. In Figure 15, I show heatmaps that were aggregated by class and activity type. These heatmaps show some interesting trends. Through the rest of this section, keywords from the taxonomy will be italicized to make the writing clearer.

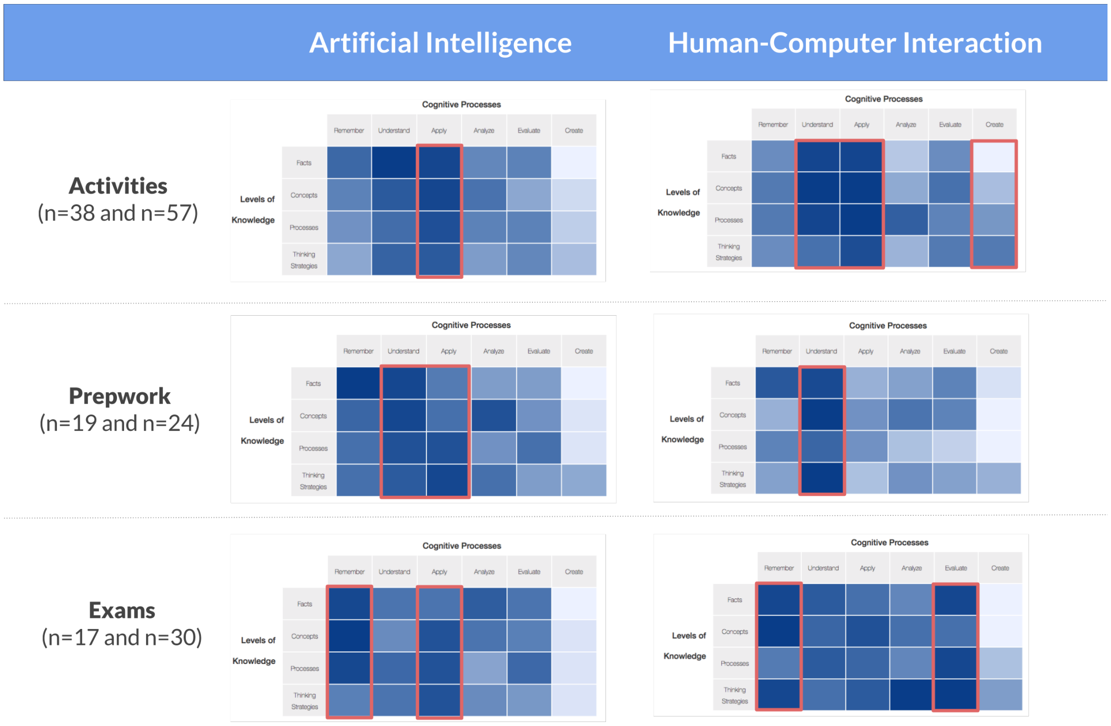


Figure 15: The results from the two summer classes in the study. Heatmaps are presented for three activity types: activities, prepwork, and exams.

Across the exams from both classes, the heatmaps showed an emphasis on *remembering* course aspects. In both exams, there were multiple choice definitional questions which may explain this emphasis. There were also some interesting differences between the two classes. In the AI exams, students had to solve problems and select the correct solution from a list. The heatmap from the AI class shows more of an emphasis on *analysis* than in the HCI class as a result. The HCI class had more emphasis on *evaluating facts* and *processes*. This may be explained by the questions that asked students to evaluate which design process could be used in a given setting. Students would have to *remember* the *facts* about each process and then *evaluate* whether it is appropriate for the given scenario.

In the prep work activities, students also indicated that they *remembered* facts,

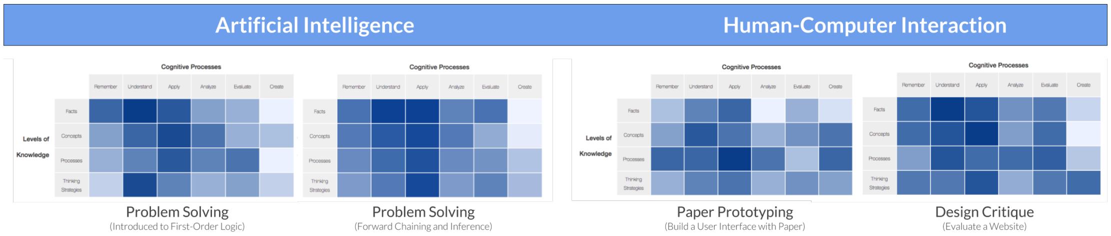


Figure 16: Four heatmaps from four different in-class learning activities.

but the main emphasis was on *understanding*. This is not surprising because in flipped classrooms, the assumption is that students build an *understanding* at home by watching lecture videos and then *apply* that understanding during class. The prep work activities also had some interesting differences between the two classes. In the AI class, students completed coding assignments at home. The class period was an opportunity for them to get some practice, which they continued at home. In the HCI class, their prep work is mostly centered around watching videos and reading from the textbook. This may explain why students in the AI class indicated that they were *applying* concepts during the prep work activities.

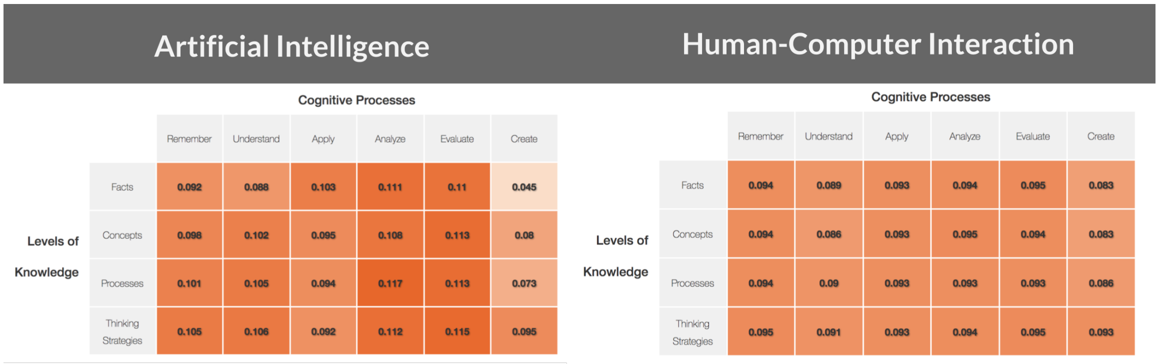
Finally, the in-class activities showed many similarities across the two classes, but also some differences. In both classes, students *applied* the course material to improve their *understanding*. In the HCI class, students indicated that they did more *evaluating* and *creating*. In the HCI class, students frequently created paper prototypes of their designs, and it is likely that this is what they referred to as *creating*. It is also interesting to see that they did not select *creating facts*. During the formative studies, when creating hints, we had trouble thinking about what creating facts might mean for students. This may suggest that some combinations of cognitive process and knowledge type either do not make sense or are uncommon.

In addition to these aggregated views, I also compared all of the in-class learning activities from the study, shown in Figure 16. In the AI heatmaps, the emphasis is mostly on *applying* what they have learned, but there is more emphasis on understanding on the day that they are initially introduced to first order logic. As students complete an associated problem solving activity, they are still trying to make sense of the concept they just learned in lecture. For the HCI class, the heatmaps are more dissimilar between tasks. In the paper prototyping task, they are *creating* paper prototypes to test their ideas. They are *applying* the prototyping *process* that they learned in the prep work. It was surprising that analysis and evaluation were less emphasized. I believe the reason is that they were not asked to start evaluating the prototypes, only to create them. Finally, the design critique was the only task that touched on many diverse parts of the matrix. Students needed to *remember* design principles, *understand* the design goals of the website, and *evaluate* whether it met those design goals and adhered to design principles.

Based on this analysis, it appears that heatmaps may effectively capture and represent students' cognitive processes. Each of the differences between the heatmaps was explained by differences between the way the courses were taught and structured. At the same time, there appears to be evidence that cognitive processes are not perfectly isolated by learning task type. It may be the case that 'prep work,' 'activities,' and 'exams' is too simple of a delineation between the different types of tasks that students engaged in. Prep work can consist of understanding new material, applying course concepts in more detail on assignments, or creating prototypes that might be evaluated in class. Despite the generic nature of these three terms, there appear to be some consistent differences across the three task types. This result provides at least partial support for the first hypothesis *H1*.

4.6.2 Heatmap Variance and Critical Mass

As mentioned in the previous section, the heatmaps are generated by crowdsourcing responses from students. The heatmaps appear to be good proxies for the cognitive processes that we would expect students to employ during the learning tasks. However, there does appear to be some variation in how students responded to the matrices. There are many reasons for these variations. First, students' experiences are probably not uniform across the class, even for the same task. For example, in a peer learning activity, cohesive teams may work very closely with each other, productive teams may split up the tasks efficiently, and dysfunctional teams may not coordinate at all. Second, students can employ different problem solving strategies, even for the same task. This is especially true in design-oriented classes like HCI. Third, some students may not buy-in to reflection or not put in the effort to respond thoughtfully when filling in the matrix. This can lead to outliers and bad data in the heatmap. Fourth, students may shift the focus of their reflection as they reflect on their learning. When students reflect on an idea or experience, their reflective frame focuses their attention on specific aspects of the concept. This frame can change as they reflect. Additionally, most concepts do not have a static, well-defined, universal meaning. Their meaning is emergent and situated in the current context. These phenomenas are known as concept shifting and they likely lead to heterogeneous reflective experiences even when reflecting on shared experiences.



Artificial Intelligence

Cognitive Processes

Levels of Knowledge	Remember	Understand	Apply	Analyze	Evaluate	Create
Facts	0.092	0.088	0.103	0.111	0.11	0.045
Concepts	0.098	0.102	0.095	0.108	0.113	0.08
Processes	0.101	0.105	0.094	0.117	0.113	0.073
Thinking Strategies	0.105	0.106	0.092	0.112	0.115	0.095

Human-Computer Interaction

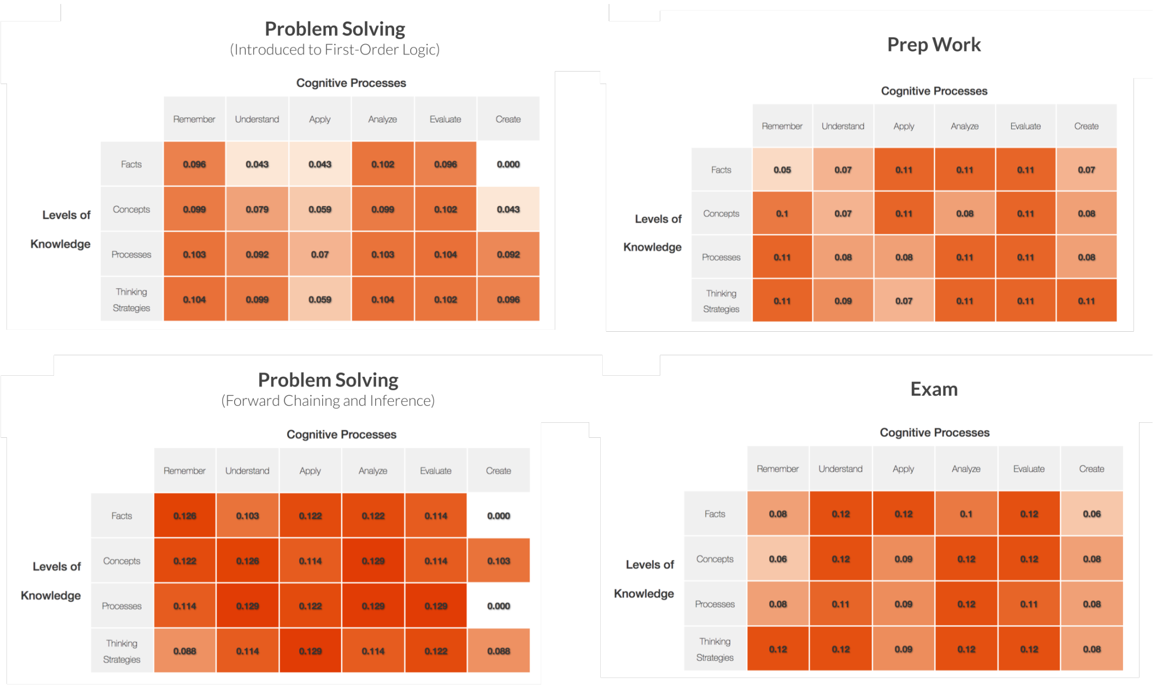
Cognitive Processes

Levels of Knowledge	Remember	Understand	Apply	Analyze	Evaluate	Create
Facts	0.094	0.089	0.093	0.094	0.095	0.083
Concepts	0.094	0.086	0.093	0.095	0.094	0.083
Processes	0.094	0.09	0.093	0.093	0.093	0.086
Thinking Strategies	0.095	0.091	0.093	0.094	0.095	0.093

Figure 17: Heatmaps that show the standard deviation for each class's responses.

As a result of these four challenges, there is likely to be at least some variation in the way that students reflect using BloomMatrix. I computed the standard deviation for each of the classes, shown in Figure 17 and for each activity. I only present the standard deviation for activities from the AI class because the standard deviation was most variable in those heatmaps. These heatmaps are shown in Figure 18. For each student in the class, each box can be either selected or unselected. The selections for each student in the class are independent from other students in the class. Each selection by each student can be considered a successive, independent trial. Consequently, combining these responses from the students for each box results in a set of binomial distributions. The standard deviation can therefore be represented by $\sqrt{\frac{p(1-p)}{n}}$.

When comparing the two classes, the standard deviation was lower for the HCI class. This is at least partially due to the increased number of responses from students in the HCI class. However, the activities in the AI class also appeared to be more focused along a single cognitive process. This is contrasted by the HCI class, where a single activity may require students to create, evaluate, and apply. Therefore, there the variability is more evenly distributed. In the AI class, the cognitive processes



Problem Solving
(Introduced to First-Order Logic)

Cognitive Processes

Levels of Knowledge	Remember	Understand	Apply	Analyze	Evaluate	Create
Facts	0.096	0.043	0.043	0.102	0.096	0.000
Concepts	0.099	0.079	0.059	0.099	0.102	0.043
Processes	0.103	0.092	0.07	0.103	0.104	0.092
Thinking Strategies	0.104	0.099	0.059	0.104	0.102	0.096

Prep Work

Cognitive Processes

Levels of Knowledge	Remember	Understand	Apply	Analyze	Evaluate	Create
Facts	0.05	0.07	0.11	0.11	0.11	0.07
Concepts	0.1	0.07	0.11	0.08	0.11	0.08
Processes	0.11	0.08	0.08	0.11	0.11	0.08
Thinking Strategies	0.11	0.09	0.07	0.11	0.11	0.11

Problem Solving
(Forward Chaining and Inference)

Cognitive Processes

Levels of Knowledge	Remember	Understand	Apply	Analyze	Evaluate	Create
Facts	0.126	0.103	0.122	0.122	0.114	0.000
Concepts	0.122	0.126	0.114	0.129	0.114	0.103
Processes	0.114	0.129	0.122	0.129	0.129	0.000
Thinking Strategies	0.088	0.114	0.129	0.114	0.122	0.088

Exam

Cognitive Processes

Levels of Knowledge	Remember	Understand	Apply	Analyze	Evaluate	Create
Facts	0.08	0.12	0.12	0.1	0.12	0.06
Concepts	0.06	0.12	0.09	0.12	0.12	0.08
Processes	0.08	0.11	0.09	0.12	0.11	0.08
Thinking Strategies	0.12	0.12	0.09	0.12	0.12	0.08

Figure 18: Heatmaps showing the standard deviation (SD) for the AI class.

were more distinguishable, as shown earlier in the heatmaps. The large variations in standard deviation between cells in the AI class, shown in Figure 18, indicate that students had higher agreement along some cells than others. For instance, students in the AI class generally agreed that they did not do much ‘creating’, especially for ‘create facts.’ Further work is needed to see whether the cognitive processes and knowledge types are independent or if there is some covariance. It is very likely that there is some covariance between the rows and columns in the matrix.

These results suggest that class size is not the only aspect that affects the variance present in the heatmaps. There are other facts such as activity type that may also affect students ability to distinguish between the different cognitive processes and levels of knowledge. More work is needed to better understand what the variance in the heatmaps represents. Does it represent confusion about what cognitive processes students are engaging in, does it represent a misunderstanding of the categories in the heatmap, or does it represent variations in students learning experiences? Based on these open questions, it is hard to make definite claims about *H2*. However, it does appear that class size may be one of the contributing factors to the variance observed in the heatmaps. Despite the variance that was present in the heatmaps, trends could still be identified as discussed in the last section.

4.6.3 Students' Reflective Preferences

On the last day of class, I asked students to complete a short survey about their experiences using the RST and completing reflective writing assignments. In the survey, students were asked open-ended questions and and asked to rate statements on a Likert scale. They were also asked which of the two reflection styles they preferred, BloomMatrix or reflective writing assignments. These surveys were analyzed using a mixed-methods approach to understand students' experiences and preferences. In total, 49 students responded to the final survey (21 and 28 students from the AI and HCI classes respectively).

4.6.3.1 Likert Scale Responses

To understand how students felt about different aspects of reflection, I provided students with a series of statements about their experience and asked them to rate them on a 5-point Likert scale (1=Strongly Disagree, 5=Strongly Agree). The statements and students' aggregated responses are shown in Figure 19. The statements included information about BloomMatrix and reflective writing to make enable comparisons. From a quick glance, the responses were mostly neutral with only a few

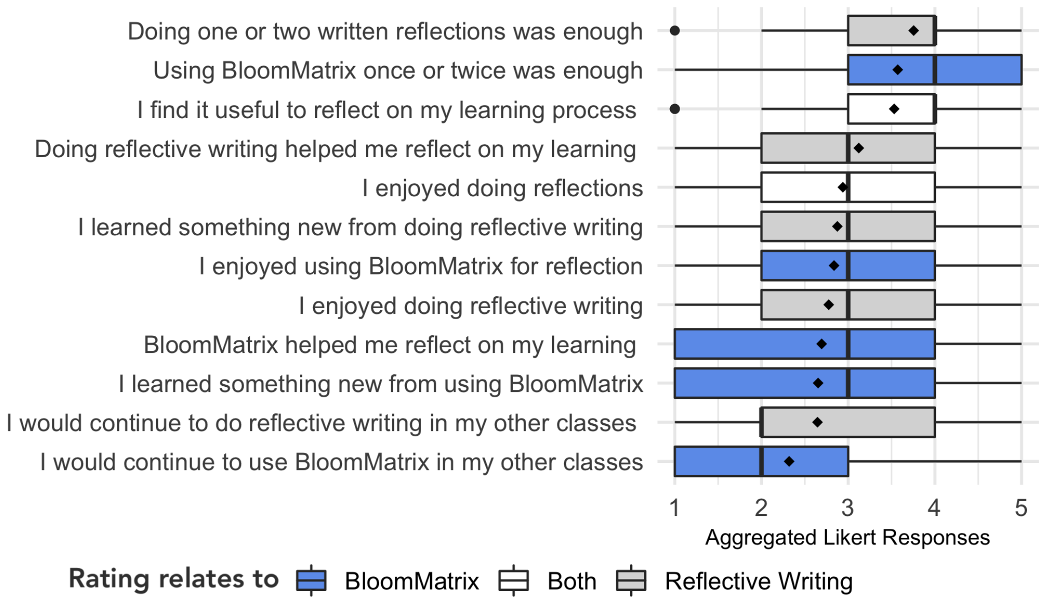


Figure 19: Boxplots of students' Likert ratings (5 = Strongly Agree) for the statements about students' experiences, ordered vertically by mean (black diamonds).

clear opinions emerging. First, students generally agreed that it is useful for them to reflect on their learning, even if it is not necessarily enjoyable. Second, students generally agreed that using BloomMatrix once or twice was more than enough. They also stated that they would not want to continue using BloomMatrix in their other classes. However, they had very similar responses for their reflective writing assignments. Despite agreeing that reflection is helpful, neither BloomMatrix, nor reflective writing appeared to meet their needs. While not strongly conclusive, these results provide more evidence that cultivating reflection in the classroom is challenging.

In addition to these statements, we also asked students which type of reflection they preferred. The results are shown in Figure 20. Overall, students indicated that they preferred the reflective writing assignments to using BloomMatrix. However, there was an interesting difference between the two classes. In the AI class, students preferred BloomMatrix, but in the HCI class, they preferred reflective writing. The

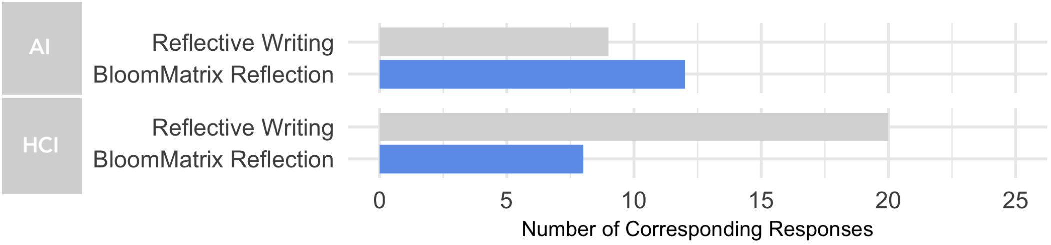


Figure 20: Students in the AI class preferred BloomMatrix. Students in the HCI class strongly preferred reflective writing. Overall, reflective writing was preferred.

difference in the AI class was marginal, but in the HCI class, students clearly preferred reflective writing. It is possible that in the HCI class, students are doing more writing assignments and they are more comfortable with them. Similarly, in the AI class there is a focus on computation and programming, but there is little to no focus on writing. It may be that these AI students saw BloomMatrix as more technical and data-driven, therefore, they were more comfortable with it. Finally, the variance was lower in the AI class and so the heatmap may have better distinguished between the different cells, resulting in more meaningful visual representations.

Considering the Likert responses and their stated preferences, students appeared to dislike both methods for supporting reflection, but did generally prefer the reflective writing assignments. This provides some evidence against *H3*. But, given that neither type of reflection was perceived by students as being particularly helpful or enjoyable, this may be more reflective of the difficulties of supporting student reflection. It is important to note that students' perceptions of whether these activities were helpful and whether they actually did help students are two different things. But these challenges may begin to explain the dearth of RSTs currently available for students. When considering these results with the evidence for *H1*, it is possible that

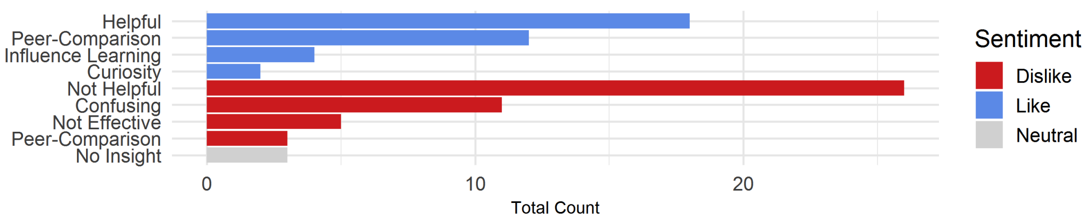


Figure 21: The dominant codes based on students' opinions about heatmaps.

BloomMatrix is better suited for instructor feedback than for student reflection.

4.6.3.2 Free Response Questions

Students answered a series of free response questions about their experiences. These questions included "in what ways was the heatmap helpful or not helpful?" and "why did you prefer the type of reflection that you indicated?". In this section, I present the resulting themes and codes from the analysis. These student responses and resulting codes shed light on the students' experiences and the preferences that they indicated in the previous section. The quotes presented in the following sections were extracted from the students' written responses. I fixed typos and made small grammatical changes, such as removing a redundant word or adding commas, where necessary to improve readability.

4.6.3.3 Opinions about the Heatmap

The first aspect that students were asked about was their perceptions of the heatmap. The dominant codes that resulted from analyzing their responses are shown in Figure 21. Based on the codes, students liked that the heatmap presented information about their peers and helped them to do comparisons, e.g.: *"I liked the heat map very much because I could see how similar to other students my thinking process was"* and *"The heat map gave me an insight on how my peers are using the content"*.

Doing comparisons allowed students to ensure that they were on the right track, for example *"seeing my results compared to those aggregated by the class lets me know if I am missing something that I should be looking into"* and *"helped me gauge if I was on the same track as everyone else or if there was another way in which I should be thinking about the problem."* Multiple students mentioned that this benefit was unique and could not be achieved through reflective writing, for example *"even if reflective writing is good, it does not offer me a point of view where I can see what the others do."* These comparisons appear to be helping some students think about the class more holistically and consider the perspectives of other students. This provides some evidence to support *H4*. Some students indicated that the increased awareness influenced their thinking. One student thought that *"it might make some people think differently about their experience."* Another student described personally experiencing this, writing that BloomMatrix challenged her to consider *"if there was another way in which I should be thinking about the problem."* The holistic perspective and new ways of thinking may have resulted in intentions to change behaviors. For example, two students described their intentions to make changes, *"and how else I should study to do well"* and *"how I should possibly change my learning approach."*

Despite these benefits attributed to the heatmap, there were students who described the heatmap as *"pointless"* or *"confusing"*. "Confusing" was one of the most dominant codes identified in students' survey responses, with 11 students expressing that idea. Students described three sub-themes when talking about why the heatmap was confusing: Distinctiveness (4), Layout/Categories (3), and Purpose (2). The main aspect that was confusing for students was that the dimensions and rows did not feel distinct enough for them to know which cells to select. One student wrote that *"The subcategories didn't feel very distinguished from one another. I wasn't sure what they were asking."* Another student echoed this opinion, *"many of the BloomMatrix categories didn't seem distinct enough."* For students who had trouble understanding the matrix, none of them mentioned the contextual hints. It is unclear whether they used them or not. A small minority of students also explained that peer comparisons were not helpful for them. One student expressed this idea, writing that the heatmap was *"not much help, seeing the hot spots of my class gives me no useful information"*. They further elaborated, *"as in 'why should I care to see how my class is doing overall?'."* Another student also had this complaint, writing *"but I didn't find it useful or helpful since it just showed what other people thought."* Although the majority of students liked being able to compare themselves with their peers, this was not every student's preference. Neither student commented on whether the individual reflection phase of BloomMatrix was helpful in any way for them. Finally, four students expressed having trouble knowing what to do with matrix, one student described it as *"vague"* and another as lacking a *"goal."*

The results from this free response question are difficult to interpret. 12 students enjoyed being able to compare their results with their peers. However, 3 students indicated that they didn't care about what other students in the class were doing. In previous sections, I was able to make sense of the heatmaps and identify trends based on the activities. However, some students in the study explicitly mentioned that they were confused by the matrix and that they did not have any insights or learn anything new. On the other hand, some students described how the heatmaps helped

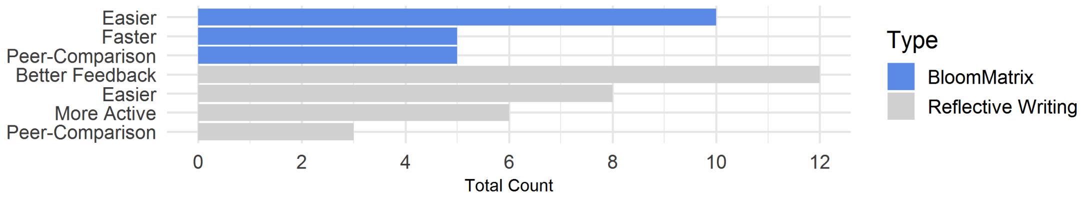


Figure 22: The dominant codes that resulted from students open-ended responses about why they preferred the type of reflection that they chose.

with their learning, broadened their perspective, and led to intentions to change their behaviors. Going forward, it is probably necessary to provide students with guidance for interpreting trends in the heatmaps. The contextual hints and responsive enabling is helpful for filling in the matrix. However, students may also need help interpreting the heatmap. One way to do this would be to create an anonymous discussion forum below each matrix where students can identify trends and ask questions. Another feature which I built, but was not used for this study, was an ability for instructors to fill out the matrix by selecting cells they expected students to select for that activity. Students could see the instructors' selects alongside the aggregated heatmap of their peers' responses. This "ground truth" might have been helpful for students to reflect on in addition to looking at their peers' responses.

4.6.3.4 Comparing BloomMatrix and Written Reflections

In the survey, students were asked to choose the type of reflection that they preferred, and the results from that question were presented earlier in Figure 20. I also asked students to elaborate on why they preferred the type of reflection that they indicated. The dominant codes for their responses to this question, separated by reflection type, are shown in Figure 22. Similar to their responses about the heatmap, students who preferred BloomMatrix liked being able to compare themselves to others. For students who preferred reflective writing, the primary benefit that students cited was being able to get better feedback and explore their ideas in more detail. As an example, one student wrote in the surveys, *"you can get better feedback with writing. Open-ended questions=more information given from user."*

Regardless of the reflection type that they preferred most, students indicated that ease of use was important to them. For both preferences students cited reasons that they thought BloomMatrix or reflective writing was easier to use. The reasons that students provided for why BloomMatrix was easier included that it was *"easier to select my thoughts if they are categorized"* and *"Gave better format to sit and think about how or why I did something."* For reflective writing, students thought that *"Questions are easier to understand"*, *"I feel it is easier to express myself"*, and *"Reflective writing lets me give simple answers for questions in a way that makes sense to me."* Ease of use was a primary concern for both preferences. I expected that BloomMatrix would be easier for students because they can quickly fill in the boxes. However, it appears from the responses in the last section about the heatmap that filling in the boxes required a lot of cognitive effort. For some students, the cognitive effort to understand the matrix may have reduced the cognitive effort they could employ on evaluating the learning activity. For RSTs that support reflecting on complex topics, such as cognition, ease of use is not straight-foward.

Students who preferred reflective writing also explained how it allowed them to be more expressive and was more generalizable to their experiences. They indicated that the cognitive processes that they experienced during class didn't always map neatly to the matrix. One student suggested that *"BloomMatrix was more of a yes or no question"* which did not capture their experience. Another said that reflective writing allowed them to be *"more verbose about the reflection, rather than filling in the pre-filled categories of the BloomMatrix"*. One student wrote that reflective writing *"allows me to actually reflect [on] what I think and not select pre-determined answers."* Two observations can be made from these responses. First, students have complex experiences during learning activities and it may be difficult to report those in a standard way. In this work, students' reflections were standardized to share their responses with the rest of the class. However, for some students this may not have been desirable. Based on this observation, future work may explore how students represent their cognition in the reflective writing assignments. There may be some standard form that emerges which could be shared across students. Second, it is possible that BloomMatrix may have appeared to be too easy for some students. Students may have felt that *'clicking boxes'* did not challenge them to deeply reflect on their experience. This second aspect speaks to a potential paradox about *'ease of use'*. As seen previously, it is preferred by students, but it may not be in their best interest and may not lead to deep critical reflection. A few of the students made statements to this point. One student stated that the main benefit of reflective writing is that it *"is more active and I actually had to really think about things more."*

Finally, students' reflective goals did not appear to be uniform across the class. Some students saw reflection as something that was personally valuable, others saw it as a way to provide feedback to the instructor. For instance, one student who preferred reflective writing wrote that "you get better feedback with writing. Open-ended questions=more information given *from the user*". It appears from this comment that the student saw reflection as a way to get data *from* them rather than an activity that was intended *for* them. In another students' response, he wrote *"I feel like it [, reflective writing,] has [a] better way of communication."* Although the purpose of reflection was conveyed to students on their first day, it is not clear how they interpreted that purpose when they did their reflections or responded to the survey. In the future, I plan to ask students about their explicit goals for reflection at the beginning and end of the semester.

4.7 Discussion and Summary

Based on the results presented in this chapter, there was mixed support for and against the hypotheses. The hypotheses are presented again below for convenience:

H1. Heatmaps will be similar for the same type of learning activity, such as watching lecture videos (prep work), regardless of content and will differ otherwise.

H2. The heatmaps will contain less variance for larger classes than for smaller classes. In larger classes, the effect of outliers on the heatmap is less significant.

H3. Students will prefer using BloomMatrix to written reflections because it is faster to fill out and easier to use.

H4. Students will talk about their learning more holistically. They will have increased awareness of the aspects that affect their learning.

The first research question was about whether the heatmaps would capture different patterns of cognition for each type of learning activity. The heatmaps showed that the trends of cognition appeared to be explained by activity type. This is an exciting avenue to consider BloomMatrix as a tool for instructor ref, rather than for student reflection. Instructors can reflect on their class using these representations the first time they teach a new course to ensure that students are receiving a holistic learning experience. They can also present these heatmaps to students at the beginning of each semester to show how different activities form a holistic learning environment for students. It could serve as visual evidence that no one activity fully targets all aspects of their cognition and knowledge.

I also explored the variations between these different heatmaps and explained some of the possible sources of these variations. By computing the variance and showing it in the heatmap, it was obvious for the AI class that students had better agreement on some aspects of the matrix than on others. However, the results about *H2* were mixed. There was some evidence that class size affects the heatmaps, as the HCI class had more uniform variance across heatmaps. However, the activity type also appeared to affect this variance. More work is needed to understand these aspects. In the future, I plan to use PCA and clustering to try to identify areas of the heatmap that have higher variance for specific groups of students or for activity types.

For *H3*, students generally appeared to prefer reflective writing. This provides evidence against *H3*. At the same time, students did not really have strong preferences for either form of reflection. Most of the criticisms about BloomMatrix were that it was confusing to use. When coding the survey questions, 11 of the 49 students who responded indicated that despite the hints, the responsive enabling, and the in-class walk-through, that they were confused or unable to delineate between the columns. It is possible that more scaffolding is needed, such as specific tasks, to help students know what to look for in the heatmaps. However, students also indicated that placing their cognition into the heatmap was challenging. The standard representation did not match their non-standard learning experiences. As future work, I plan to consider ways to construct standard visual representations from non-standard responses from students. Such a system would leverage the open-endedness of reflective writing with the ability to compare oneself with their peers, which 8 of the 49 students mentioned as a reason for their preference and 12 mentioned as the reason they liked the heatmaps.

Finally, when considering *H4*, there was evidence that BloomMatrix led some students to consider their experiences more broadly. There was also some evidence that students wanted to make changes based on what they saw in the heatmaps. These changes were modest and mostly related to their study habits. However, no students mentioned behavioral change or broadened perspectives that resulted from using reflective writing. With this comparison in mind, there appears to be some evidence for *H4*. However, it is important to keep in mind that intentions to change do not necessarily translate to actual behavioral change. It is also important to note that only some of the students experienced this increased awareness or increased agency.

In addition to these other findings, my main take away from this work is that supporting reflection is difficult, especially around topics that students may not have a good operationalized understanding of, such as cognition. Students frequently cited *Ease of Use* as something that they valued. For instance, being able to get quick insights and do less work was frequently mentioned in student responses. However, reflection is often not quick or easy. It is difficult work and as described by students, often not enjoyable. Although students acknowledged that they believed reflection were beneficial, they did not seem to like either of the two types of reflection presented to them. Students juggle many different things and it is likely that neither BloomMatrix nor reflective writing had a high enough return on investment for most students to see them as being valuable. In this study, each of the instructors hoped that BloomMatrix might help students value the prep work. For each of them, conveying the value of these prep activities was difficult. While it is unclear whether reflection achieved that goal, it is interesting to consider that students do not complete this activity which is explicitly tied to their grade. If students are not willing to complete prep work activities which directly affect their grade, it is not surprising that they will devalue reflection which has no explicit effect on their grade and future careers. Finally, it could be argued that HCI students were primed to consider ease of use because it is a concept that is discussed frequently in the course material. However, the fact that both classes mentioned this aspect, leads me to consider that this is an important aspect to students in general.

Tying back in to *Ease of Use,* students are skeptical of any additional demands on their time. Therefore, it is important to build RSTs that can minimize the effort put in while maximizing the insights that students receive. But it is also important to consider reflection as a literacy that should be explicitly taught alongside the other digital literacies, such as design thinking and computational thinking. Reflection should be a more explicit part of students learning and they should know how it ties into their eventual careers. For this to be possible, reflection needs to be treated as something more than just a quick activity to get through among many other learning activities in which they participate.

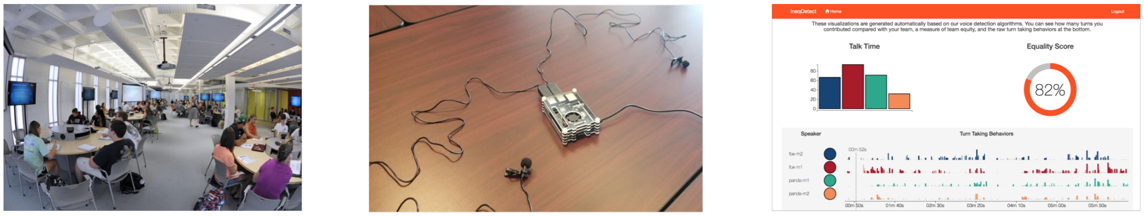
CHAPTER 5: INEQDETECT: REFLECTING ON COLLABORATION

IneqDetect is a system that records and visualizes students' group conversations. By reflecting on these visualizations after class, students can identify and improve on problems with their group dynamics, such as conversational inequality. During class, students wear lapel microphones which are attached to small hardware devices that are placed on the table. At the end of class, the recordings are sent to a server and automatically processed to determine which student is speaking during a given time segment. These segments are visualized on a scrollable time line along with a barchart of the total talk time by student and the amount of conversational inequality in the group. An overview of the system and study context are shown in Figure 23.

In this chapter, I will start by presenting the motivation that inspired this research. I will describe the system and present the the theories, models, technology, designs, and field notes which informed the design of the IneqDetect system. Finally, I'll present the study design and the results from deploying IneqDetect in four different classes in the College of Computing and Informatics.

5.1 Motivation

Traditional lecture classrooms are slowly being supplanted by more active learning environments where students have opportunities to interact with the material and each other during class-time. Moving homework and assignments into the class-



Active Learning Classroom

Multiple Raspberry Pis Record Conversations

These visualizations are generated automatically based on our voice detection algorithms. You can see how many turns you contributed compared with your team, a measure of team equity, and the raw turn taking behaviors at the bottom.

Student's Conversations are Visualized for Reflection

Figure 23: An overview of the IneqDetect system. In active learning classrooms, Raspberry Pi devices are used to record students' group conversations. At the end of class, students reflect on visualizations of these conversations.

room gives instructors an opportunity to support students as they struggle with the material. Misconceptions can be addressed directly and students can reach out to their peers for emotional and conceptual support. Integrating aspects of collaborative learning can be especially helpful in large classes where instructors may have trouble providing one-on-one support for individual students. In these social learning environments, students develop social skills, negotiate their understanding with team-members, and provide intellectual and emotional support for their peers. Despite these many benefits, social learning environments are also a microcosm of society. And there are problems that can arise when students interact with each other. For instance, sociocultural inequities have been observed in students' group discussions [100] and in pair programming activities [153]. Sociocultural inequities are conversational disparities that can exist between students. Various social factors can serve to elevate the voices of some students and systematically silence the voices of other students. For example, students can also become convinced of incorrect information by more persuasive students. While often unintentional, these negative social interactions can play into stereotypes of Computer Science (CS), such as it being competitive, asocial, singularly-focused, and primarily male [99].

I also observed these problems in two initial studies that I conducted to understand how students collaborate during peer instruction learning activities [106, 105]. Peer instruction is an instructional style where students work together to answer quiz questions [42]. The groups in these studies were in two conditions: co-located or on-line. The on-line groups communicated using Google Hangouts. In both cases, students were presented with a series of questions, discussed each question as a group, and then answered individually. Students were encouraged, but not required, to reach a consensus before answering. The first study took place in a laboratory setting which was repeated in a classroom setting in the second study. While conducting these studies, I observed many instances where students did not contribute equally. There was also evidence of both collaborative and anti-collaborative behaviors in the groups. In these studies, performance and communication patterns varied widely from team to team. This work inspired me to start thinking about tools that could help to improve the group dynamics in these teams. I developed IneqDetect as a tool to investigate and operationalize group dynamics in these groups, and also to provide students with the agency to begin to address these problems on their own.

Given the many opportunities afforded by collaborative learning, it is essential to find ways to support collaboration while also promoting sociocultural equity in groups. IneqDetect provides students with tools to investigate and address these problems themselves. By reflecting on their group conversations, students may become more aware of how they interact with their peers. They may also identify problems, such as conversational inequality, that they can address themselves. By tracking their interactions from week-to-week they can also experiment with different ways of interacting. For instance, if they identify that one student is not contributing much to the conversation, they could try asking for their feedback or getting to know them better to make them feel more comfortable in the group. In these ways, IneqDetect captures social behaviors and presents them to increase awareness and agency within groups. Ideally, this can lead to more equitable groups and behavioral changes that improve group dynamics.

5.2 Theory, Technology, Designs, and Field Notes

IneqDetect is a reflection support tool (RST) designed to record and visualize conversations to support reflection and improve conversational equality within groups. In Chapter 2, I presented reflection and RSTS. In this chapter, I provide additional theories and related work that are relevant to the design and development of IneqDetect. These include social theory, as it applies to collaboration and equity, speaker recognition algorithms, observational studies about collaboration in CS, and related designs that visualize verbal and textual conversations. These aspects will be presented in the "theory, technology, designs, and field notes" sub-sections. This vocabulary was borrowed from the RtD framework as a way to indicate the ways in which multiple research domains have contributed to this work.

5.2.1 Theory: Social Theory and Equity

Broadening participation in Computer Science Education has been a long standing goal for computer science educators and administrators. Conversations about the gender gap [162, 37], a lack of racial diversity [69], and disparities between students' socioeconomic status [109] are common and often revolve around the ideas about equitable access to education opportunities and preparedness. These inequities are often explained by systemic, structural barriers, where students do not have physical access to artifacts, tools, and institutions that are necessary for education [109, 74]. Considering these structural barriers is important to improve equity for all students, but these barriers alone do not tell the whole story. Inequality in the classroom also has components that are sociocultural [124], perceptual [30], pedagogical [164], and experiential [143]. Sociocultural inequities, such as equal access to the conversational floor, have been identified as a challenge for student dyads as they learn collaboratively [100]. In that work, members of inequitable teams were more likely to rush and focus on completing the learning activity rather than on the process of learning collaboratively.

Taking a sociocultural perspective is an important complementary avenue to investigate because identity is internalized through social interactions and is situated in a cultural context. Social norms and culture can have a very strong impact on student persistence and success, especially as students are actively developing their identities as computing professionals. For instance, men and women often have different and incorrect perceptions of what computer scientists do [30]. And these perceptions about computer science are often formed during students' initial experiences with computing [30] which are increasingly rooted in social interactions with their peers. For this reason, the goal of broadening participation in CS must account for how pedagogical practices that instructors employ affect students' identities, attitudes, and abilities to express themselves in their learning communities. Furthermore, as seen in previous work [100], not accounting for communication patterns may also reduce the effectiveness of collaborative educational activities.

5.2.2 Technology: Speaker Recognition

There are a variety of techniques that have been used to identify who is speaking during a given time. These techniques include speaker classification, thresholding individual microphone signals, blind signal separation, and bone conduction. Speaker classification, also known as speaker diarization, is an unsupervised process of identifying each speaker within an audio stream and determining at which points in time they were speaking [6]. This approach is often used with a single audio track, such as for news broadcasts [13]. Another approach for detecting who is speaking is to record audio on multiple microphones and then use a threshold to determine whether or not the microphone is active, a process known as *Voice Activity Detection* (VAD) and *Speech Activity Detection* (SAD) [73]. Finally, blind source separation can be used with multiple microphones to help triangulate a speaker based on the intensity of their voice as measured by at least three microphones [166].

The approaches listed above employ microphones or microphone arrays to detect who is speaking at any given time. Physiological sensors, such as bone conduction, are another approach which can be more robust, less obtrusive, but also do not typically result in data that can be transcribed. For example, Skach et al. detect conversation based on variations in pressure that is applied when a person sitting on the pressure sensing seat cover speaks [150]. These types of solutions could be seamlessly integrated into existing classrooms that feature desks or tables and chairs, but students are also limited to conversing in a seated position. This might discourage students from getting up and helping others or meeting students at other tables. Finally, physiological sensors may also be preferred because they are not visible to students. Students may behave differently when they know that their conversations are being recorded. A microphone is a constant reminder to students that they are being recorded, but physiological sensors can be integrated in ways that do not draw attention to them. However, this approach raises ethical issues when students forget that they are being recorded.

5.2.3 Field Notes: Turn-taking in CS Education

Many ethnographic approaches exist for understanding group collaboration. Lewis and Shah conducted observational studies of students working in dyads and observed that when inequities existed within the pairs, students were more likely to rush through activities [100]. In their work, they suggest that attitude and pedagogy can affect equity and interactions. They designed their pedagogy to encourage equity, but in spite of this, some pairs were still inequitable. In addition to attitude, Deitrick et al. conducted observational studies to show how the ability to align across a set of dimensions can affect collaboration [46]. For instance, students could not align on a style of communication, on their intended goal, or on how they would achieve the goal. In some cases these problems can also be addressed with pedagogy, such as making turns explicit or specifying the goals in more detail. But these solutions limit the freedom students have to engage in authentic problem solving. For these reasons, reflection may be helpful as a way to give students agency to foster equity within their groups.

5.2.4 Designs: Visualizing Conversations

Visualizing conversations helps to support real-time or summative analysis of conversational content and behaviors. Most visualizations of conversations focus on representing either spoken, verbal communication or text-based computer-mediated communication (e.g.: discussion forums or chats). Some visualizations also include gestures and physical interactions, such as the Collaid system [110]. Analyzing verbal conversation typically focuses on turn-taking behaviors and the volume of speech, whereas text-based conversations often leverage the threaded nature of discussion forums to show who responded to whom. Text-based conversations also do not need to be transcribed. Therefore, text-based conversations often represent what was said through topics, relationships, and sentiment.

Systems that visualize verbal data include the Conversational Clock [17] and VizScribe [31]. The Conversational Clock [17] allows students to observe a summary and the details of their conversation in real-time. Using a clock as a visual metaphor, each minute of turn-taking segments is wrapped in a circle around previous minutes in the conversation. This conveys a visual history of collaboration, but prioritizes recent parts of the conversation. Students use the visualization to monitor their conversation in real-time. VizScribe shows designers conversations over time in a summative representation that they can review after collaborating. These two systems attempt to support in-action and on-action reflection respectively. In the case of the Conversational Clock, speakers rarely looked at the visualization and listeners often looked at the clock. This suggests that more complex visualizations have the potential to be distracting when the current task, such as speaking, demands cognitive effort.

Visualizations that use IRC or Discussion forums as the source of conversation include Coterie [152], PeopleGarden [165], and Loom [26]. Judith Donath describes these three representations in some detail along with issues to consider when visualizing social data [48]. She describes how visualizing conversations is challenging because there is a tension between traditional visualization principles, representing the semantic meaning of conversation, and the culturally determined nature of social data. For example, she reviews PeopleGarden, a system which represents each discussant by a flower, where the height represents the amount of talk contributed. She explains how this might be motivating for some users but in some contexts, such as in heated arguments, it might be inappropriate. Furthermore, she explains that the mapping between flower and discussant could be confusing for users.

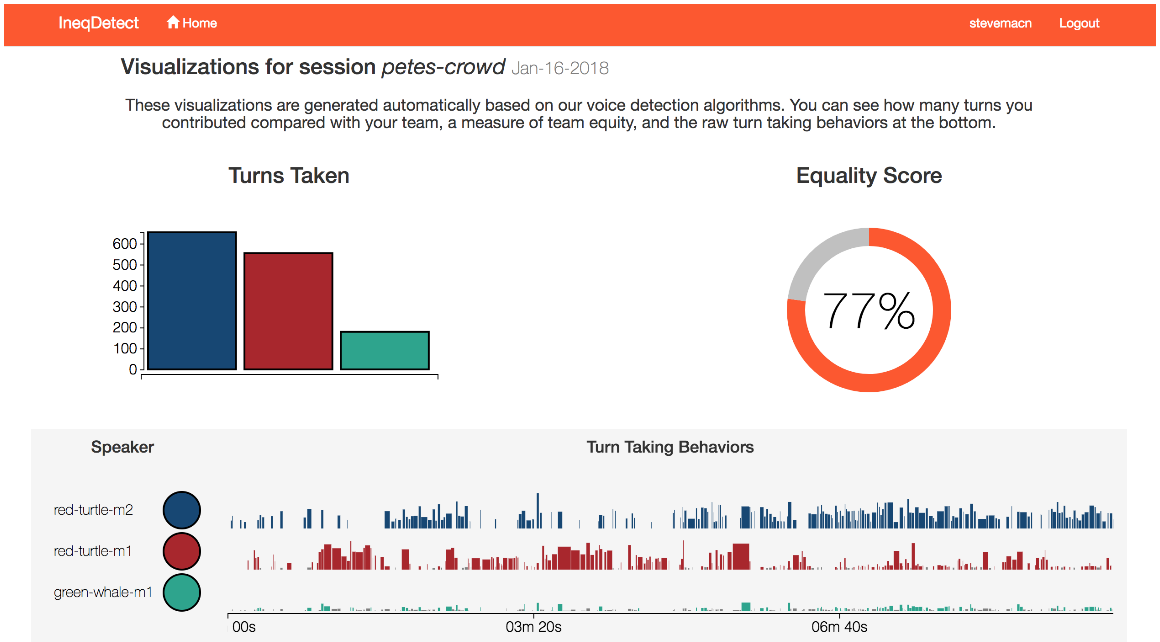
5.3 The IneqDetect System

IneqDetect visualizes conversations to support student reflection. It records conversations on hardware devices and converts the recordings into segments of detected voice, which can be visualized for student to explore after learning activities. The visualization presented to students can be seen in Figure 24. In this section, I describe how the system and its implementation.

5.3.1 System Design

IneqDetect consists of multiple lapel microphones connected to Raspberry Pi devices [4]. Each Raspberry Pi can record up to two speakers. For larger groups, multiple

[4] https://www.raspberrypi.org/



Visualizations for session *petes-crowd* Jan-16-2018

These visualizations are generated automatically based on our voice detection algorithms. You can see how many turns you contributed compared with your team, a measure of team equity, and the raw turn taking behaviors at the bottom.

Figure 24: The visualization dashboard that is presented to students after they are finished collaborating in their groups. The top left shows time that each person spoke, top right shows a measure of conversational equity (Gini coefficient), and the bottom shows turns detected over time for each team member.

Raspberry Pi devices were used and digitally synchronized. At the end of class, the devices send the recordings to a server to be pre-processed, analyzed, and visualized. On average, the visualizations were ready to be viewed after five to ten seconds. At the end of class, students review these visualization to see when they and their group members spoke. They can also view summary statistics of their interaction, such as the amount of conversational inequality that was detected within their group. The system diagram in Figure 25 shows how the various components fit together.

The Raspberry Pi devices are placed on each table and share a portable battery. In the classrooms in our department there is not a reliable power source that is close to each desk. The battery is capable of powering up to three Raspberry Pi devices for a few hours, which reduced the need for constant charging. Due to bandwidth and

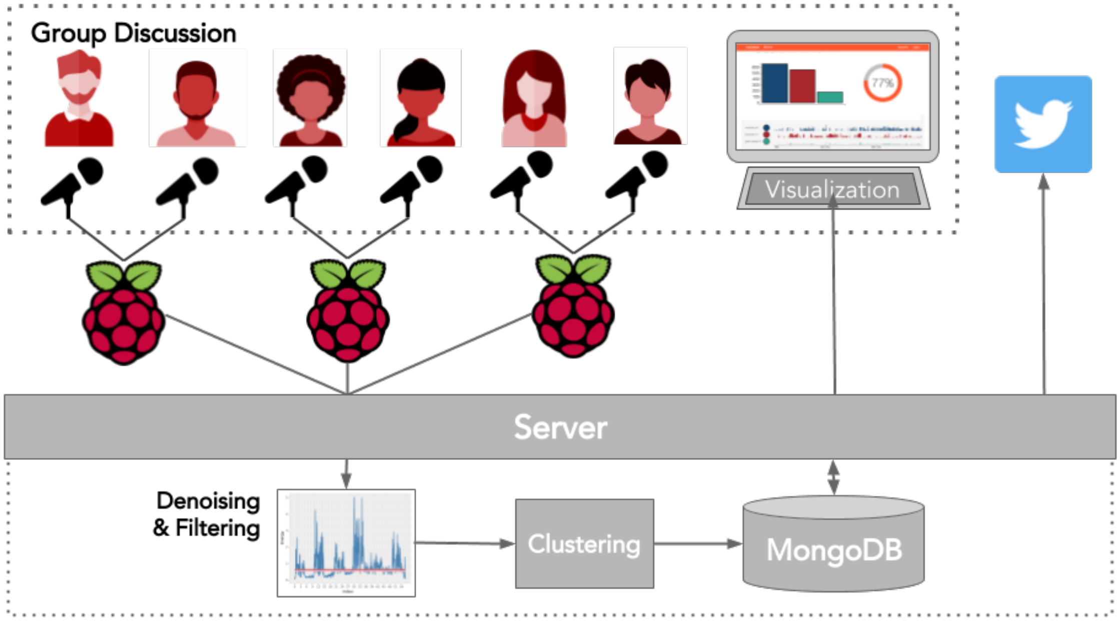


Figure 25: A system diagram of the hardware devices, microphones, server, and visualization dashboard. Students' discussions were recorded, filtered and de-noised, clustered to remove cross-talk, and then saved to a database as segments. The visualization represents these segments for students to reflect on as a group.

power restrictions, each Raspberry Pi can handle up to two microphones. For a team of six students, three Raspberry Pis are used. The lapel microphones are corded, which means students need to stay within a few feet of their corresponding Raspberry Pi. Each Raspberry Pi is labeled with stickers that feature cartoon animals. These animals help students identify themselves in the visualization, but also protect their anonymity. Recording can be started or terminated on all devices by tweeting commands on the IneqDetect Twitter account, using the Twitter API [5]. This was done to provide instructors with the ability to start and stop recordings on their own.

The server that receives the recordings from the devices processes them and also hosts the web app and visualization dashboard. The server communicates with the

[5] https://developer.twitter.com/en/docs.html

devices using web sockets and sends the stop and start signals to coordinate the devices. At the end of the class, a signal is sent to all of the devices and each device sends its recordings to the server. Each of these recordings is processed by the server and then visualized to support reflection. Students can log in to view the visualization dashboards at any time. In my studies, students viewed the visualization dashboard on a single computer to support group reflection and to encourage discussion.

5.3.1.1 Analyzing the Audio Recordings

To analyze the audio recordings, a variety of techniques were used. These techniques accounted for the real-world environment in which the students' conversations were recorded. The first challenge when recording in classrooms is that there is often a lot of background noise. This background noise is not constant and it varies in intensity throughout the class period. There are times when the class is nearly silent and other times when students are moving around and shouting over the background noise to be heard by the team. The second challenge is that many of the techniques which have classically been used to triangulate audio sources are ill-equipped to deal with phase shifts, echoes, and offsets that are caused by the physical structure of the classroom. In our pilot studies, I found that common techniques, such as independent component analysis (ICA), were ineffective.

After the pilot study, I analyzed the data using a signal processing approach with adaptive filtering. The first step was to denoise the audio signal. I used spectral whitening and a fast Fourier transform (FFT) on windows of the audio signal as part of this pre-processing step. From this signal, the voice features were extracted as Mel-frequency cepstral coefficients (MFCC). I used only the first 12 coefficients, which is common for voice activity detection (VAD) [86]. Using these coefficients, the energy was computed across these speech features during a given time interval. This energy-based approach is common for VAD [158]. Finally, the energies were compared across all of the microphones using a moving window. Initially, any energy signals that were two standard deviations from the mean were removed. This was ineffective in the classroom, where the noise detected in the classroom varied widely. Instead, an adaptive threshold was applied to these windows using k-means clustering, with k determined using the elbow method. This reduced the amount of cross-talk that was picked up on each microphone. The resulting solution was much better at distinguishing between speakers, even when they sat close together. In many classes, students were seated approximately three to four feet apart.

5.3.1.2 Visualizing Students' Conversations

The visualization dashboard for IneqDetect was designed with Schniederman's *Visual Information-Seeking Mantra* in mind, "overview first, zoom and filter, then details on demand" [148]. There are three components in the dashboard, as seen in Figure 24. Two of these components provide summaries, which include a measure of conversational equality in the group's discussion and the total talk time by speaker. The equality measure was computed using the inverse of the Gini coefficient and represented as a percentage. The Gini coefficient is a measure of inequality, which is typically applied in financial contexts, but also in evaluating conversational inequality [153]. Comparisons between speakers can also be made using a bar chart, which represents the total talk time for each speaker. Finally, a scrollable timeline is presented along the bottom of the dashboard. The timeline shows areas where each speaker's voice was detected. The timeline can also show overlaps where multiple speakers were detected at the same time.

This approach quantifies collaboration, rather than using qualitative aspects of conversation. Previous work used the Google Hangouts API to determine who was speaking in on-line conversations [153]. Lewis and Shah also manually coded turns to quantify collaboration [100]. This was done to promote student self-disclosure and because other aspects of conversation, such as paralinguistic cues, are less interpretable by students. Especially in the context of equity and conversational equality, access to the conversational floor is a clear measure

5.4 Research Questions

IneqDetect was designed to improve conversational equality by making students more aware of group dynamics through reflection. By visualizing students' conversations from week-to-week, students were afforded an opportunity to reflect on and experiment with different ways of interacting with their groups. By deploying IneqDetect in the classroom, I intended to explore the following research questions:

R1. **Awareness** Are students aware how much they talk within their groups? Does IneqDetect improve their awareness?

R2. **Equality** How equitable were the students' group conversations? Does IneqDetect improve conversational equality?

R3. **Hypotheses** Do students generate and test their own hypotheses?

R4. **Insights** Do students obtain insights from using IneqDetect?

R5. **Agency** Do students change their behaviors after using IneqDetect?

R6. **Enjoyment** Do students like using IneqDetect?

5.5 Hypotheses

Based on these research questions, I formed the following hypotheses about students' group dynamics and about their use of IneqDetect:

H1. Students will not be able to accurately estimate how much they speak.

H2. IneqDetect will improve conversational equality in groups.

H3. Students will experiment with their group dynamics using IneqDetect.

H4. Students will obtain insights about their group dynamics.

H5. IneqDetect will influence students to change their group dynamics.

H6. Students will enjoy using IneqDetect.

H7. Students will prefer using IneqDetect compared to reflective writing.

5.6 Studies in the Wild

To help students reflect on and improve their group dynamics, I deployed IneqDetect in four classes in my college. These deployments help me to explore the research questions and analyze my hypotheses presented in the previous section. Across these classes, the study design was adapted to meet the individual needs of the classes and the instructors. As a result, there are some significant differences between how the study occurred in each of these classes. For example, in three of the classes students used IneqDetect multiple times to allow them to monitor their group work week-to-week. It allowed me to see the effect that IneqDetect had on students over time. In one of the classes, a different approach was taken: a different group was chosen each week. In this class, many different groups were able to use the devices and reflect on the visualizations. These groups were not able to see how their group dynamics changed over time but they were able experience the tool and provide feedback. In addition, these classes differed in size, activity types, and team structure. Some of the classes featured first-year students and the others featured entirely masters-level students. While the team sizes were the same for most of these classes, the way that they were formed and their gender and racial compositions varied widely. To account for these many differences, I avoid using inferential statistics to analyze the data. Rather, I triangulate the data by presenting multiple different observations and perspectives, and I acknowledge the nuances in the data where appropriate. This research seeks to generate new hypotheses and gain a better understanding of group dynamics and RSTs, rather than validating theories and proving causal relationships.

5.6.1 Setting and Context

The studies presented in this chapter took place in four classes, summarized in Table 5. All of these classes were offered by the College of Computing and Informatics, but participation is not limited to students who are majors within that college. Students in courses like Introduction to Computer Science I may be from business or other domains. Similarly, Human-Computer Interaction courses include students from design domains such as architecture. These classes differed in size, activity types, and team structure.

For classes to be considered for the study, the instructors needed to feature activities where students worked together in pairs, groups or teams during class. Every class in this study featured some aspect of active learning in the class. Three of the four classes in this study took place in classrooms that were specifically designed to support active learning. The active learning classrooms each had desks with wheels that can be reconfigured into makeshift tables for students to sit together. A extra desk was positioned in between the students during the study to hold the IneqDetect devices. Two of the classes had group sizes of four students, one of the classes had group sizes between 5-6 students, and the last class had group sizes that ranged from 3-5 students. A description of each of these classes is presented later in this chapter.

Table 5: An overview of the classes in which the IneqDetect study occurred.

Course	**When**	**Groups Studied**	**Group Size**	**Class Size**
Intro to Computer Science	Spring ‘18	2	6	54
Systems Integration	Spring ‘18	4	3-5	24
Intro to Game Des. & Dev.	Summer ‘18	1	4	12
Human-Computer Interaction	Summer ‘18	1	4	13

5.6.2 Introduction to Computer Science

The CS1 course is the first programming class for students in the College of Computing and Informatics. The preparedness levels of students in this course vary widely. The majority of students in this course are majors in the college. However, the course also attracts some students from business and other colleges. Some students have years of programming experience and others have never done any programming. These differences in preparedness are challenging to manage. The pace of the course can be slow and tedious for some students, while other students struggle to keep up with the material and workload. When collaborating, these differences can be difficult to manage and students with more experience are likely to take control of the conversation or do all of the work themselves. These problems can be exacerbated when competition is introduced through timed activities or by tying the activities to a significant portion of the students' grades. Groups did not have formal roles and were formed using the lightweight teams philosophy [94].

For these reasons, this class was a great candidate for using IneqDetect. I expected that the conversational inequality in this class would be very high and that IneqDetect might, therefore, be most helpful in this class. The system was completed one month into the semester, and this was the first class to use it. In the first week of the study, adaptations were made to ensure that it was more robust to variations in classroom noise. To focus on IneqDetect, only one reflective writing assignment was given to students in the class. After this first assignment, data was only collected from groups participating in the study. There were two groups that used IneqDetect in the class. The first group consisted of of five students, two of whom were women. In the second group, there were six students, three of whom were women. All students in both groups consented to all aspects of the study.

5.6.3 Systems Integration

Systems Integration (SI) is a graduate-level course that is highly technical. The class in the study allowed students to choose their own technologies to build a large software system. Students were taught about the model-view controller, application programming interfaces (APIs), and microservice architectures as ways to structure their projects. The class met twice per week. Each week, the first class was a student-led interactive workshop about a new technology, and the second class was an open period for students to work together on their projects with the two instructors moving from group to group to answer questions, debug problems, and provide guidance. Impromptu mini-lectures and discussions were common during the second class period to communicate one team's insights to the rest of the class.

IneqDetect was used in this class by multiple different teams. Students used IneqDetect during the open class session when they worked on their class projects. They did not use IneqDetect during the workshop days. The student population for this class was almost entirely international students with a roughly even gender distribution. This class was interesting because students were older and had more experience, many of the students were international. The class activities were also much more open-ended than in any of the other classes where the studies occurred. Finally, by studying multiple different groups I was able to obtain the baseline conversational equality for multiple groups. Groups did not have formal roles and students were grouped by interest and complementary ability.

This class was taught in the spring of 2018, and I was one of the two co-instructors for this course. To ensure that students were treated fairly, my co-instructor collected the consent forms on my behalf and no extra credit was given to students for participating in the study. In this class, I only conducted sporadic participant observations, because I had to provide help to multiple groups throughout the recording period. I was unable to sit with the group for the whole period as I did for the other classes in this study. Students did not complete reflective writing assignments in this course.

5.6.4 Introduction to Game Design and Development

In the Introduction to Game Design and Development (GDD) class, the instructor taught in a hybrid lecture and active learning style. In most classes, the instructor would use the first half of class for discussions, lectures, and peer instruction. In the second half of class, students would work in teams on their assignments and team projects. This was the only class in the study that featured teams with defined roles. These roles reflected the types of roles that students might hold at a game company: designer and artist, developers, and game producer. Students in the teams adopted these roles but defined what those roles meant for themselves. In most of the teams, the game producer acted as the leader and project manager. The class consisted of only 12 men. The group selected in this class consisted of four men. I believe that all of the students in the class were majors in the college.

Students used IneqDetect while working on their projects. These projects were somewhat open-ended with milestones and deliverables clearly defined. Conversations related to project management, artistic direction, program implementation, project updates, and brain-storming. Conversations were most animated when discussing artistic direction and brain-storming game ideas. In this class and in the Human-Computer Interaction class, students used IneqDetect between two and three times over a one to three week period. All students in the class also completed reflective writing assignments. Students received extra credit for all reflection activities, which was determined by the instructor. During the intervention week, students who did not use IneqDetect did reflective writing at the end of class on those days.

5.6.5 Human-Computer Interaction

The Human-Computer Interaction (HCI) course is a user-centered design course in the College of Computing and Informatics. There is not a programming component. Instead, the focus is on learning and applying design methods in the context of technology. Students gather requirements from potential users and build various low and high-fidelity prototypes to explore different design goals. Students evaluate these prototypes using a variety of evaluation techniques with real and imagined users. The activities in the class are a combination of highly-structured and open-ended. The open-ended activities are usually related to the students' group projects. Groups in this class did not have any defined roles.

In this class, IneqDetect was used for a variety of different activity types during class. These activities included user research, paper-prototyping, and heuristic evaluations. The group selected for the study consisted of four men. There were only two other groups in the class. One group did not have every member consent and the other group did not have every member present on the first day that IneqDetect was used in the class. The instructor for this class asked that students receive extra credit for each reflection activity. During the intervention weeks, students who did not use IneqDetect completed reflective writing assignments instead. These students were given the option to stay and do the reflection in class or complete it at home.

5.6.6 Recruitment

The recruitment methodology for this study was very similar to the one presented in Chapter 4. I used snowball sampling to find instructors that were teaching in the

Spring and Summer semesters. I started by asking instructors in the CCI Center for Education Innovation (CEI). Instructors were recruited from this center because many of the associated instructors use active learning in their classrooms. These instructors are also excited about innovative pedagogies and educational technology. At the time that I started recruiting for my dissertation studies, IneqDetect was still being actively developed. It was not implemented until mid-way through the Spring semester. As a result, I was only able to recruit two instructors. One of these instructors agreed to use both RSTs presented in this dissertation, IneqDetect and BloomMatrix. I was co-teaching in the Spring, and I asked my co-instructor whether he was interested in participating in the study. I recruited the last instructor from my lab. They heard about the study and told me that they were teaching in the summer for the first time. They expressed interest in trying IneqDetect.

After recruiting the instructors, I visited each class to present reflection, IneqDetect, and an overview for the study. Students were provided with a consent from that had three options. First, students could decline to participate in the study. If they did so, they still completed the reflective writing assignments. However, the data was not collected or used for research. Second, students could consent to share their reflective writing assignments and survey data, but not use IneqDetect. Finally, students could consent to share their reflective writing assignments and survey data, and also use to IneqDetect in their groups, if the group was selected. For IneqDetect to be used by a group, all of the students in that group must have provided consent to use IneqDetect. If multiple groups were eligible to use IneqDetect, a group was selected randomly with priority given to groups with all members of the group present on the first day when

IneqDetct would have been used.

5.6.7 Procedures and Data Collection

The study had three main stages: activities that happened before the intervention, the intervention, and activities that happened after the reflection. This sequence is shown in Figure 26. The first stage focused on understanding students' existing reflective practices and abilities. In this stage, students completed a survey and reflective writing assignments to get a baseline of their reflective abilities. The number of reflective writing assignments was dictated by the instructor. The second stage was an intervention which focused on scaffolding reflection to improve student awareness of their team's collaboration with the goal of improving equity within the groups. In this stage, students either used IneqDetect or completed reflective writing assignments. The reflective writing in this stage contained prompts that were intended to focus students' attention on their team's collaboration. For each week that students used the IneqDetect system, they were compensated with a $5 gift-card from Starbucks. In the third and final stage, students complete a survey, engage in more reflective writing, and have the option to be interviewed about their experiences with reflection. Students who participated in the interview receive a $15 gift-card from Starbucks.

These procedures were adapted to meet the needs of each course in the study. A variety of factors in the classes called for adaptations. For instance, the length of the courses in which the study took place varied from 16 weeks in the spring to 5 weeks in the summer. The course topics, instructor, and pedagogical techniques also varied from course to course. All of these factors led to variations in terms of how

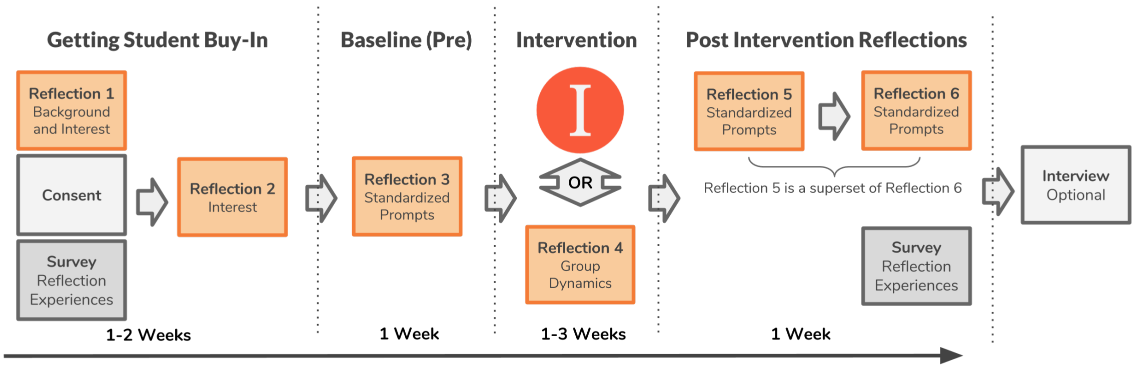

Getting Student Buy-In
Baseline (Pre)
Intervention
Post Intervention Reflections
Reflection 1 Background and Interest
Consent
Survey Reflection Experiences
Reflection 2 Interest
Reflection 3 Standardized Prompts
I
OR
Reflection 4 Group Dynamics
Reflection 5 Standardized Prompts
Reflection 6 Standardized Prompts
Reflection 5 is a superset of Reflection 6
Survey Reflection Experiences
Interview Optional
1-2 Weeks
1 Week
1-3 Weeks
1 Week

Figure 26: The schedule of how study activities were sequenced throughout the semester. Pre- and post-surveys help to show how attitudes toward reflections have changed. Baseline and post-intervention reflections are used to show how reflective writings change after using the RSTs.

much time was spent in each stage of the sequence. For the shorter summer session classes, everything was condensed to fit into a five week period. Additionally, some classes run longer than expected, which can reduce the time for reflection at the end of class or the reflection activity may be skipped altogether. At three times across the HCI and GDD classes, the learning activity took longer than expected and a few students elected to stay after to complete the reflection while others left as scheduled. These confounds would have made it difficult to conduct the studies as traditional controlled studies.

After the first two phases of this study, I made some adaptations to ensure that the study protocol above could be followed more accurately in the two summer courses. These adaptations included customized presentations to introduce reflection in a way that related to the course material. In the HCI class a short lecture was given about color theory. In this presentation, many students were surprised to learn that color perception is much more complex than many people assume. This ties into the design aspects of the HCI class, but also show that reflection can help us to consider familiar things in a new and exciting way. I also created every reflective writing activity for the instructor. I presented the activities at least three days before the class and asked for feedback before assigning it to students. I attended every class period for both classes. In the HCI class, I served as a teaching assistant, giving students design feedback and reviewing their submissions. In both classes, I provided feedback for each reflective writing assignment that they completed. Most of my feedback consisted of encouragements and pointing out something that I liked about their reflection.

The data collected in this study consists of surveys, written reflections, interviews, and participant observations. In the following subsections, I describe this data, how it was collected, and how it was analyzed.

5.6.7.1 Survey Data

I collected survey data from students at the beginning and end of the semester. This data gave an overview of how students felt about reflection and how that changed over time. In these surveys students were asked to respond freely to questions. Students were provided with statements and were asked to rate them on a 5-point Likert scale, for example, *"I find it useful to reflect on my learning process."* In the first survey, students were asked about their existing reflective practice. These questions included *"What aspect of your own life do you currently reflect on? (Religion, Fitness Apps, Classroom Experiences, Relationships, Friendships, etc)"* and other questions related to the specific reflective goals for the course. For example, in the AI course, students were asked to imagine utopian and dystopian futures for artificial intelligence.

I also collected survey data on the days that students used IneqDetect. This data included asking students to estimate how much they spoke during the conversation as a percentage. Students were asked to rate statements on a Likert scale about their experiences with reflection and IneqDetect. I analyzed the survey data using inferential and descriptive statistics. To make comparisons between the two conditions correlations were computed using Spearman's Rank Correlation test. This test was used to account for the ordinal Likert-scale data. Spearman's Rank Correlation test is "appropriate when one or both variables are skewed or ordinal and it is robust when extreme values are present" [116].

Finally students were asked to provide information about their experience using IneqDetect. I expected that not every student in the groups would volunteer to participate in the follow up survey. As a result, I asked a series of open-ended questions to get a better understanding of students' experiences using IneqDetect. This data was coded using *First Cycle* and *Second Cycle* coding methods. In the first cycle, the data was labeled using a combination of *In Vivo coding* and *Structural coding.* In the second cycle, patterns were identified from the codes by grouping them into themes. To complete this coding, I recruited an external evaluator to apply the coding scheme. The questions asked were primarily about what students liked and disliked about using the system. Therefore, having an external evaluator reduces potential biases that I might have introduced as an investigator. This also helped to provide another coding perspective that triangulates the codes that I discovered and categorized with the undergraduate student.

5.6.7.2 Individual Interview Data

Students who used IneqDetect were also offered the opportunity to talk about their experiences in a follow-up interview. This provided me with detailed accounts of the students' experiences and helped to contextualize students' experiences. Students were compensated with a $15 gift-card from Starbucks for participating in the interview. On average, the interviews lasted about 30 minutes. Six students volunteered to participate in the interviews (5 men, 1 woman). At least one student volunteered from each team in which IneqDetect was used. In the GDD and HCI classes, two students from each of the two teams volunteered. For small projects, 6-10 participants is the recommended number of people to interview [27].

To analyze this data, I partnered with an undergraduate student to code the interview data. We started by transcribing the audio recordings from the interviews into transcripts. We segmented the transcripts based on pauses. We then individually reviewed the transcripts using *First Cycle* coding methods [138]. To do this, we applied *In Vivo coding* and *Structural coding* to label the utterances. In Vivo coding uses the students' own words to label their utterances. After coding each recording, we met to discuss our codes. We kept codes that were common and we negotiated codes that did not match. If consensus was not reached, we dropped those codes. Then we recoded the transcripts. After coding all of the transcripts, we compiled themes that appeared across all of the sessions. And we pulled quotes from the transcripts that exemplified each theme. The quotes found in subsequent sections were extracted using this process.

5.6.7.3 IneqDetect Data and Equity Scores

IneqDetect records and analyzes students' conversations. From this analysis, I extracted the equality score and the total talk time for each speaker in the group. I used these measures to better understand the group dynamics within each group. I used this data to triangulate students' experiential accounts of their group dynamics. For instance, I can compare the amount that each student was detected to speak with their estimates of how much they spoke. I can also compare their perceptions of conversational equality with the amount measured by IneqDetect.

5.6.7.4 Participant Observation Data

For every class in this study, I attended all or nearly all of the class sessions throughout the semester. The only exception was the CS1 class, where I attended about half of the class sessions. When I attended classes, I took notes about students' group interactions. For groups that used IneqDetect, I typically sat near that group to make observations about their interactions. There were exceptions in the Systems Integration class and the HCI classes where I had to balance my role as participant observer with my role as a teaching assistant or as an instructor. These observations also included students' discussions about the IneqDetect visualization at the end of class. During the participant observations, I made notes about whether there was a clear leader or not. I also made observations and estimates about how equal I thought the conversation was and how much each participant spoke. I used this data to provide additional context to students' interview data.

5.6.7.5 Intervention Procedure

Each day that IneqDetect was used, students followed the sequence outlined in Figure 27. First, students engaged in a collaborative learning activity with the group members. This included activities such as collaborative quizzes, paper prototyping, heuristic evaluation, and problem solving. This was determined by the instructor and not by me. Their conversations during this learning activity were recorded using the lapel microphones and Raspberry Pi hardware devices. At the end of the learning activity, the recording ended and students participated in a short survey about their experience. This survey also asked students to estimate how much they thought they talked compared with their group members as a percentage. The purpose of this survey is to capture their perceptions before those perceptions are biased by their peers' comments and the IneqDetect visualization. Next, students view the visualization as a group and discuss the results. Students were prompted to start discussing when I say, "You have a few minutes now to review the visualization." After a few moments, if they have not already started talking as a group, I say "Do you see anything interesting?" When conversation stagnates, I follow up by saying "Are the results what you expected?" or "Do these results suggest to you that your group should make any changes going forward?" If students broached these topics on their own, I would skip that prompting question.

5.6.8 Evaluation

To evaluate the data collected in this study, I used a mixed-methods approach. This mixed-methods approach was necessary to account for the size of the study,

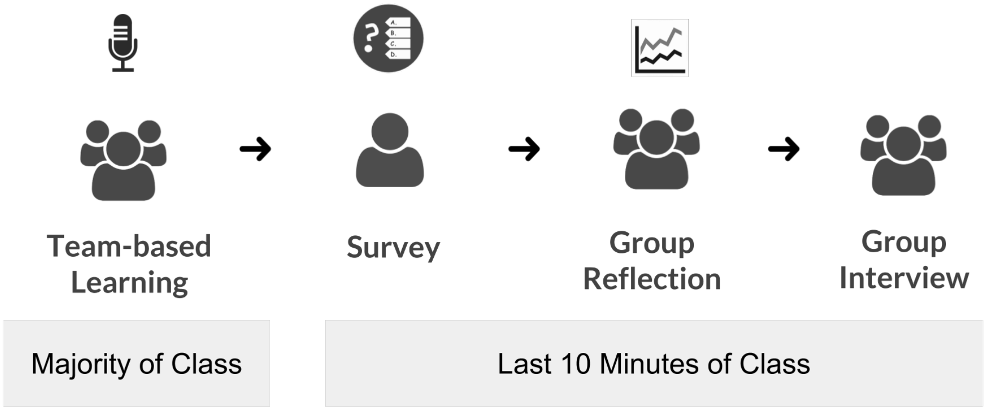


Figure 27: The sequence for the in-class portion of the study. Students collaborated on active learning activities and then at the end of the class they take a survey, review the visualization as a group, and answer questions about their reflection.

variations between the classes in which the study occurred, and to account for the real-world nature of the study. In any study about group dynamics, it is difficult to make generalizations. Accounting for and modeling personalities, communication styles, and power dynamics is very challenging. This challenge is compounded by different activity types, varied activity lengths, and distractions that inevitably occur in real-world classes.

To adapt to these challenges, I use the data to triangulate students' experiences during the study. Using multiple sources of data provided multiple perspectives and reduced biases that are inherent in any one data collection and analysis method. To reduce my own bias when analyzing the data, I coded the data with an undergraduate student and I started with the interview data and *In Vivo* coding to obtain themes that were rooted in the students own words and accounts of their experiences. Using the resulting themes from the interview data, I introduced other data sources to provide additional context and verify observations made by the students themselves. This grounded approach was used to ensure that students had a chance to convey their own stories. This also ensured that quotes were not cherry-picked from the interviews to support my hypotheses. When possible, I interweave the stories of two students within the same group to provide a more detailed account. In total, 6 students participated in the interviews.

In addition, to these qualitative aspects of the mixed-methods approach, I also used quantitative methods to analyze some aspects of the data. The survey data is presented as aggregates and some tests for significant differences between the two conditions were performed. Differences between Likert responses were computed using Spearman's Rank Correlation test.

The various in-the-wild, classroom environments made it impractical to use inferential statistics for hypothesis testing. This limits the generalizability of the results presented in this chapter. As a result, this work is exploratory in nature. The goal is to better understand the research questions that I proposed earlier in the chapter, and to potentially generate new research questions in the process. In keeping with research-through-design and design-based research, this work helps to better understand the many complex factors that affect collaboration, sociocultural inequities, and reflection support tools in real-world settings. This this process, I have captured and distilled the experiences of students in the class to develop new and evolving understandings of these complex factors.

5.7 Results

In this section, I present the results from analyzing the data collected in this study, based on the procedures outlined in the previous sections. Some of these results, such as the themes, have previously been reported in my first author paper, *"IneqDetect: a Visual Analytics System to Detect Conversational Inequality and Support Reflection during Active Learning."*

5.7.1 Triangulating the Themes from the Interviews

As, mentioned earlier, the interview data served as the foundation for this analysis. In total, 6 students volunteered for interviews (5 men, 1 woman). Survey data was also collected to provide additional information about students' experiences. 15 students responded to the survey (7 and 8 students from HCI and GDD respectively). Coding the interview data resulted in the following four themes:

Table 6: The themes that emerged when coding the interview data.

Themes
Estimating Turn-Taking and Accuracy
Students' Perceptions of Accuracy
Roles: Leaders and Non-Leaders
Motivation, Focus, and Behavioral Change

5.7.1.1 Estimating Turn-Taking and Accuracy

I expected that students would have a difficult time estimating how much they contribute to a conversation. If this hypothesis is true, then conversational inequalities might emerge in groups without students even being aware of them. Alternatively, students may be aware of the inequality and ignore or normalize those disparities. Five of the six students interviewed described being surprised by the results in the visualization. This suggests that students may have a hard time estimating how much they contributed to the conversation. GD-1 said that they were *"shocked at first. I didn't know that I talked that much!"* HCI-3 was also *"surprised ... [that] I think I* *talk a lot more than I do."* This initial surprise was followed by a reflection about how much they spoke, *"when the equality score came up I was surprised at first how low it was [for me] but then I was like that's about right because I'm never like one of the super talkative ones."* CS-1 was the only student who said that they were not surprised by the result, but they still indicated that on some days the results were unexpected. For instance, CS-1 said, *"the amount of talking within each day you know sometimes I actually thought I was going to be an average speaker but then sometimes I notice sometimes, some days I'm talking more or less."*

To better understand students' surprise, I surveyed students about how much they thought that they contributed to the conversation compared to their group members. Spearman's Rho indicated no significant correlation between their estimates and the data recorded by IneqDetect ($Rs = 0.311, n = 32$). I did not ask the participants to explain the source of the discrepancies between their expectation and the results as detected by IneqDetect. However, GD-2 provided an unsolicited explanation, *"There are instances where I thought I talked a lot, talked more than others. Essentially because of the high you get when you're talking, more so when you're leading the conversation."* Another discrepancy is related to how students remember the contributions of others. Students discussed how some types of communication, such as higher-quality contributions, were also perceived as being greater quantitatively.

5.7.1.2 Students' Perceptions of Accuracy

IneqDetect represents students' conversations visually. During the interview, students described how accurately IneqDetect captured and represented their conversations. The majority of students thought that IneqDetect was highly accurate, but three of six students interviewed indicated that it was not always accurately representing the meaningful parts of their communication. For instance, GD-2 said that IneqDetect was *"100% accurate"* at distinguishing who was speaking, but around *"30% was irrelevant"* conversation related to *"jokes and side conversations."* GD-1 stated that he was *"fairly confident"* that the results were accurate, but that when considering *"accuracy versus inaccuracy, they're talking also about appropriateness in some ways, right? I mean if it's recording you while you're talking about football, it's not accurate."* He went on to say that in case, *"it could be seen as accurate, since it's accurately recording [the voices]... but it's not an accurate representation [of collaboration]."*

Students were very confident in the results. One reason was that they could identify specific landmarks from the conversations. For instance, HCI-2 indicated this, saying *"pretty confident. That's where we all paused and we were working. That's where we were yelling about Star Wars."* He went on to say, *"Pretty accurate ... some points I couldn't place exactly."* For two of six students interviewed, these reference points helped them to navigate the results. Another aspect that made students more confident in the results was that the results were consistent across similar sessions. CSI-I explained that he thought IneqDetect was accurate *"because the result was consistent, constant results made me know."* Finally, HCI-1 stated that although the results *"matched what was going on in my head"* (their expectations) about *"60%"* of the time, they thought the results were about *"90-95% accurate"* overall.

5.7.1.3 Roles: Leaders and Non-Leaders

‘Roles’ was a theme that came up in most groups. This was surprising, because only the GDD class structured collaboration with explicitly defined roles. Despite the lack of structured group work, students in many of these groups defined and performed roles on their own. Who performed which role was not explicitly agreed upon within these groups, and there were cases where the perception of who performed which role was not uniformly agreed upon by all of the students in the group. The main role that emerged in most groups was a leadership role. Students generally described performing this role themselves or instances where other students in their groups performed this role. HCI-3 described how others students assumed leaderships roles and how those changed depending on the type of the collaboration, saying *“HCI-2 talked most on hardworking days, HCI-1 talked more joking.”* On the other hand, GD-1 described his own role as a leader and frequently referred to the roles that others held in the group. Specifically, he mentioned the designer eleven times, the artist four times, producer twice, and the software developer twice. In his interview he used designer and artist to describe the same role. His teammate did not mention any of the roles specifically during the interview.

To further analyze the leadership roles that emerged within the groups, I asked students in the survey about their perception of themselves as the leader of their own group. Students who more strongly agreed with the statement that they were leaders within their group, estimated that they spoke more within the group ($Rs = 0.32$, $p < 0.05$). Their perception of themselves as a leader was also correlated with the amount of talk time that IneqDetect detected from them ($Rs = 0.40$, $p < 0.05$). These self-perceived leaders also strongly agreed with the statement that they spoke more than they wanted to speak ($Rs = 0.55$, $p < 0.05$). I did not ask students whether they chose the leadership role or whether they preferred leadership roles.

However, the interview data provides some evidence that leadership roles were not always preferred. GD-1, a self-identified leader, described his reluctant acceptance of the leadership role that was assigned to him, *"I didn't know what was expected [of me], because this is a new class [and] I'm a cyber-security major; so, I don't know the first thing about assigning roles to make a game."* He explained that *"I am the only graduate student in the class so it [, the leadership role,] was kind of forced on me, but I never say no to leadership positions."* Consequently, he assessed his leadership in the project negatively, saying that *"as a leader, I didn't do a good job of assigning roles."* Despite his assessment, he was the strongest leader in any of the groups that I observed during my participant observations. The other leader that was interviewed was HCI-2. HCI-2 did not describe himself as a leader, but HCI-3 identified him as one of the two leaders in the group. The closest HCI-2 came to describing himself as a leader was in saying that *"[HCI-3 and I] were more focused on getting stuff done and then going about our business."* HCI-2 did not use the word leader or role at any point in the interview. His group member, HCI-3, used the word *leader* eleven times, and used the word *roles* three times. In this group, HCI-3 also indicated that he was inspired by the leadership exhibited by HCI-2. He went on to say that he wanted to take on more of a leadership role in the future.

5.7.1.4 Motivation, Focus, and Behavioral Change

There were a variety of benefits that students associated with IneqDetect. Many students described that they experienced improvements in motivation and focus. Five of the six students interviewed described ways the ways that their group dynamics changed as a result of using IneqDetect. CS-1 talked about how IneqDetect motivated him to stay focused on course materials, saying that IneqDetect *"motivated me to talk about the topic at hand."* CS-2 agreed that IneqDetect was *"keeping me more focused."* She explained that her team thought it improved their focus, saying *"a lot of people have been talking about that aspect."* She also said that IneqDetect was causing her group members to explain ideas in more detail to increase the amount of time that they contributed. She said *"the whole semester up to that point they were kind of like I'll just put C or whatever."* After using IneqDetect, *"they would have a bit more of an explanation."* She further speculated that this may have encouraged her peers to do the prep work, saying that *"one or two of them might have actually even read through the book a little."*

Motivation also came in the form of competition. GD-2 said that wearing the lapel microphones *"gave legitimacy, it made it all feel so real."* For him, using IneqDetect *"turned this into a fun activity to challenge ourselves to talk more."* While *"fun"*, he also described it as *"strangely competitive."* In a group of four men, he went on to say *"take four dudes give them all a microphone and you're going to find a competition."* Although he mostly explained the competition positively, he also described it as a potential stressor, *"it was motivational and kind of a worry."* His teammate, GD-1, did not describe any aspects of competition within the group. Instead, he described collaborative behaviors that were supported by IneqDetect. He said that based on the visualization he discovered that *"our artist/designer didn't get as much time as I would like ... [and so] after the first project, I changed gears from leadership."* After noticing that the designer was not getting as much time to talk, GD-1 made a change the following week. He said that *"As you saw in the second time we recorded ... I took a backseat."* He said that this experiment was successful and that the designer did talk more in that week. I did not make notes about this during my participant observation. However, IneqDetect showed that in the second week his contributions went down a lot and the designer's contributions went up.

Finally, there was an interesting change in group dynamics that occurred in the HCI group. At the beginning of the study, HCI-1 was the primary leader in the group. Both students described him as someone who cracked jokes and derailed the conversation. For example, HCI-2 said, *"he would like crack jokes a lot more and kind of get us off subject more but not to a point that it was like bad or annoying."* HCI-3 said that this changed after the team viewed the visualization. He said the group observed that *"HCI-2 talked most on hardworking days, HCI-1 talked more joking."* After observing this trend, HCI-3 said that *"the group like unanimously decided and HCI-2 became the main worker ... HCI-2 took [the] leadership [role]."* He said after that week that *"HCI-1 stopped talking."* Observing this shift in his group was motivational for HCI-3. He indicated that he had not previously considered himself a leader, but that he wanted to *"... take more of the role HCI-2 did."* This was especially interesting because HCI-2, the student who took on the leadership role

Table 7: An overview of the changes in conversational equality for teams that used IneqDetect Longitudinally. For 3 of the 4 teams, the equality worsened.

	Gender	First Use	Second Use	Third Use
GDD	(4M)	0.12	0.38	
CS1 (1)	(4M/2W)	0.04	0.16	0.25
HCI	(4M)	0.18	0.34	
CS1 (2)	(3M/3W)	0.6	0.28	

did not appear to notice this shift. He said over time that the team became more focused on the activities, but when specifically asked whether he observed any changes in the group, he said *"After? Honestly, no."* I was surprised that both students had such different views of the same experience. Based on my participant observation notes, HCI-2 did become involved and HCI-1 became less involved after the first time using IneqDetect. These changes were also detected by IneqDetect. After the first week, HCI-2 started speaking more and HCI-1 started speaking much less.

5.7.2 Evaluating the Conversational Inequality

To evaluate the conversational inequality detected within the groups, I extracted the data from IneqDetect. This data included the time-series data about which group member was detected speaking during each time interval, an aggregation of their speaking throughout the session, and the inequality measure that compares each of the speakers. I used this data to investigate the questions, "How equitable were the students' group conversations? Does IneqDetect improve conversational equality?"

Based on this data, the inequality detected across the four teams with low initial inequality went up (0.12 => 0.38, 0.04 => 0.16 => 0.25, and 0.18 => 0.34) and the inequality in the one team with high initial inequality went down (0.60 => 0.28). This was measured across the four teams which used IneqDetect repeatedly. This data is summarized in Table 7. Across all eight teams that used IneqDetect, the variability in conversational inequality for the first time each group used IneqDetect was relatively high ($n = 8$, $mean = 0.25$, $sd = 0.21$). In the four teams, the conversational inequality appears to have regressed to this mean in subsequent weeks, with the high inequality teams becoming more balanced and the low inequality teams becoming less balanced. One explanation can be inferred from comments made by HCI-2. He said that his groups' dynamics were good and that *"[The IneqDetect results] would be more valuable had I been in like a different type of group... [with] some guy or some girl that was really shy and never spoke up."* He suggested that easier classes do not need IneqDetect as much as in harder classes, saying "I think that we were able to joke around and have a good time and still do the work that we need to do ... but there are some classes that require a lot more focus ... if I was sitting at a C or a D ... in that scenario it [, IneqDetect,] would be very useful for saying, dude, I need to stay on subject more cause I'm going to fail Calc 2." It is also possible that an awareness of inequality is not enough to change it. Conversational equality is complex, and students may not know how to make changes to improve it.

Equality was not a theme that emerged in the interviews. Mostly, because only a few of the participants mentioned it. However, there were some instances where students mentioned equity or conversational equality. CS-2 talked about power dynamics that existed between different genders and how these dynamics, *"shuts down some of the girls, or I feel like they get interrupted, stuff like that."* She went on to say that team structure and gender representation affected these dynamics, *"if they [, the groups,] were more mixed, the girls would be more comfortable interjecting, cause*

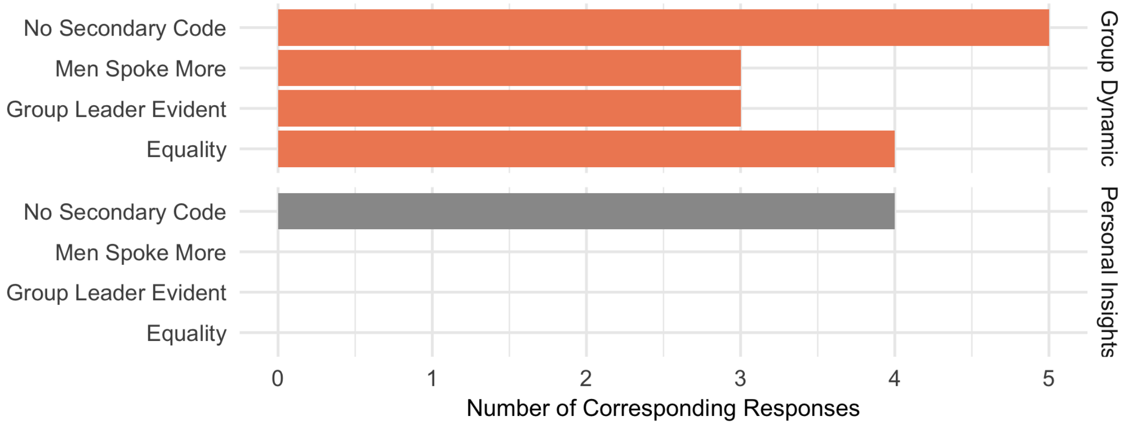


Figure 28: The secondary codes are grouped by the primary codes. These codes were extracted from students' responses about insights that they obtained from IneqDetect.

I know they [, the girls,] know it." She said the power dynamic results from *"not just the guys interrupting, [but] it's girls not feeling brave enough to speak up."* She went on to say that these aspects were observable in the visualization and that the girls were empowered to speak up more in subsequent weeks. Conversational equality was also discussed as a result of using IneqDetect, HCI-2 said *"I've noticed about myself [that] I have a tendency to just talk over people."* He said that was one insight that he had from using IneqDetect, *"there were other times where all four of us were like talking, and we were shouting over each other, which I thought was funny [, but I was] maybe not like super surprised by it."* Finally, as mentioned earlier, competitive and anti-competitive behaviors may have affected the score. GD-2, who described how IneqDetect motivated him through competition, stated that he believed *"I think everyone had this general understanding that I'm going to talk more than you guys."* And as mentioned earlier he described gendered aspects, saying *"take four dudes and give them all a microphone and you'll find a competition."*

Although equality and equity did not emerge as a theme in the interviews, they represented two of the most common themes related to insights in the surveys. These aspects, such as conversational equality and gender imbalances are presented in Figure 28. Insights about gender imbalances were mentioned by three students out of twenty-one total responses. When accounting for the fact that eight responses came from teams that consisted entirely of men, 23% of the respondents in mixed teams mentioned these gender imbalances. The three responses were that *"Guys talked more"*, *"The boys typically spoke more"*, and *"we found that the males of the group tend to speak more."* In all three cases, this gender imbalance was the only insight that those three students reported having received from IneqDetect. Additionally, students also discussed conversational equality as a main area of insight received from IneqDetect. For example, one student said that they *"tried to determine whether I participated equally in the conversation."* Other students realized that it was more or less balanced than expected. For example, one student said, *"That despite the feeling that it was equal it wasn't"*, and another student said, *"I thought it was more balanced than it was."* Other students described non-specific insights related to conversational equality. Examples included *"The visualization provided some terrific insight to speaking habits and the ability to see whether or not team members were speaking in equal amounts"* and *"It is interesting how the speaker voice levels are visualized and comparing the voice to a group."*

5.7.3 Comparing IneqDetect and Reflective Writing

In this study, all of the students completed reflective writing assignments, and one team in each class also used IneqDetect. Of the students who used IneqDetect, 71%

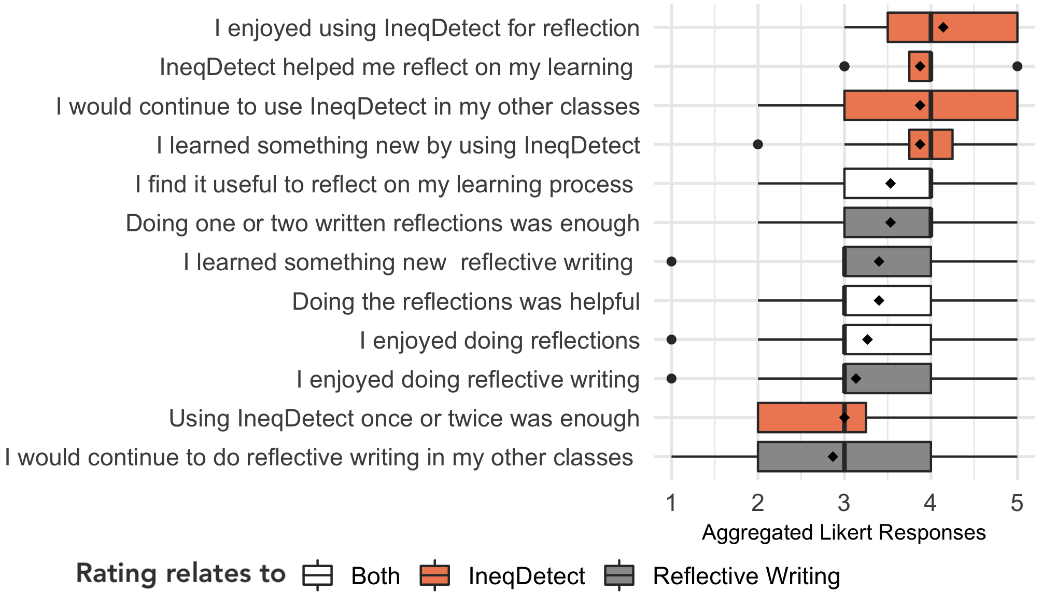


Figure 29: The Boxplots aggregate students' Likert ratings (5 = Strongly Agree) for statements about their experience and IneqDetect. Black diamonds represents the mean, the bold line is the median, and the box represents the inter-quartile range.

of them preferred it to reflective writing. Students were also asked to rate a series of statements on a 5-point Likert scale. These ratings are presented in Figure 29. Students rated IneqDetect as being more enjoyable, more helpful for reflecting on learning, and more conducive for generating insights than reflective writing. Students were also much more likely to use IneqDetect in future classes. Students generally agreed that one or two reflective writing sessions were enough, but did not indicate strongly either way whether using IneqDetect once or twice was enough. This may indicate that students received different benefits from IneqDetect each week, even within the same group. None of the differences between conditions were tested for significance given the low number of students who used IneqDetect.

Of the students who preferred reflective writing, GD-1 indicated that the structure present in reflective writing was preferred. He said that *"written reflections were a* *little better [than IneqDetect], they were specific."* He said that *"reflection specifics help"*, and IneqDetect would be improved *"if you're given a set of tasks."* CS-2 said that she liked IneqDetect and received benefits, but when asked whether reflective writing might have provided similar benefits, she agreed that *"[reflective] writing might have had a similar effect."* These comments suggest that future work should include tasks that students can complete when doing their reflections. It also suggests that it may be beneficial to add paralinguistic and prosodic features that are more difficult for students to perceive without computational assistance. However, this data may also be harder to interpret than the conversational inequality; which was described as confusing by one student in the survey responses.

5.7.4 IneqDetect as a Reflection Support Tool

IneqDetect received mostly positive results as an RST. It received higher ratings than either reflective writing or BloomMatrix. In comparison, students rated it as more being enjoyable, more helpful for supporting reflection, more conducive to learning something new, and students were more likely to use it in future classes. These results have been shown previously in Figure 29. When asked directly, 71% of students preferred it to reflective writing. Students who used IneqDetect did not use BloomMatrix, but by a transitive argument, it was strongly preferred to BloomMatrix as well. Students also indicated many instances where they obtained insights, formed their own research questions, and even changed their behaviors. All of these aspects make a compelling case for IneqDetect as an RST. However, two of six students also indicated that it was difficult to know what to do without an explicit reflective task.

It is understandable that students struggled to know how to reflect on the way that they interact with their peers. 4 of 6 students who were interviewed indicated that they did not reflect on their social interactions with others or on their group dynamics. When asked in the interviews what students reflected on, one student said grades and another said life and stressors. Only one student explicitly mentioned that they reflected on their social interactions. When asked explicitly about the way that they communicate, HCI-2 said *"usually I'm not overly mindful of what I'm saying or what I'm doing I just kind of speak my mind."* It is not surprising that IneqDetect did not therefore lead to a deep reflection for him. He said viewing the visualization resulted in *"not a huge reflection. Made me think honestly that we should focus more in class."*. GD-1 said that he sometimes reflects on collaboration. He said that he sometimes *"think[s] like am I speaking too much?"*, but added the caveat, *"so I do think of it, not as often as I should, but I do."* GD-2 said that it is a *"regular everyday occurrence of just trying to meet new people"*, but that *"there's not much reflection honestly."* He went on to say that reflecting on how you communicate can make it unnatural and difficult and for that reason he generally avoids it. Despite avoiding reflection about group dynamics, GD-2 also said that IneqDetect was a *""great tool to give feedback, the data it provided was great ... forced you to reflect on things."* Finally, it is worth considering that IneqDetect might be most useful for inequitable teams or in difficult classes. Earlier, the one highly inequitable team dramatically reduced the conversational inequality in the second week. However, that is only one data point. HCI-2 suggested that IneqDetect would be most useful in harder classes, such as *"Calc 2."* As mentioned earlier, he thought it would be a great tool for when you're *"sitting at a C or a D"* and need to convince group members to stay on task.

The most positive outcome for an RST is that it can support iterative critical reflection that leads to behavioral change. In the interview data there were some instances of this which were described in previous sections. For instance, I previously described how GD-1 used IneqDetect to identify that one group member was contributing less than others. He used this insight to experiment with his own behavior in an attempt to encourage that student to contribute more to the conversation. I also presented the example described by HCI-1 about his teammate who took on a leadership role as a result of using IneqDetect. He even described his own intentions to make changes to his career and behavior as a result. In addition to these examples, I also asked students about the questions that they explored using the IneqDetect visualization. The themes that emerged from their responses are shown in Figure 30. These themes were grouped based on whether the focus was placed primarily on answering questions about themselves or about their group. Across both focuses, many students asked questions related to how much they or their peers spoke during collaboration. This makes sense because IneqDetect provides metrics to easily answer this question. Students also frequently made comparisons between group members to understand who spoke and when they spoke. The next most common theme that emerged was trying to identify leaders within the group. These themes show that the data and visual representation likely frame the kinds of questions that students think to ask. 'Peer comparisons' was also a theme that emerged in Chapter 4. It is possible that both RSTs framed reflection around this aspect, but it is also possible that students are curious and interested in how their peers perform and behave during class.

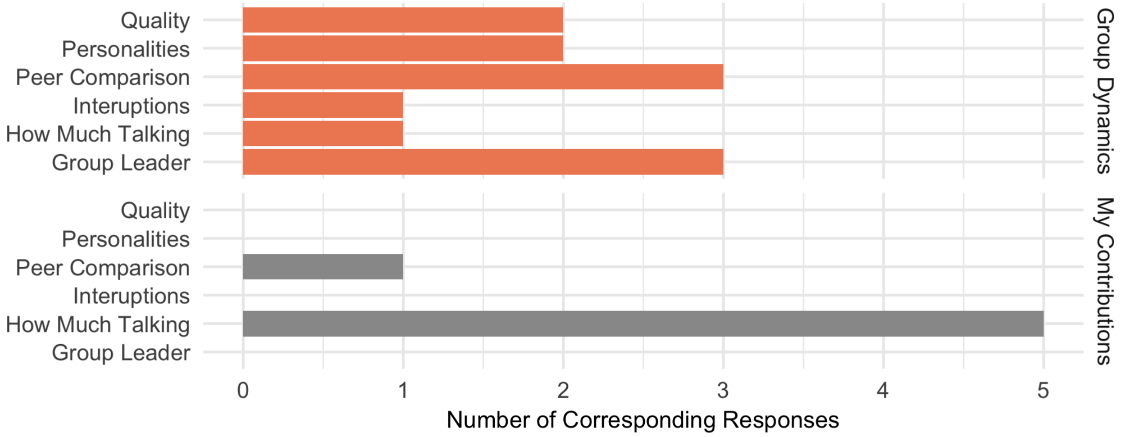


Figure 30: The themes that were distilled from students' responses about what questions they tried to answer using IneqDetect. Themes are grouped based on whether they are related to the individual or the group.

Finally, students were asked to share the aspects they liked and did not like about using IneqDetect. Students' responses from these two questions were coded and are shown in Figure 31. Students were generally positive about IneqDetect's features. They really liked the visualization and the ability to make comparisons between their group members and themselves. For example, one of the students said *"The visualization provided some terrific insight to speaking habits and the ability to see whether or not team members were speaking in equal amounts."* Like in Chapter 4, students talked again about *Ease of Use.* The second most common reason that students liked IneqDetect was that it was *"easy to use."* The main two things that students disliked about IneqDetect were the wires and the microphones. These aspects also relate to *Ease of Use.* One student felt *"stuck"* and another mentioned that it was *"difficult to move."* Similarly, multiple comments about how students liked the visualization and making peer comparisons talked about instant insights or being able to see the group dynamics directly. These aspects also speak to ease of use.

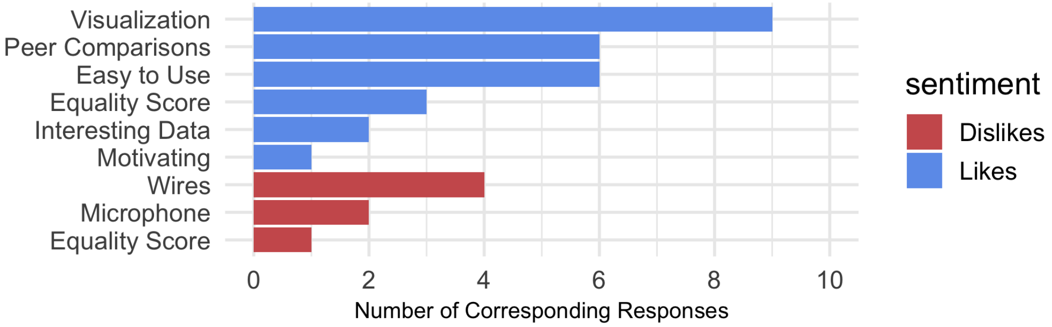


Figure 31: Students were asked to describe the things that the liked and did not like about IneqDetect. The themes that emerged are grouped by sentiment.

This reiterates the idea that students want something that can generate meaningful insights, while minimizing effort.

5.7.4.1 Suggested Improvements for IneqDetect

Students were asked to provide suggestions for ways to improve IneqDetect going forward. Students provided a few suggestions that have been summarized in Figure 32. The most common suggestion from students was to make the microphones wireless so that they could move around the classroom. In the previous subsection, the number one thing that students disliked about IneqDetect were the wires. Only one student who provided this feedback elaborated on why the microphones should be wireless, saying *"... because it would allow for discussion to be recorded while writing on the whiteboard."* In that group, there was a week where they were asked to work out problems on the white board. The student who wrote on the whiteboard removed his microphone for most of that class period.

Students also suggested adding new features to the existing system. One of these features was to detect the topics of the conversations automatically and visualize them. This is a feature that is currently in progress. Students also indicated that it

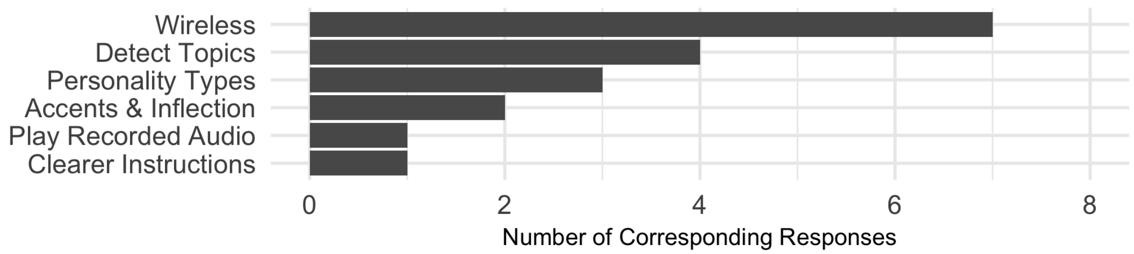


Figure 32: Themes related to students suggestions for improving IneqDetect.

should capture or respond to different personality types. For instance, one student said *"I think the system does a great job at capturing relevant data and displaying it. It may not take into consideration the natural personality of a participant. For example, if someone is quiet-natured (introverted), they will not speak as much."* It is possible that IneqDetect might be personalized to different types of speakers. Introverts could receive subtle encouragement or suggestions for how to get more involved in the conversation. Extroverts could receive suggestions about how to facilitate conversations with others and how to identify and include the voice of others.

5.8 Discussion and Summary

IneqDetect was designed and deployed to improve students' group dynamics. I expected that presenting students with visual representations of their group dynamics would allow them to analyze and improve these dynamics on their own. I generated some research questions to explore these aspects and formed hypotheses based on my expectations. The hypotheses are presented again below:

H1. Students will not be able to accurately estimate how much they speak.

H2. IneqDetect will improve conversational equality in groups.

H3. Students will experiment with their group dynamics using IneqDetect.

H4. Students will obtain insights about their group dynamics.

H5. IneqDetect will influence students to change their group dynamics.

H6. Students will enjoy using IneqDetect.

H7. Students will prefer using IneqDetect compared to reflective writing.

To evaluate the first hypothesis, I asked students to estimate how much they thought they spoke and compared those results to the amount that IneqDetect detected that they spoke. There was no correlation between these two variables. This suggests that students are generally not able to estimate their contributions to the conversation quantitatively. This aligns with my observations in my previous studies of student groups in active learning classrooms. It also aligns with the participant observations that I made in class. Some students indicated that they were aware of their inability to accurately estimate how much they contributed to group conversations, and that their perception might not match reality. Students suggested that higher quality contributions appeared to constitute more talk time and that they had different perceptions of time when they were speaking compared to when they were listening. Students also ascribed different value to different aspects of the conversations. For instance, in a single group, one student said that silence was when the work was happening, and his group member equated more conversation to more productivity. Understanding students' perceptions of collaboration is an interesting area for future work.

Across all of the teams that used IneqDetect the variability in conversational inequality detected by the IneqDetect system was high. This was measured based on the first time the team used IneqDetect. For the groups that used IneqDetect multiple times, I compared the change in conversational inequality that occurred after the first use. In teams with low conversational inequality, it got worse. In teams with high conversational inequality, it got better. As mentioned by one student, IneqDetect may be best suited to help groups with high inequality or for difficult classes. In easier classes, conversational inequality may have less obvious repercussions because bad team dynamics do not have a big impact on students grades. This result was unexpected and goes against my second hypothesis that inequality would decrease after using IneqDetect. My hypothesis was based on the idea that if students reflect on their groups inequality, they would improve the inequality through positive reactivity [117]. However, identifying inequality and addressing it are two different things. It is possible that the increased awareness and behavioral changes described by participants did not specifically affect conversational inequality. More work is needed to understand what types of behaviors are conducive to support equitable social interactions in the classroom. To improve conversational equality, students may need explicit guidance about how to change their behavior in groups. It is also important to note that only four groups were studied to evaluate this hypothesis. More teams need to be studied longitudinally to corroborate this observation.

While conversational equality did not improve for most groups in the study, students described many benefits that they ascribed to IneqDetect. Students indicated that they were more motivated and had a renewed focus on course topics. They also changed or intended to change their behaviors. The changes included wanting to speak more or less, take on leadership roles, or be more inclusive of other group members. Students also provided some instances where they explored questions about their group dynamics, including one example where a student created their own experiment to get a team member more involved, and he evaluated it using the data presented by IneqDetect. IneqDetect led to changes in leadership roles within one group. The questions that students explored using IneqDetect were primarily about their group dynamics. However, students also explored insights about themselves as well. These findings provide a lot of evidence to support hypotheses three, four, and five. Students also described many insights about themselves, others, and societal issues such as gender inequality that they identified when reviewing the visualizations. Students were generally surprised by the results presented in the visualization, but also had a lot of confidence in those results. Discussions with the students about accuracy resulted in questions about what types of metrics should be included when evaluating group collaborations.

One interesting aspect was that the observations made by students were not always shared by group members. Even in the same group, students had drastically different perceptions of the same shared experience. This was interesting because students discussed the results in the visualization as a group at the end of each class period. It was surprising that this shared representation didn't always result in consensus about the group dynamics. This leads to many new research questions about shared representations of group behaviors, about how to support group sense-making, and about the possibility of introducing reflective consensus building activities.

Finally, I observed mostly positive results related to IneqDetect's effectiveness as an RST. Students used IneqDetect to reflect on many different aspects of their collaboration and to answer multiple diverse questions. IneqDetect led to behavioral change and to critical, and in at least two cases, potentially life changing insights. For example, one student questioned his approach to group work and indicated that he wanted to take on more leadership roles. He reflected on his current work ethic, conversational style, and then related those aspects to his future career. This was compelling evidence for IneqDetect as an RST. However, some of the students also reported that IneqDetect didn't lead to deep reflections for them. They described the results as interesting or useful, but did not have the same kinds of profound insights. In particular, a three of six students struggled to understand what conversational inequality meant within their groups or questioned the value of conversational equality. Students had trouble knowing what patterns to look for or how to change their behaviors in response to those patterns. Students suggested that additional information and specific tasks might help them in these regards.

In this chapter, I presented IneqDetect and a detailed account of students' experiences using the system across four classes and eight groups. Students in these classes enjoyed using IneqDetect more than reflective writing. IneqDetect provided students with many benefits and identified many interesting insights about themselves and others. The insights gleaned from IneqDetect led to changes within the groups, and personal changes for the students. However, these changes did not include improvements to conversational equality. The studies introduced new avenues for research, such as understanding students' perceptions of collaboration, understanding how students engage in collaborative reflection, and how to help students negotiate conflicting perceptions of a shared experience. Finally, I obtained additional evidence that students highly value *Ease of Use* across RSTs. Consequently, developing effective RSTs appears to require providing students with the most insights for lowest investment of time and energy. Supporting collaboration and reflection are difficult. Each students' experience is different, and it appears that is true even within a single team. Going forward, it is necessary to create tools that support both group-level and personal insights, that minimize the effort expended by students, and also provide explicit structure and recommendations to help students identify trends and enact meaningful changes to their group dynamics.

5.8.1 Limitations

Given the number of participants who used IneqDetect more than once, I captured a rich detailed account of their experiences. However, this makes generalizing the findings from this chapter difficult. Furthermore, generalizing social theory is always difficult, because team formation, situation, and task each have a strong affect on collaboration [111]. I deployed IneqDetect in four CS classes, but did not change the structure of those classes or the existing learning activities that students engage in. This resulted in many different types of collaborative experiences. Some classes featured unstructured groups, one class had a team with defined roles, and the composition of the teams consisted of many different ages, genders, and backgrounds. This in-the-wild approach was adopted to capture the many varied ways that collaboration can happen in real classes. As a result, this chapter generates new knowledge and avenues for research. But it may not generalize to every classroom experience. Findings need to be replicated in controlled lab studies and in massive classroom deployments.

CHAPTER 6: STUDENTS' REFLECTIVE WRITING ASSIGNMENTS

In Chapters 4 and 5, I presented two reflection support tools (RSTs) that helped to scaffold students' reflections. They captured and visualized aspects of students' learning experiences to help students identify insights and improve their learning. In those chapters, students identified insights and made changes to their study habits and group dynamics. The tools were evaluated based students' preferences and on the RSTs' abilities to help students identify insights.

In this chapter, I analyzed students' reflective writing assignments to see what effect the RSTs had on students' reflective practices. Students completed 1653 reflections across the four classes that used the RSTs longitudinally. From these reflections, reflective features were extracted and analyzed. These features included the depth of reflection, the topic of reflection, whether students' attention was focused internally or externally, and the sentiment of the reflection. Reflections were also coded for instances of agency, awareness, and insights. These coded features that were extracted from the reflections were compared longitudinally across three conditions: BloomMatrix, IneqDetect, and reflective writing only. The students in the *reflective writing only* condition were in classes where IneqDetect was used, but they were not randomly selected to use IneqDetect. This chapter presents the results from these analyses and a discussion about the impact of RSTs, reflective practice, and study designs for evaluating RSTs.

6.1 Motivation

The FDR New DEAL Model presented in Chapter 3 describes how multiple RSTs can be combined to frame reflection along multiple dimensions. In the model each RST frames reflection around a specific aspect, such as cognition or group dynamics. Insights from an RST about group dynamics might prompt reflection about cognition or other aspects using a second RST. Using the model results in an ecology of multiple RSTs that frame reflection along multiple different aspects. This model is similar to the "Learning Loop Complex" devised by Russell et al. [137] and adapted by Pirolli and Card in their "Notional Model of Sensemaking" [128], presented in Figure 33. In that model, analysts oscillate between foraging for information and making sense of that information. Through this process, the analyst develops a better understanding of the problem, data, and solution. Similarly, the New DEAL Model supports iteratively foraging for insights using RSTs and then making sense of those insights.

In previous chapters, the RSTs were evaluated for their abilities to generate meaningful insights, and students described instances where these RSTs accomplished that goal. Those in-the-moment insights were valuable to students, but it is not clear how these tools effected students' reflective practices. Did students continue to reflect on their group dynamics and cognition even after they stopped using the tools? Furthermore, did those tools frame students' reflection about aspects of their learning such as their cognition and group dynamics in subsequent reflections?

In this chapter, I evaluate the tools' impacts on students' reflective practices as measured by changes to their reflective writing. Reflective writings were evaluated

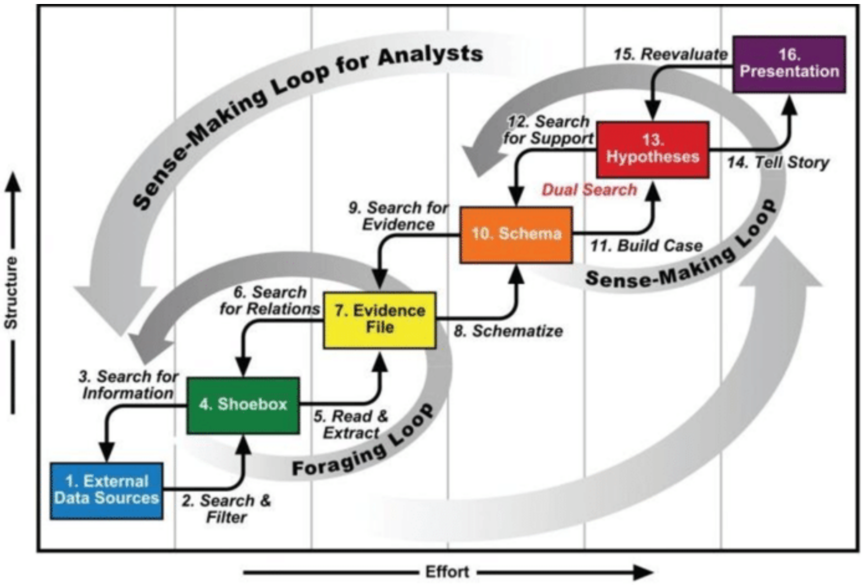


Figure 33: Sensemaking Loop for Evidence-based Data Exploration [128].

for evidence of agency, awareness, and insights. I also look at the reflection quality as measured by a stage-based model of reflection, the sentiment, and focus of students' attention. Collecting and analyzing this information helps to improve our understanding of RSTs and the long-term effects of using RSTs in the classroom. Supporting these features that are derived from students' reflective writing is important because the goal for this research is not only to provide in-the-moment support for reflection, but to help students to build long-term reflective practices and engage in reflective thinking in every aspect of their lives. Based on these aspects, I establish the following research questions:

R1. **Baseline** How do students reflect on their learning without intervention?

R2. **Prompts** How do reflective prompts frame and affect students' reflections?

R3. **Framing and Changes** How do RSTs frame students' reflections along each dimension (e.g.: topic, awareness, agency, insights, sentiment, and depth)?

6.2 Hypotheses

To address the research questions, I plan to evaluate the hypotheses presented in Table 8. That table contains my expectations for the first reflective activity (baseline reflections), and the changes after students use IneqDetect, BloomMatrix, or complete reflective writing without an intervention. The students who completed the reflective writing without an intervention were in the IneqDetect class, but did not use IneqDetect.

Table 8: An overview of the expected values for each reflection feature. BloomMatrix, IneqDetect (Reflective Writing), and IneqDetect (Intervention) are the three conditions in the study. Baseline represents the first reflection.

Condition	**Topic**	**Depth**	**Insight**	**Aware**	**Agency**	**Affect**	**Focus**
Ineq-Ref	Concepts	Med	Low	Low	Low	Neutral	Both
Ineq-Int	Group	High	High	Med	High	Positive	External
Bloom	Cognition	Med	Med	High	Med	Positive	Internal
Baseline	Concepts	Low	Low	Low	Low	Neutral	Both

6.3 Methodology

The data presented in this chapter was collected in classes where students used IneqDetect or BloomMatrix. The classroom contexts, descriptions of participants, and the study procedures were presented in Chapters 4 and 5. In those classes, I also designed and implemented reflective writing assignments, often in collaboration with the instructors of those classes. The reflective writing assignments were completed during class time or after class as specified by the instructor. Students completed reflections at multiple points throughout the semester. These varied slightly by class and condition. However, common reflective prompts were used before and after the interventions in an attempt to isolate the effect of the RSTs, and reduce the effect

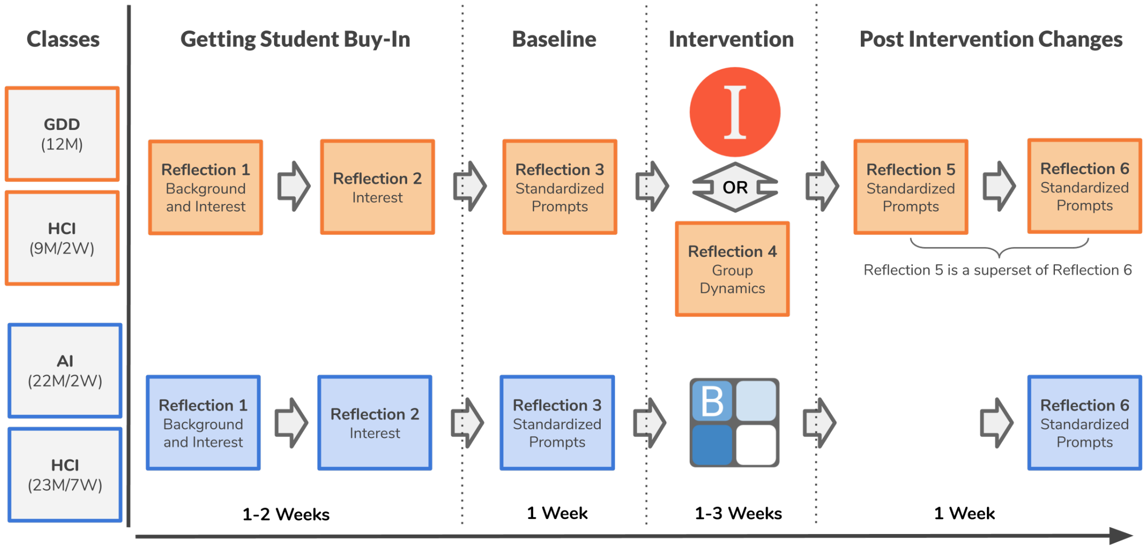

Classes
Getting Student Buy-In
Baseline
Intervention
Post Intervention Changes
GDD (12M)
HCI (9M/2W)
AI (22M/2W)
HCI (23M/7W)
Reflection 1 Background and Interest
Reflection 2 Interest
Reflection 3 Standardized Prompts
I
OR
Reflection 4 Group Dynamics
Reflection 5 Standardized Prompts
Reflection 6 Standardized Prompts
Reflection 5 is a superset of Reflection 6
Reflection 1 Background and Interest
Reflection 2 Interest
Reflection 3 Standardized Prompts
B
Reflection 6 Standardized Prompts
1-2 Weeks
1 Week
1-3 Weeks
1 Week

Figure 34: The schedule for reflective writing assignments broken down by study. Classes using IneqDetect are on top, classes using BloomMatrix on the bottom.

that prompts might have on students' reflections. An overview of when students engaged in reflective writing assignments is outlined in Figure 34.

To understand students' reflective experiences and the impacts that RSTs had on students' reflective practices, I coded their reflections using manual and automated methods. This resulted in a vector of reflective features for each reflection. These features included agency, awareness, insights, depth of reflection, the topic, and direction. By analyzing these features, I am able to establish baselines for reflection, see how students' reflections change as a result of the interventions, and determine how the reflective prompts affect students' reflection.

To establish a baseline and to encourage student buy-in, the first two reflection activities were designed around the course material. The intention was to get students to have some early successes with reflection and become engaged with the material. These baseline reflection prompts included asking students to imagine design futures, considering the impacts of technology, and reflecting on their career goals and the goals of the class. Some of these prompts also attempted to tie students' reflections into the course material and eventual intervention. In the AI class, which used BloomMatrix to reflect on cognition, students were asked about how their brain is different from that of an AI agent. In the IneqDetect classes, students were asked to speculate about what roles they usually take in a group, and why they fulfill these roles. These prompts were designed to get students to buy-in to reflection and to see reflection as something that is valuable for them.

To make comparisons between the RSTs, and to understand the effect that these tools had on students' reflective practices, I used a pre- and post- study design. Students completed a series of reflections before and after the interventions and the reflective prompts for these reflections were the same across all the conditions. This is a novel approach to evaluate reflection support tools. Existing approaches to evaluating reflection often ask participants to reflect on their experience with an RST or personal informatics system, and then they only evaluate those summative responses. My approach includes the existing reflective practices of individual students and controls for them in the model. I used a multivariate mixed-effects model and fit the data using maximum likelihood estimation. The fixed effects were the reflection condition and the repeated measure of time. The random effects were the students nested within their classes.

Finally, to evaluate the effect that the prompts had on students, I compared the prompts for each of the different features of reflection (e.g.: awareness, agency, depth, etc). I made these comparisons using boxplots and visual comparisons, and I also com-

Table 9: Coding Scheme for Evaluating Reflection

Dimension	**Values**
Attention	Internal, External
Topic	Social, Cognitive, Concepts
Direction	Retrospective, Current, Prospective
Depth	Description, Explanation, Question, Transformation, Critique
Sentiment	A continuous value [Negative (-1) to Positive (+1)]
Awareness	5-point scale [Limited Awareness (1) to Highly Aware (5)]
Insights	5-point scale [Limited (1) to Profound (5)]
Agency	5-point scale [Low (1) to High (5)]

pared across the different prompt themes. To obtain the prompt themes, I grouped the prompts thematically and created a new meta-level prompt that is a generalization of all the prompts within that theme.

6.3.1 Adopting a Coding Scheme to Evaluate Reflection

After reviewing the different evaluation methods for reflection, presented in Chapter 2, I chose to adopt and supplement the Fleck and Fitzpatrick model [58]. Justification for that decision was presented in Chatper 2. To triangulate the different aspects of reflection, I coded the data based on the *Attention*, *Topic*, *Depth*, *Agency*, *Awareness*, and *Direction* as they are expressed in the reflective writing activities that students completed. These dimensions, along with their possible values, are presented in Table 9. The sentiment and word count were also extracted from the text automatically. I used sentiment analysis based on dictionary lookup, while accounting for valence shifters, such as negators, amplifiers, de-amplifiers, and adversative conjunctions [132, 149]. Finally, I also analyzed the data for evidence of insights identified by students, changes that students intended to make in the future, and behavioral changes that they noticed in themselves and others as a result of using the RSTs.

My coding scheme for evaluating each of the dimensions is outlined below:

- **Attention (Focus)** Inspired by Rotter's *locus of control* [136, 97], this code determines whether the student is reflecting on aspects that are internal or external to themselves. As an example, "I talked much more than my team members, I should try to get their feedback too" would be coded as internal. Observations that reference other people or insights and changes that students ascribe to others are coded as external. For instance, "I talked much more than my team members, they need to stop being so quiet during team activities."
- **Topic** Each RST presented in this dissertation is intended to scaffold a specific kind of reflection. BloomMatrix encourages students to think about their cognition. IneqDetect cues students to consider their turn-taking behaviors and the social interactions that occur within their team. The possible values for this dissertation include social interactions, cognition, or course topics. This code is intended to determine whether and to what extent the RSTs frame reflection.
- **Depth (Quality)** Reflection depth is a measure of the type and complexity of the reflection. It can be used as a proxy for the quality of reflection [14], and is traditionally measured using stage-based models. I adopt Fleck and Fitzpatrick's model [58] because of its widespread use [35, 14]. I code the reflection data line-by-line along the five levels outlined by the Fleck and Fitzpatrick's model. Areas that do not contain any evidence of reflection are left blank.
- **Awareness** It is possible that students may be narrowly focused on one aspect of their learning but have deep and critical insights about that narrow focus. Awareness is a measure of how many distinct things they mention in their

reflections. Discussing more aspects and making more connections results in a high value on this 5-point scale.

- **Insight** Students can identify insights that are very specific but also very valuable for them. In these cases, their awareness might be very low because the insight is only related to a specific context, but within that context the insight might be transformative. Furthermore, change and intention to change are not necessarily required for the insight to be valid. Identifying an important insight results in a five on this 5-point scale. A rating of 1 indicated that students blamed other students or the instructor.
- **Agency** Inspired by Paulo Freire's work about how reflective action can be a tool for liberation and the many connections between behavioral change and reflection. Agency is measured on a 5-point scale between low-agency (1) and high-agency (5). It refers to instances where they describe actions or plans that they or others have taken to affect the learning process.
- **Direction** The direction of reflection refers to whether the students are focusing on past behaviors and experiences, current observations and insights, or future plans and intentions. Each reflection is coded as retrospective, current, prospective, or left blank if none of these classifications apply. Retrospective indicates that students are focusing on the past. Prospective indicates that students are imagining or speculating about the future.
- **Sentiment (Affect)** To understand students' affective states, I also extracted the sentiment from their writings. Sentiment was measured as a continuum between positive (+1) and negative (-1) poles. Zero represents neutral.

I used these coding schemes to triangulate the multiple aspects of reflection that exists in students' writings. This was important because reflection is not one dimensional and the interactions between these many dimensions may provide insight into new methodologies for evaluating reflection and may shed more light on the nature of reflection. This new understanding can inform the design of future RSTs to ensure that students have tools that provide a holistic view of their learning.

6.3.1.1 Inter-rater Reliability

To understand how robust the coding schemes were to apply, I recruited an external coder, and computed the inter-rater reliability (IRR) based on the overlap between their codes and my own. Unlike percent agreement, IRR accounts for the number of classes, the balance of the classes, and the distance between classes. To compute the IRR, I used Cohen's weighted Kappa (κ) with square weights to account for the ordinal data [36, 59]. The data meets the five assumptions for Cohen's Kappa: the data was ordinal and mutually exclusive. The raters were independent and rated overlapping data. A total of 50 reflections were coded by both raters, and the results from computing Cohen's Kappa on the reflective features is shown in Table 10. According to the guidelines for interpreting Kappa values [93], < 0 is poor, $0.00 - 0.20$ is slight, $0.21 - 0.40$ is fair, $0.41 - 0.6$ is moderate, $0.61 - 0.80$ is substantial, and $0.81 - 1.00$ is almost perfect. Based on these guidelines, the agreement between the coders was mostly moderate, with substantial agreement for depth and focus.

It is important to note that Cohen's Kappa accounts for possible agreement that might occur by chance, and reduces the observed agreement to account for that

Table 10: Inter-rater reliability results between two coders. Kappa values were obtained using Cohen's Kappa and ratings indicates the number of paired ratings. Weighted Kappa was computed for ordinal codes.

Measure	**Depth**	**Insights**	**Aware.**	**Agency**	**Topic**	**Direction**	**Focus**
Kappa	0.666	0.581	0.597	0.473	0.624	0.445	0.691
Agreement	68%	68%	66%	50%	83.7%	75.6%	84.6%
Ratings	50	50	50	50	49	41	39

chance. Therefore, the number of overlapping codes between coders can affect the Kappa value. The two codes with the lowest agreement were agency and direction. Agency was challenging to code for because some reflections contained instances of helplessness with indications agency and intentions to make changes. Direction was also hard to code because the prompts were often leading. Students often matched their response to the tense of the prompt. But in some cases they answered a prompt about the future in present tense. In those cases, it appears that students are talking about future aspects, while using present tense. For these reasons, these two codes were slightly lower in agreement than the others.

6.4 Results

The results are presented below in three sections. First, I evaluate the prompts and whether and how they effect reflection. Then, I present a temporal analysis of how students' reflective writing changed throughout the semester. Finally, I present the results from the between and within subjects study.

In total, 2206 were reflections submitted across the four classes, 553 of these reflections were blank. The remaining 1653 reflections were coded by two coders. The completion rate for reflections was 74.9%. An overview of the reflections collected by class and condition is shown in Figure 35. Reflections from the two 1212 classes were

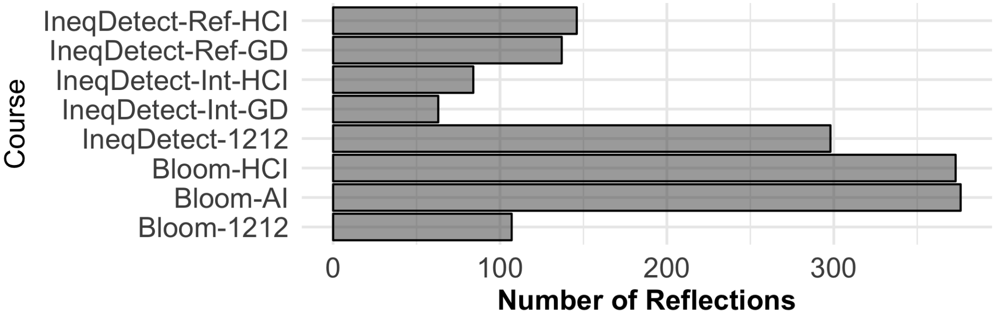


Figure 35: An overview of the reflections completed in each class. Reflections collected in the two 1212 classes were not used for analysis.

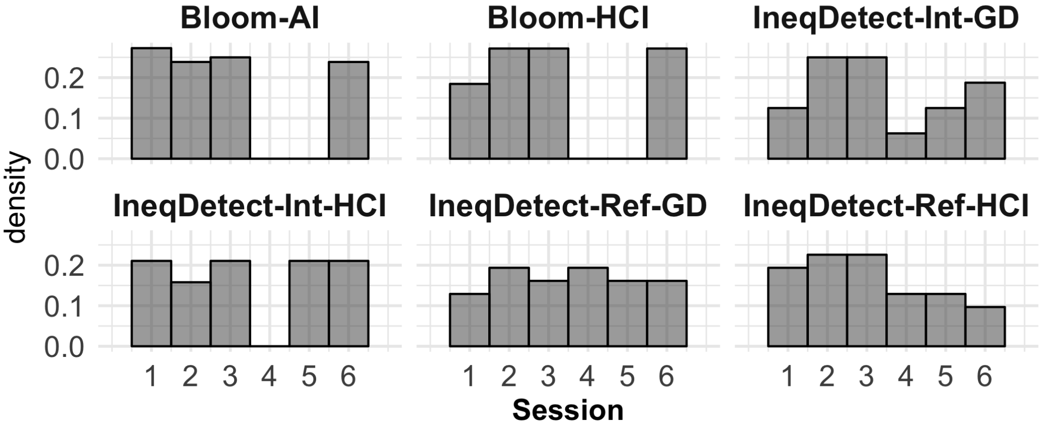


Figure 36: Density plots of the reflections in each class. They show what percentage of total class reflections were submitted in that week. Changes from week to week indicate more or less response. Gaps indicate areas when students used the RSTs.

not used for the analysis because they did not follow the complete study protocol. The remaining 1169 reflections were used for the analysis. An overview of when each reflection was asked in each class is shown in the Appendix in Tables 12, 13, 14.

6.4.1 The Framing Effects of Reflective Prompts

The goal of the first analysis was to understand the effect that the reflective prompts had on students' responses. I expected that the way that prompts are phrased would prime and bias students' reflections. I expected that prompts would have an effect on the length of students' reflections, the sentiment of their reflection, and on the coded

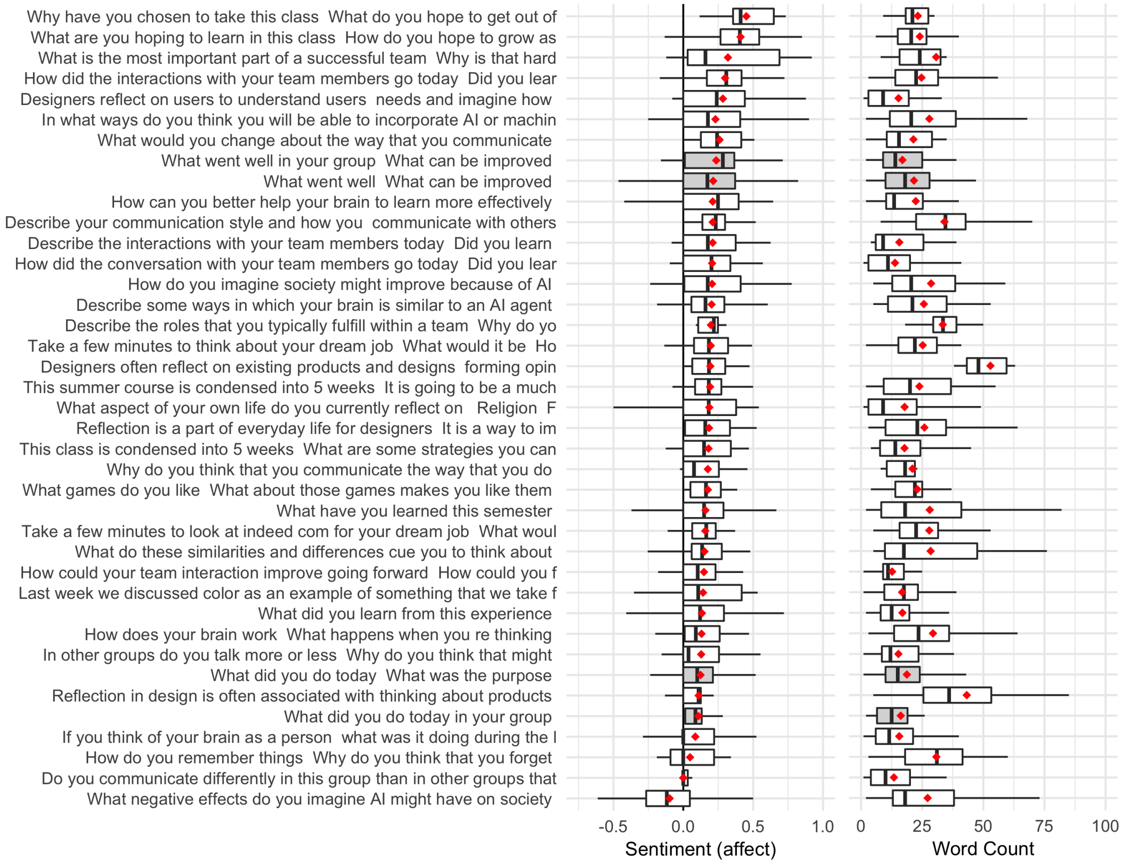


Figure 37: Boxplots for sentiment and word count split by reflection prompt. The prompts were sorted by average sentiment. The gray boxplots indicate the common questions that were asked both before and after the intervention.

reflection features. To analyze these effects, I visualized the reflection features for each of the reflection prompts. In this case, visualizations are preferred to statistical approaches for multiple reasons. First, the prompts were not always given at the same time or to the same students. Second, depending on the reflection feature, the scales of the data vary and they contain a mix of continuous, discrete, and non-ordinal categorical values. Converting categorical values to a numeric scale is not appropriate when the categories are non-ordinal and distance between categories is not known.

Sentiment and word count were the two reflection features that were extracted automatically from the students' reflections. These features are plotted for each reflection prompt in Figures 37 and 38. In Figure 37, the boxplots are sorted by sentiment and there does not appear to be any visual correlation between the two reflection features. From this graph, it appears that sentiment is affected by the reflection prompt. It is reasonable that a prompt with a negative affective prime, such as *"What negative effects do you imagine AI might have on society?"*, would lead to negative responses from students. In general, across reflective prompts the responses were generally skewed toward positive responses, but prompts about why students are taking the class, their hopes for the future, and how they can improve elicited the most positive sentiment. It is also important to note that neutral sentiment in this context can include equally balanced positive and negative sentiment. It does not necessarily mean that the reflection was devoid of sentiment.

In Figure 38, the reflection prompts are ordered by word count. What is interesting about these results is that the prompts that elicited the lowest word counts were designed to improve students' metacognition. These prompts that elicited low word counts were also about students' experiences in the class related to specifics about their learning, in-class experience, or group dynamics. Students wrote more prolifically in response to reflective prompts about their opinions. Prompts that elicited the highest response from students asked them to reflect on their future careers, their opinions about themselves and concepts, and to speculate about the course material, themselves, or the future. This may suggest that students are more responsive to open-ended reflections about provocative concepts rather than highly specific reflections about their experiences in the class. When reviewing the word count of students responses, there were some notable outliers. These students were really engaged by the reflective prompt, but typically due to their interest in the topic. Two examples selected from the five longest reflections are shown below:

- How does your brain work? What happens when you're thinking? **363 Words** *"Combining all the processes that we are 'unaware' of, how much actual info is the brain processing at a moment? Is there really no connection between 'trivial' information such as blood being pumped to certain areas and how the brain process relevant information such as whether one should eat dinner now or later? Given my background in Psychology, the base answer is that neurons transmit electrochemical signals to other neurons, and it seems to be that the various branching properties of neurons produce thought and memory. Furthermore, more used connections somehow become 'stronger' while less used connections 'weaken.' It is clear that the brain has a short-term memory (that holds about seven discrete pieces of information, on average), and a long-term memory. On a slightly more metaphysical level, there are two primary schools of philosophical thought into the realm of human cognition. The first of these schools utilizes the construct of a higher 'mind' that exists as the location of human cognition. This school of thought can trace its origins to Plato and Aristotle (Plato to a greater extent than Aristotle), and posits that the human mind holds an internal, complete representation of its environment. Thought and planning are achieved through manipulating this internal environment. More modern perceptual-psychology has largely overridden this concept of a 'mind' within the human brain. Milner, Gooddale, Noe, and other cognitive scientists* *have begun conducting experiments that are beginning to show that thought and environmental interaction are much more closely linked to an individual's physical environment and complete body-system than on an abstract 'mind' found somewhere in the brain. For example, to drink from a cup, the brain does not recognize, 'That is a cup; therefore, to drink from it a follow a long table of actions.' Instead, the mind-body system likely assigns certain 'affordances' to objects, such as the cup affords holding liquid, it affords being picked up, thrown, etc. The simpathetic nervous system is responsible for implementing these affordances and the physical body systems (musculature, skeleton, etc.) carries them out. Equating all of this to AI and computer science could mean the difference between logical if-else statements and the implementation of heuristics, reflex agents, and percepts."*

- What games do you like and what about those games makes you like them? **150 Words** *"All 3 games in the Dark Souls trilogy are my favorite games. I've played RPGs a lot in my gaming life, and Dark Souls is the greatest of all. I feel that it not only tests you as a player of the game itself, but as a gamer in general. Dark Souls introduces you to a highly advanced and extremely difficult environment that's impossible to get used to without a full understanding of how your character works and how to handle everything. The series itself is also incredibly well-done as a collective storyline ... Dark Souls 3 is my favorite of the three because of its combat system, bosses, lore and music – so much so that I sometimes read Dark Souls lore online and listen to the game's soundtrack* *in my spare time!"*

These reflections demonstrate students' passions about a topic. The prompt gave them an outlet to express that interest and to share their passion. What is interesting about these reflections is that in both cases they appear to be written for the reader, not for the writer. These reflections contain words and phrases like *you* and *for example* which suggest a dialog. Furthermore, students provide additional background information, such as their degree in Psychology or their previous gaming experience, to provide context for the reader. For some students, they may see reflection as a way to communicate with the instructor in a way that is not currently supported in typically classroom formats. For these students, instructors' feedback about the reflection would likely be very motivating. For other students, this aspect may be less important. Automatically identifying students who are looking for feedback based on their word use would be a scalable way to provide those students with what they need. It is interesting to consider what the other students would want to receive from reflections.

I also coded students' responses using a multi-dimensional coding scheme that included depth, awareness, insights, and agency. All of these four reflective features were coded on a 1 to 5 ordinal scale. The reflective prompts are shown with these features in Figure 39. Based on the graph, it is clear that it is challenging to encourage deep reflections, but that the reflection prompts do appear to affect the depth of reflection. Prompts that were specific about experiences in the classroom led to shorter reflections, and also more shallow reflections. It makes sense that word count

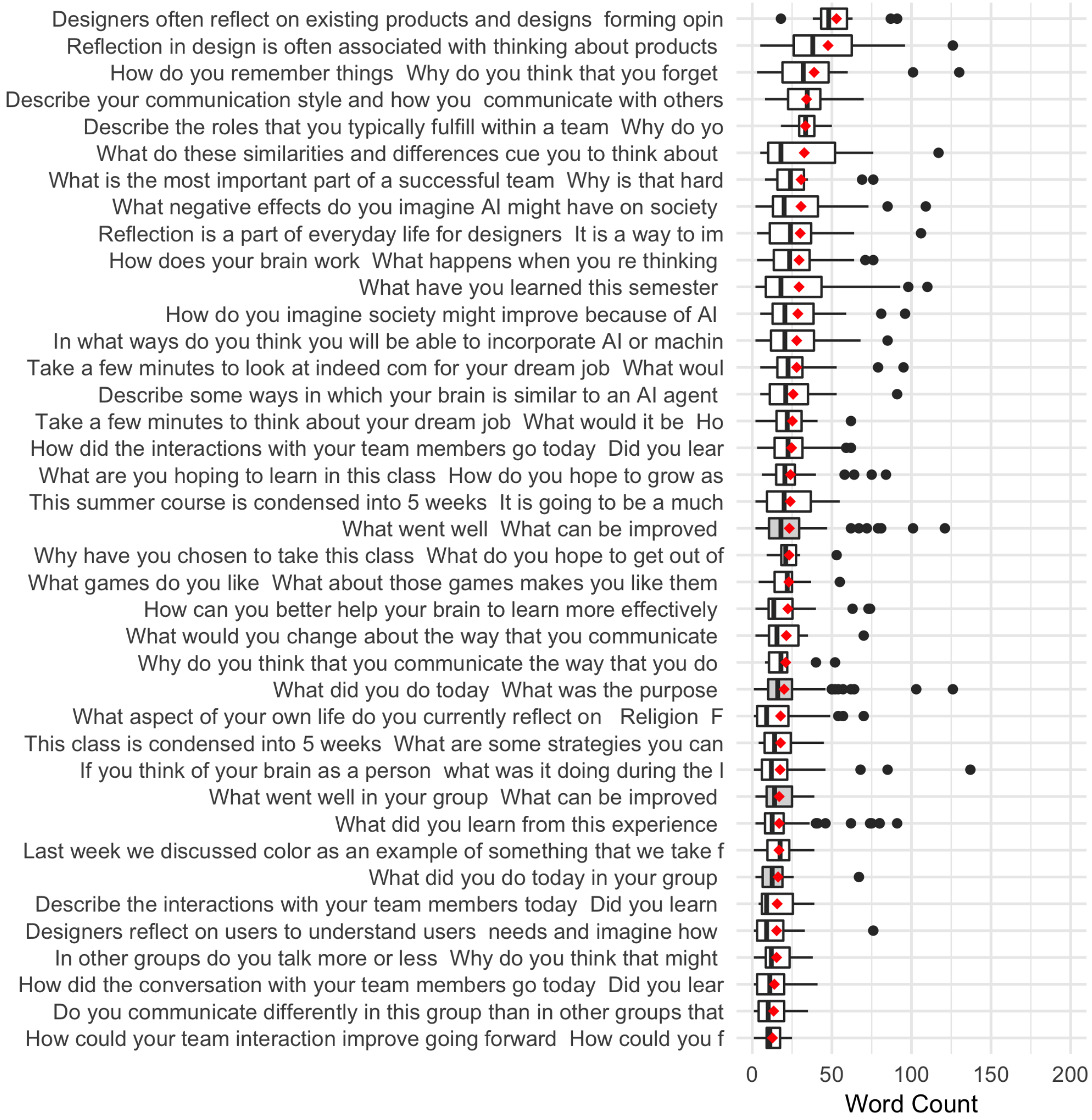


Figure 38: Boxplots for each reflective prompt sorted by word count.

and depth would be related, especially for terse reflections. However, none of the top ten longest reflections were rated as being the highest quality reflection. Most of the longest reflections were rated as a 3 or 4 because they did not include critical perspectives, intentions to change, or a broadened perspective.

The prompt that elicited the lowest reflection depth asked students to consider what they could do differently to improve their learning. Most students answered this prompt by describing things like, *"prepare better before class"*, *"more trial and error instead of just being lost"*, and *"write down more notes on what I did."* The most common words remaining after removing stop words included *practice, read, study, and notes.* To characterize these reflections, most students provided feedback about the class or wrote what they expected the instructor might want to hear. They did not question their own ideas or engage in any analysis about their behaviors, their experiences, or their learning.

For more open-ended reflective prompts, the reflections were deeper and students had more critical insights. For example here is an response from the reflection prompt that resulted in the highest depth, *"I would like to have a more professional way of communicating when it's appropriate. As of now, whenever I speak in a professional setting, it feels like I have to put on an act, and it usually feels like that act is not very good. But if I was able to think on my feet, and deliver statements in a more direct way, I could easier form professional speaking habits."* The student is aware of how they currently communicate, they have intentions to change, but it is not clear that they know how to make that change. For this reflective prompt, the most common words included *team, communication, assist, hard, teammate, and task.*

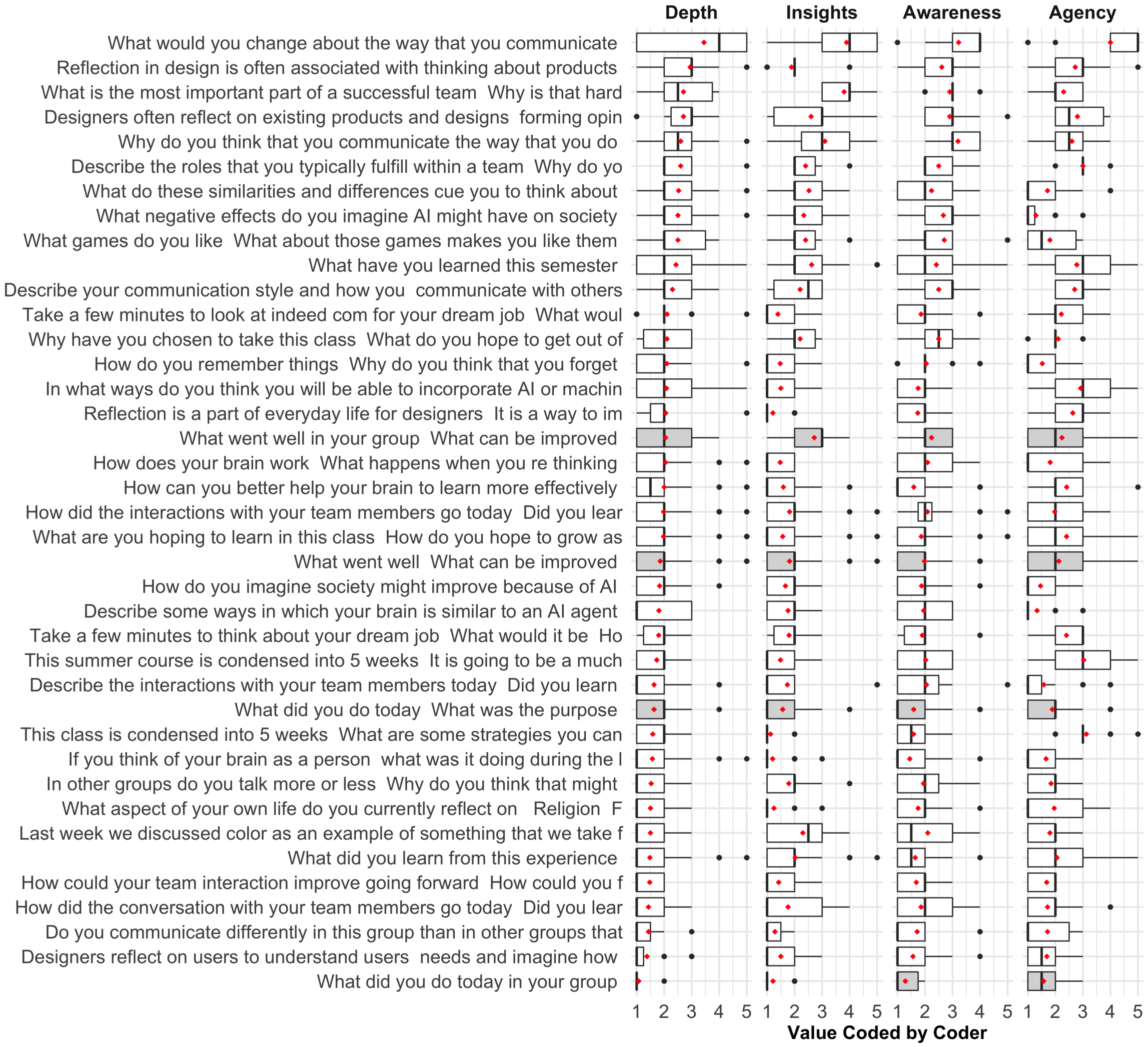


Figure 39: Boxplots for each of the manually coded reflection features. Boxplots are sorted by prompt based on the average depth. The gray boxplots indicate the common questions asked before and after the intervention.

Overall, there were few reflections with a depth of 5, which indicates that a student engaged in a transformational reflection that resulted from challenging their own assumptions. Prompts that had a higher-than-average depth of reflection, typically had few shallow reflections. This shows that the prompt can improve reflection, but that reflections with higher depth are extremely rare. These life changing reflections are rare and accounted for only 2.36% of the reflections submitted. These 39 highest quality reflections appeared across 9 prompts. It is possible that prompts increase the likelihood of these deep reflections, but it appears there may also be other latent factors, such as students' mood, experience, writing ability, and reflective capacity, which are not measured in this dissertation. Below are some examples of the highest quality reflections. Quotes have been slightly adapted for grammar, spelling, and to make them more concise.

- *"This semester has been a fantastic stepping stone into the world of AI. Looking back, I didn't retain much of the coding aspect of the algorithms; I went to the tutoring center early in the semester, only to find out that there were no tutors that could help me. On the other hand, I understand all the algorithms and have retained my knowledge of search algorithms like: Breadth-first, depth first, etc. This semester has really shown me that the field of AI is full of mysteries and unsolved problems. I am excited to continue in this field of study and to discover new, innovative ways of thinking."*
- *"I think I communicate this way because I sometimes struggle with ideas and while I say my idea, others can add onto it and then as a whole we can group*

together to determine what needs to be done/outcomes."

- *"I have learned that I can perform under pressure, this was a 16 week class that we have completed in 5 weeks! I'm sure a normal class might have more small assignments throughout the semester, ... So I am proud of myself for making it this far, and now all I have to do is pass the final and I'm done! Other, more concrete, things I have learned this semester is that AI can be as simple as a few if statements or as complex as a machine that learns from its own mistakes. I have learned that implementing these algorithms is only as easy as the framework built around the algorithm makes it. Finally, I have learned that I definitely want to learn more about AI and am glad that I get to in the fall!"*
- *"I ran into many impediments while working on my programming assignment, however those impediments forced me to develop a deeper understanding to solve my problem. Due to my learning disability reading text does virtually noting for me, so like usual I will use the key terms form the book and the section titles to find content online to explain the information in a way I do understand."*

Returning to Figure 39, awareness and insight appear to be related to each other and to depth. This is not surprising because both of these two reflection features are components of depth. For instance, being aware of many aspects of one's learning without insights would not likely lead to a deep reflection. Similarly, having a narrow but important insight does not always constitute a deep reflection. These two codes were added to better understand the components of reflection depth. The wording of the prompts appears to have a small effect on targeting either insight or awareness.

One prompt tells students to consider that the *class is condensed into 5 weeks* and asks them consider how they might adapt their learning styles to adjust. This prompt elicited a lot of awareness of the various ways that they can prepare, but generated few new insights. The most frequent words extracted from these reflections included *class, read, time, schedule, daily, ahead, study, and review.* Students talked about preparing for class, doing the prep work, reviewing materials daily, and keeping a study schedule. They were aware of many aspects that affected their learning but did not analyze them or have particularly insightful comments about those aspects. For the most part, these reflections were phrased as advice for another student, using distancing words like 'you.' Few reflections used the word 'I.' An example of a typical response was *"Put time into lessons everyday, work on assignments right after reading the text."* On the other hand, there were few examples of prompts that targeted insight without also influencing awareness.

Agency is another component of depth. It captures students' self-efficacy and their interest or intention to make changes to their behavior. This feature was least correlated with depth, insight, and awareness of the four features. Additionally, prompts appear to strongly affect this feature. Providing prompts that ask students what they would do differently in the future led to the highest agency. Examples of these kinds of prompts include *what could you do differently?*, *how can you incorporate course topics in your career?*, or *what would you change about the way you communicate?*. The prompts that led to the lowest agency included asking students to imagine the negative consequences of AI or asking them about what they did in their groups.

In Figure 40, the reflection prompts for topic, attention, and direction are pre-

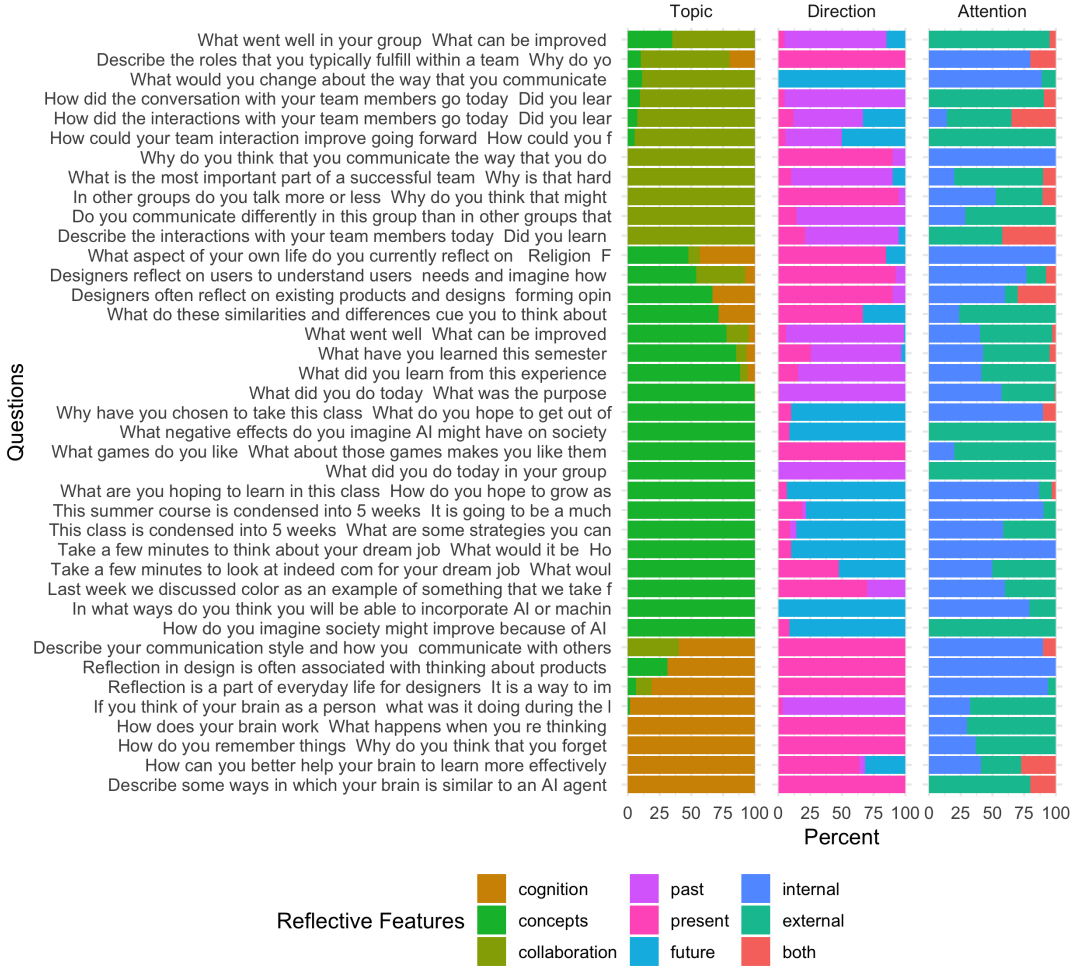


Figure 40: Reflection prompts for topic, direction, and attention broken down by the proportion of their possible codes. Sorted by the topic.

sented. They are sorted by topic which represents whether students were talking about cognition, social aspects, or concepts. Conceptual aspects were the most common with 44% of the reflective prompts priming students to reflect about conceptual aspects. All five of the prompts that were coded only as cognitive were about the brain or how the brain works. Four of them mentioned the brain explicitly. All of the eleven prompts that had at least 50% of responses about social included the words *team, communicate, or group* in the prompt. Across the three reflective features in Figure 40, there does not appear to be a correlation between the different reflection features. Therefore it may be possible to prime these different features independently.

Students' reflections were also coded based on whether their attention was focused internally or on external aspects. The results are shown in Figure 41. Reflective prompts about external aspects and group work appeared to elicit the most external responses. Internal focus was most common for prompts that asked students why they do things or asked them what they could change or what they reflect on. One prompt reiterates the potential impact prompt phrasing may have on priming students, *If you think of your brain as a person, what was it doing during the learning activity?*. One might expect that this would prime students to reflect internally about their brain. However, many students personified their brain as an external character and described what it was doing. It is likely that asking students the same question without mentioning *"as a person"* they would have reflected more internally. Understanding these nuances is important for designing reflective prompts.

Reflection prompts could be phrased in past, present, and future tense. The verb tense appears to have primed students to think about the past, present, or future.

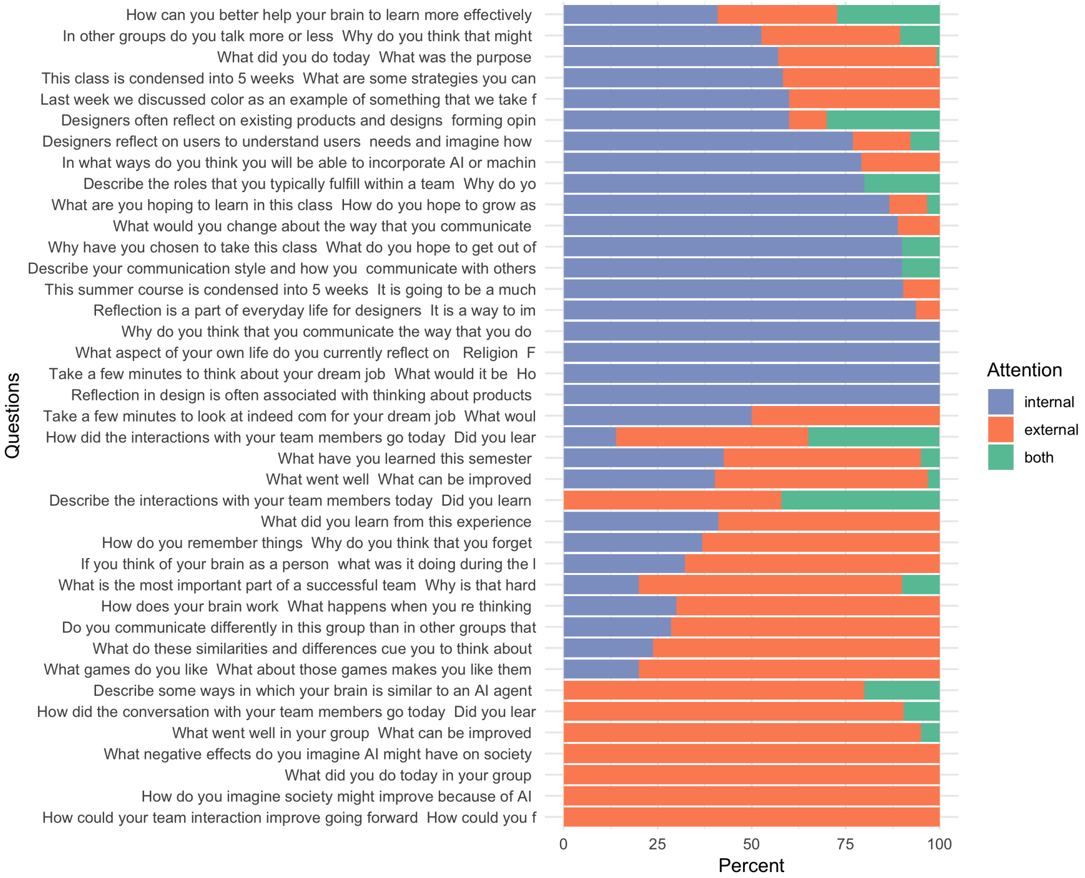


Figure 41: Reflection prompts for attention sorted by dominant code.

I was interested to see whether students used the same tense that was presented in the reflective writing prompt in their responses. The results from the study, shown in Figure 42, mostly support this expectation. All of the reflections that were coded unanimously as focusing on the present, used present tense in the reflection prompt. Similarly, the prompts that were coded as mostly focusing on the future asked students to reflect on their future careers, their strategies to succeed in the course, or the future impacts of AI on society. Finally, the one reflection prompt that elicited unanimous focus on the past was about students' past experiences.

Some of the reflections talked about past, present, and future aspects. This was

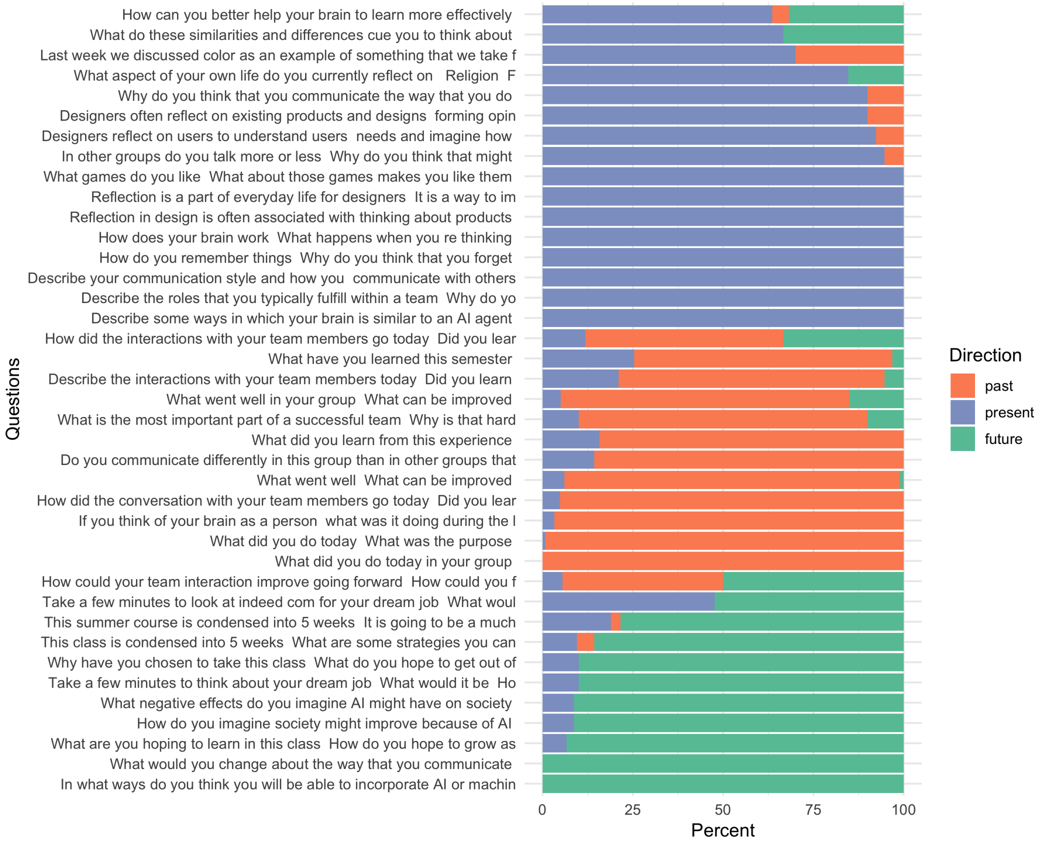


Figure 42: Reflection prompts for direction sorted by dominant code.

more common when students talked about group dynamics. In those cases, students talked about how their group has interacted, how it currently interacts, and how they expect things to change in the future. These were challenging to code when there was not a prevailing focus. Many of these were left uncoded.

Finally, when students were asked what went well and what could be improved, over 90% of the responses focused on the past rather than the future. It is not clear what this result means, but it was an interesting trend that appeared in the data.

6.4.2 Temporal Trends in Students' Reflective Practices

In the previous section, the relationship between prompts and the reflection features were explored. It is clear from those results that reflection prompts can influence and prime students' reflections. In this section, I wanted to see whether students' reflective practices change over time. Specifically, I wanted to see whether doing reflections more frequently led to better, deeper and more comprehensive reflections. An alternative hypothesis would be that students lose interest in reflection throughout the semester and experience some form of 'reflection burn out.' There is some evidence of this 'burn out' effect. In each class, I asked students informally about how the reflections were going. In one BloomMatrix class students indicated that they were overwhelmed with the amount of reflection. After discussing with the instructor, we dropped reflection five. In the other BloomMatrix class, the activities were slightly delayed and reflection five was dropped to ensure that all of the reflections were completed before finals week.

To understand students' temporal experiences, students' responses for each reflection session were grouped and visualized for each class. These line graphs show for each reflective feature how the classes' average responses varied over time. The number of reflections varied between two conditions. As mentioned, reflection session five was not included in both BloomMatrix classes. In addition, these students used BloomMatrix instead of doing reflection four. This resulted in six total reflections for IneqDetect students and four reflections for BloomMatrix students. Consequently, the line graphs do not have points for weeks four and five for the BloomMatrix classes.

In Figure 43, classes are compared across the six reflection sessions. In the first two reflection sessions, there was a lot of variability. These weeks also had slightly higher values for each of the reflection features, which makes sense because during these sessions the reflective prompts were designed to get students to buy-in to reflection. These prompts were also highly varied, asking students to speculate about the future,

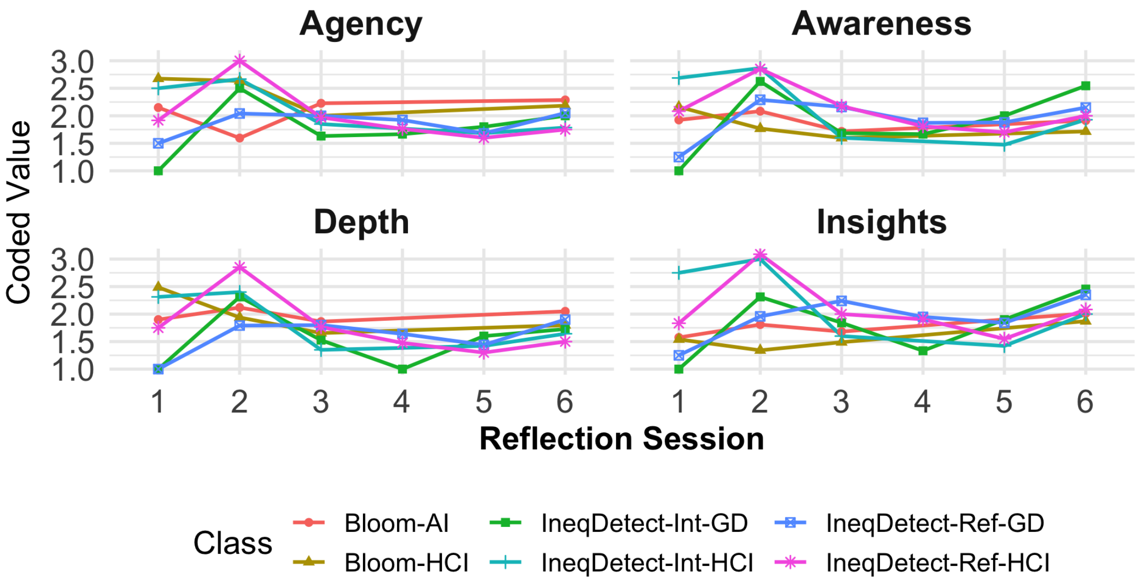


Figure 43: A time-series line graph that shows how agency, depth, awareness, and insight varied throughout the semester. Values represent the average across all the prompts for each reflection session.

to reflect on their behaviors, and to consider interesting aspects about the course material. Surprisingly, students' responses did not tend to decline throughout the semester. Despite some indications from students that the reflections were becoming tedious, the quality appeared to remain stable through the later sessions. In sessions three, five, and six, the same common prompts were asked repeatedly. It might be expected that students would get bored with those prompts and disengage, but this does not appear to be supported by the data which remains consistent through those periods of time.

The word count in the first session was also highly variable, and it was slightly elevated in the second reflection session. The graph for word count and sentiment are shown in Figure 44. For most of the classes word count also remained relatively stable through the second half of the semester. Sentiment had the least variability

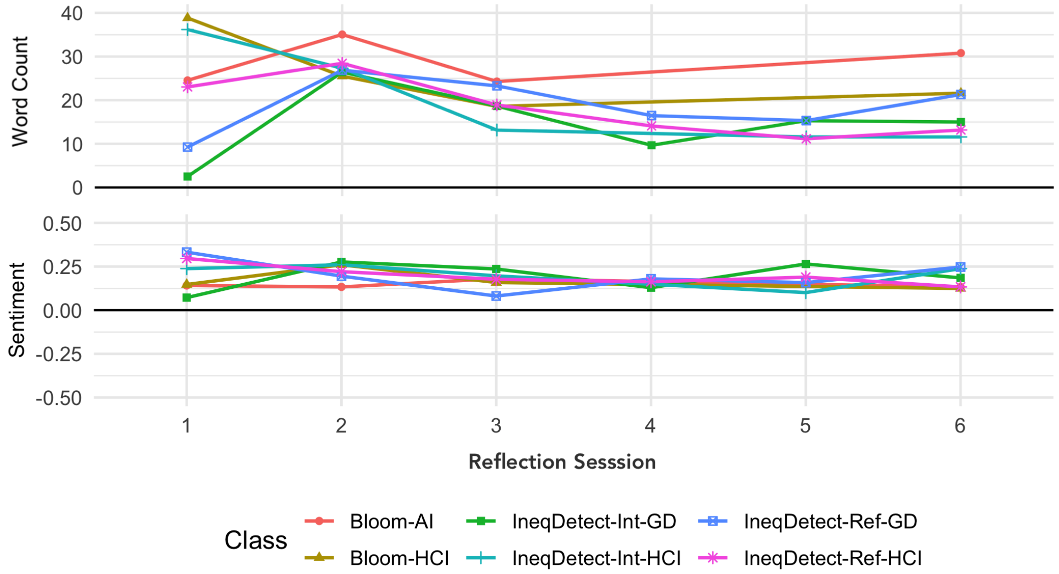


Figure 44: Time-series line graphs that show how word count and sentiment varied throughout the semester. Values represent the average across all the prompts for each reflection session.

throughout the semester. In the last section, the prompts did appear to have some effect on the sentiment of students' reflections. However, few prompts asked students to reflect on negative aspects of their experiences. Therefore, it is not surprising that there was very little variation throughout the semester.

Students can reflect on the present moment, on their past experiences, or speculate about the future. These categorizations are not mutually exclusive. Students can reflect on the past and extrapolate lessons learned to future contexts. We coded students' reflections based the temporal focus. Reflections that did not clearly have a prevalent temporal focus were left uncoded. The results are presented in Figure 45. In the first few sessions, the emphasis appears to be on the present and future. Students were asked to reflect on their future careers, the course concepts, and the course itself. At the end of the class, students were asked more about their experiences that

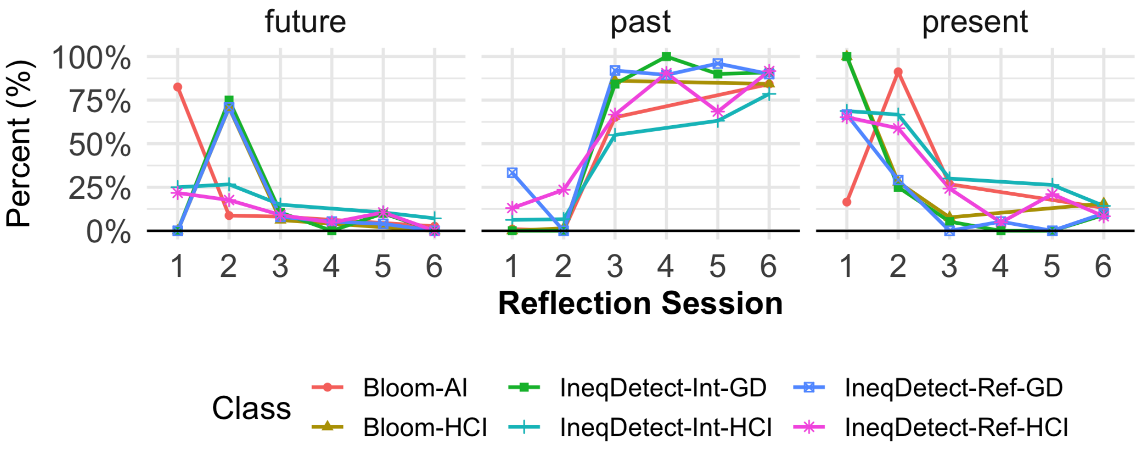


Figure 45: Line graphs that show at each point in the semester whether students focused on the past, present, and future. The graph is discontinuous in areas where no emphasis was placed on that aspect.

day, that week, or that semester. These trends appear to come through in the graphs temporally. However, coding for this reflective feature was challenging. Students often mismatched their verb tenses with the topic of reflection. Most commonly, students used present tense to refer to the past or future. For example, “In general, I can improve most activities by paying better attention to each individual task instead of attempting to multi-task (especially as multi-tasking is not truly something humans are capable of doing, instead we switch between tasks).” Students often answered questions phrased in future tense with present tense responses.

Reflections were also coded based on the topic of students’ reflections. We captured three possible topics; cognition, social interactions, or concepts. The results from these codes are presented in Figure 46. Reflections about cognition were the least prevalent focus. In the first two reflection sessions, students in the two BloomMatrix classes were primed by prompts that focused on cognitive aspects. Students in the IneqDetect condition were primed by prompts that focused on their social interactions

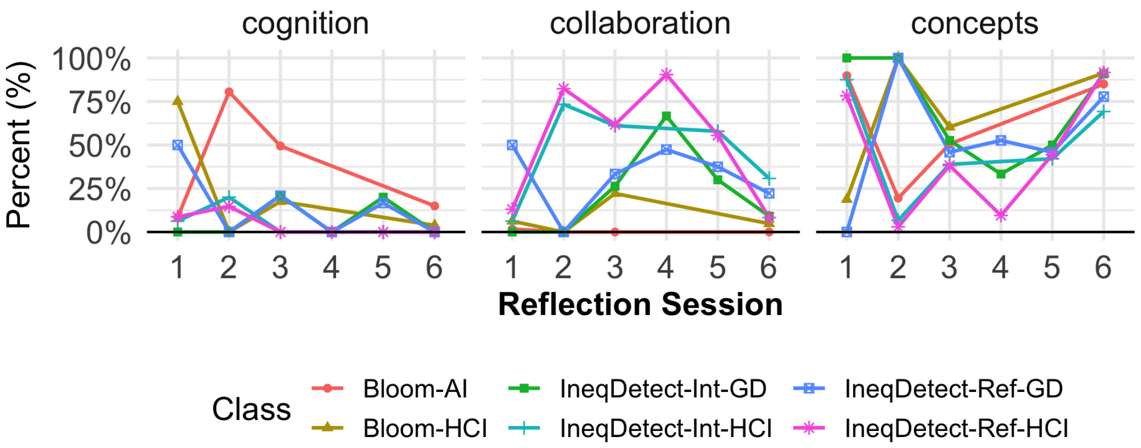


Figure 46: Line graphs of students' focus on cognitive, social, and conceptual changed throughout the semester. Separated by class.

and group dynamics. These prompts appeared to have some effect, but there is much more variation than for the temporal focus. The clearest trend in the graph is that the BloomMatrix students very seldom reflected on the social aspects, whereas, students from the IneqDetect class reflected more on social interactions and group dynamics. The effect appears to have faded leading into week six. This is surprising, because the prompts for weeks five and six were the same. One possibility is that there is a carry-over effect. In week five, students were still considering the social aspects from IneqDetect and the week four reflections which were focused on social aspects. But in week six, enough time had passed where students started to reflect on concepts again. Generally, there appeared to be a tendency toward concepts, unless cognitive or social aspects were primed.

The final coded value was the locus of attention. Students could focus their intention on themselves or on external aspects. External aspects could include things like course material or their team members. In practice, many reflections had aspects that

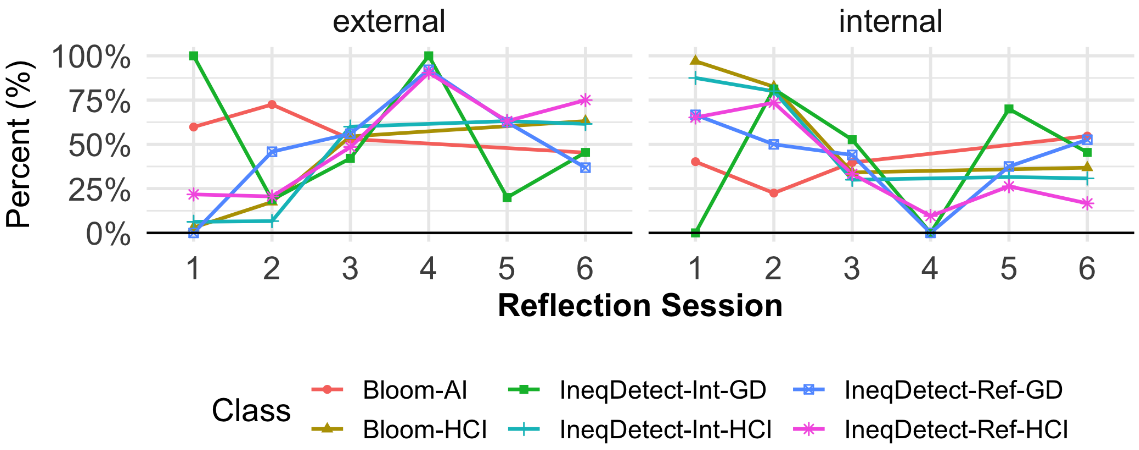


Figure 47: A line graph that shows how attention shifted between internal and external focuses across the semester for each class. Values do not always sum to 100%, because for some reflections there was not a dominant code.

were internal and aspects that were external. If there was not a clearly predominant term, it was left uncoded.

6.4.3 Interventions and Comparing Conditions

To understand the effect of the interventions in the different classes, the experiment was designed to have between and within subjects components. The between subjects component was whether students used IneqDetect, BloomMatrix, or reflective writing only to support their reflection. Students in the IneqDetect classes who did not use the system were in the reflective writing only condition. The within subjects component was to repeat the same exact reflection activity before and after the interventions. In sessions three, five, and six the same identical prompts were asked to all students in each condition. Session three is the ‘pre’ condition that happened before any interventions, and sessions five and six were combined into a ‘post’ condition.

To analyze these differences I plotted the data and also planned to use a multivariate mixed effects model to analyze the data. In the model, students were nested within

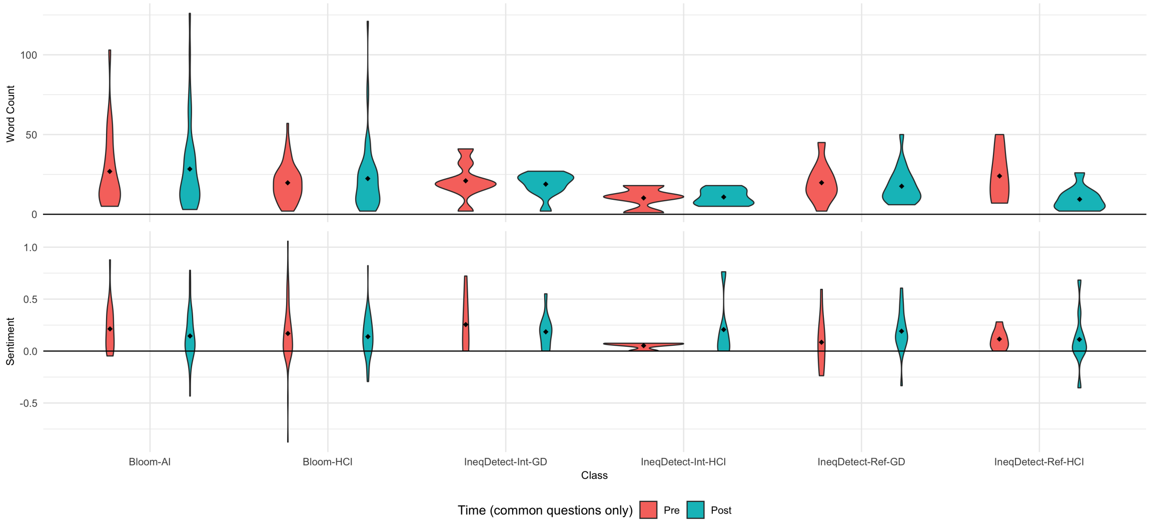


Figure 48: Violin plots showing sentiment and word count for the common reflection prompts that were given before and after each intervention in each class.

their class as random effects. The fixed effects included the reflection session and study conditions; IneqDetect, BloomMatrix, and reflective writing. This model was not used because after visual comparison, the differences appeared to be very slight. It is unlikely that these small effect sizes would be significant after accounting for multiple comparisons across all of the dependent variables. For these reasons, visual comparisons are the primary source of analysis in this section.

In Figure 48, sentiment and word count are presented as violin plots. They are separated by class and colored by whether they happened before or after the intervention. There appear to be some minor differences in word count between some of the classes, but the distributions are pretty similar for the most part. It was hypothesized that after the intervention, students would write more. Based on the results, this does not appear to have happened. This may be more evidence that prompts have a strong priming effect on students' reflective writing assignments.

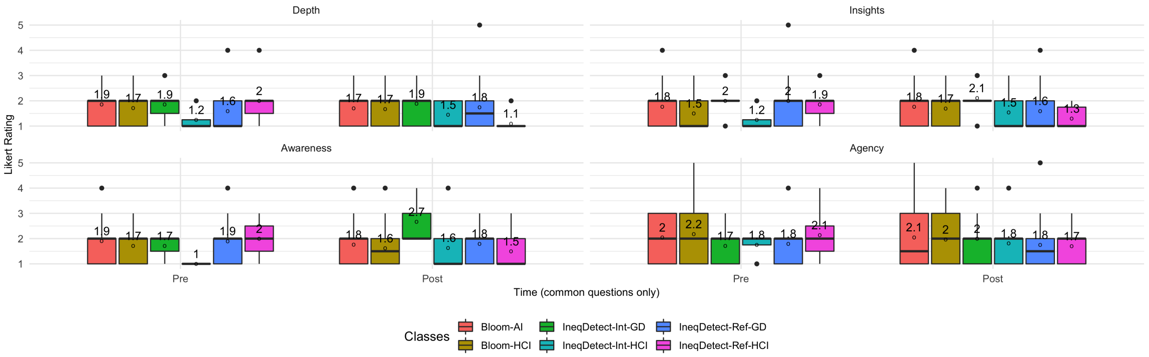


Figure 49: Box plots for each of manually coded reflection features from the common reflection prompts that were given before and after each intervention in each class.

Similarly, students' other reflective features did not appear to change much between the pre- and post-conditions, as seen in Figure 49. I had hypothesized that the interventions would improve these reflective features between the pre- and post-conditions. There are a number of possible explanations why this did not happen. First, as shown repeatedly in this chapter, the reflection prompts appear to have a strong priming effect on students' responses. This means that any effects from the intervention are being overshadowed by this priming effect. Second, reflective writing may not be an effective proxy for reflective practice. Previous chapters show instances of deep insights, higher agency, and increased awareness, but it is possible that these reflective breakthroughs were limited to when students used the tools. The tools may have increased the students awareness and insight, but not on these common reflection questions which were not directly related to the tools themselves.

6.5 Discussion

In this chapter, I explored the reflective writing assignments to understand students' reflective practices, their experiences with reflection, and to determine whether the RSTs had any effect on students' reflective practices. I made an assumption in this chapter that reflective writing assignments would be an appropriate proxy for students' reflective practices. Based on my analysis, this does not appear to be the case. Reflective prompts appear to have a strong priming effect on students' reflective writing assignments. This can bias students to reflect in ways that may not be representative of their reflective practice. For instance, previous chapters have included numerous examples where students used RSTs to increase awareness, identify insights, and make changes to their behaviors. These reflective insights may or may not have changed students' reflective practices as a result, leading students to reflect more frequently and deeply on cognitive and social aspects of their learning. They may also have been limited to the immediate context and may not have had a lasting impact on students' reflective practices. In the same way, it is not clear whether or how reflective writing influences students' reflective practices. Furthermore, given the impact that reflective prompts have on students' reflections, are the reflective responses actually representative of students' underlying reflective practice? In this section, I will discuss the insights from this chapter related to these aspects.

6.5.1 Reflective Prompts Prime Reflective Writing

The first finding from the analyses was that the reflective prompts appear to have a strong effect on students' reflective writings. There appeared to be some evidence that the prompts primed and influenced each of the features that were coded from the reflections. Asking students about specific learning experiences reduced the depth of the reflections. Asking students about the positive or negative aspects of their learning experience led to positive or negative affect in their responses. Asking students what they would change or how they could do things differently improved their agency, and asking students to focus on a specific aspect of their learning, such as cognition, social interactions, or course concepts primed students to focus on those aspects.

It appears that these priming effects tended to overshadow students' reflective practices. There are a variety of reasons for this. First, many reflections appeared to present a previously identified insight rather than an active reflective process. Second, reflective insights did not appear to influence reflective writing assignments. For instance, students did not mention IneqDetect or BloomMatrix by name in any of the reflective writings. Third, many reflections had evidence that they were written for the reader, for instance, providing extraneous details that would only be helpful for someone reading the reflection. In these cases, students were not reflecting on their experiences as much as they were communicating their thoughts and experiences to the reader. For these many reasons, it is possible that written reflections are not a good proxy for reflective practice. They may assess students' capacity for reflection, but it is not clear that they measure and capture students' reflections in real-time. On the other hand, interviews included many instances where students said something and then questioned their comment and revised it or provided more clarification.

6.5.2 Deep Reflections Are Rare

The prompts appear to strongly influence students' reflections, but only to a point. The highest-rated, deep, critical reflections were extremely rare, and only accounted for 2.36% of all the reflections. Prompts that elicited deeper reflection on average did so by reducing the number of shallow reflections, not by producing a disproportionate number of highest quality reflections. It appears that open-ended prompts elicited longer reflections which were also deeper on average. But to consistently encourage deep reflections, it may be necessary to create new interactive reflections that challenge students to reflect more deeply. Because most models of evaluating reflection describe reflective stages, it may be possible to create adaptive prompts that start by trying to get students to describe what happened, then ask them to analyze it, then ask them to connect it to other experiences. These adaptive prompts could take the form of a reflective chatbot. In addition to moving students to deeper stages of reflection, the chatbot could also help students create a plan to make changes. The chatbot could even prompt students to reflect on their changes and goals. This form of *Dialogical Reflection* would be more similar to a conversation that one might have with a friend or therapist who is trying to help put an experience into perspective.

6.5.3 Interest and Engagement

There were a few instances where students got really engaged by the reflective prompts. This did not always result in a deep or insightful reflection, but some reflections were a hundred or more words. In most of these cases, students appeared to be sharing personal information about their background, sharing their interest, or demonstrating their expertise on a specific topic. Not all students reflected in this way, but at least some students appear to value a conversation with the instructor or TA that reads the reflections. This may provide more evidence that a *Dialogical* reflective chatbot might be an effective way to support these students. Having a chatbot that could engage them in their interests and help them make connections between that interest and the course concepts could be very valuable.

6.5.4 Limitations and Future Work

The results presented in this section are suggestive rather than conclusive. This in-the-wild study was conducted with many assumptions that have been previously discussed. These assumptions were necessary to make, but in some cases they did not appear to be supported by the results. Future work is needed to replicate these studies in more controlled environments with few variations. For instance, more work is needed to understand how the phrasing of reflective prompts influences these reflection features. Capturing students' reflective practices is also still an open question. I had assumed that reflective writing could be a proxy for reflective practice, but considering this chapter in light of previous chapters, it appears that triangulation is always necessary to understand reflective practice. Students reflect in many different ways and writing may be capturing other aspects such as reflective capacity, interest in the topic, and preferences for disclosure. These aspects do not necessarily capture students' evolving reflective practice. Finally, this work introduced a repeated measures study design for evaluating RSTs. This did not appear to be effective due to the influence of reflective prompts. In the future, students may need to be asked to reflect on the RST itself rather than on their experiences.

6.6 Conclusion

In this chapter, I presented the study design and the results of collecting reflective writing assignments in the classes that used either IneqDetect or BloomMatrix.

The study was designed to make comparisons between these conditions and also to understand the changes that these RSTs had on students' reflective practices. To analyze the reflective writing assignments I used manual and automated qualitative coding techniques to extract reflection features. Through this process, I obtained insights about reflective writing prompts, students' reflective writing practices, and about study designs for evaluating both RSTs and reflective practices.

Analyzing the collected reflections showed a strong priming effect for the reflective writing prompts. This priming effect made it difficult to interpret differences between the different conditions. The prompts were created primarily to help students learn, and therefore they were highly customized for each class. Therefore the priming effect can be see as confound, de However, it generated many new avenues of research. Some of the open questions include understanding how reflective writing prompts prime reflection, methods for capturing and evaluating reflective practice, and the design of studies to evaluate RSTs. For future work, I intend to explore new ways to support reflective writing using technology, such as, by building a chatbot to promote *Dialogical Reflection.* I also intend to create standard instruments to capture, isolate, and evaluate reflective practice.

CHAPTER 7: DISCUSSION, CONCLUSIONS, AND FUTURE WORK

The goal of this dissertation was to provide students with reflection support tools (RSTs) that would help them to identify insights about their learning while improving their agency and awareness. The tools presented in this dissertation also served as probes into a potential ecology of data-driven RSTs that support multi-dimensional reflection. This holistic perspective differs from existing models of reflection, and it challenges existing notions of RSTs. Many existing RSTs frame reflection on a single focal point, such as student affect [11]. These RSTs are seldom intended to be used in tandem with other RSTs. In most cases, they are only evaluated for their ability to generate insights, but not their ability to develop a student's reflective practice.

To address these two gaps in research, I designed and deployed two RSTs and created a model for supporting holistic reflection along multiple dimensions. In Chapters 4 and 5, the RSTs were evaluated for their ability to generate insights for students along each dimension. In Chapter 6, both RSTs were evaluated for their ability to support and develop students' reflective practices. These tools were intended to be used as part of an eventual ecology of RSTs, guided by the New DEAL Model, presented in Chapter 3.

I created IneqDetect and BloomMatrix to start building an ecology of tools that could help students to reflect on their learning holistically. IneqDetect captures and visualizes students' conversations in groups to improve conversational equality.

Table 11: An overview of the main takeaways from this dissertation work.

Takeaways
Reflective Prompts Frame Reflective Writing
Deep Critical Reflection is Rare
Awareness, Agency, and Insights
Automatically Captured Data was More Trustworthy
Reflection as a Conversation
Data as a Shared Representation
Strategies for Evaluating Reflections

BloomMatrix helps students reflect on their cognition and on an aggregated representation of other students' cognition. These tools were successful in supporting reflective insights, increasing awareness, and improving students agency. However, these tools did not appear to have a significant effect on students reflective practice longitudinally. In this section, I discuss why this may not have been a reasonable goal to achieve. I summarize the lessons learned from this comprehensive multi-classroom study which triangulated students' reflective experiences through surveys, interviews, and reflective writing assignments. I conclude with a series of design implications and a list of future research areas that have surfaced from this work.

7.1 Takeaways

This dissertation resulted in many insights about reflection and about the design and evaluation of RSTs. These takeaways are summarized in Table 11

7.1.1 Reflective Prompts Frame Reflective Writing

Across all of the classes, the reflective prompts appeared to strongly influence the reflective responses from students. For instance, open-ended prompts that elicited students' opinions about themselves, their teams, or concepts encouraged longer and deeper reflections; whereas, specific prompts about students' experiences in class often led to terse, shallow reflections. The verb tense of the reflective prompt also influenced whether students focused on the past, present, or future. Similarly, phrasing appeared to shift students' attention internally to themselves or externally to concepts and others. Finally, certain prompts appeared to encourage or discourage agency in students' responses. In survey design, the way that a question is phrased can have a strong priming effect on the person who responds. This is well understood and has been studied extensively, but this amount of rigor has not been applied to understanding reflection prompts.

Although additional work is needed to evaluate the effect of reflective prompts in a controlled laboratory study, these differences appeared to have a strong impact on students' reflections in this dissertation. This means that caution needs to be applied when using reflective prompts to evaluate RSTs. It also may suggest that in-the-moment reflection may not be representative of a student's overall reflective practice. Students who reflect deeply or shallowly in the moment may be reacting to the prompt more than expressing their own, internal capacity for reflection, if one exists. The framing effect of prompts also challenges us to think more about whether questions affect reflection during interviews, and whether tasks and representations shape reflection when using RSTs. Future work is needed to better understand the framing effect of reflective writing prompts and RSTs. Guidelines need to be created for the design of reflection prompts. Finally, standardized prompts should be used in studies to make it possible to compare between studies. In this dissertation work, the standardized prompts were *"What did you do today? What was the purpose?"* and

"What went well? What can be improved?"

7.1.2 Deep Critical Reflection is Rare

Deep critical reflections only accounted for 2.36% of all the written reflections in this dissertation. Some prompts appeared to be more likely to support these types of reflections than others, but in general the reflective prompts that had the deepest reflections on average did so by minimizing the number of shallow reflections. This finding suggests that the design of reflection prompts is important, but only to a point. To ensure that students consistently have deep, critical reflections, more support is needed for students. One possibility includes prompts that can adapt to scaffold students toward deepening their reflections.

Students described deep critical reflections enabled by RSTs. Students that used IneqDetect had critical insights about the way that they communicate, which led to intentions to change their leadership styles, speak more or less, and adopt new team roles in the future. Students described being explicitly inspired by the visualization and by their team members. For BloomMatrix, critical reflection was less common, but students did indicate intentions to make changes to their study habits. Future work is needed to understand the factors that lead to deep, critical reflections.

7.1.3 Awareness, Agency, and Insights

IneqDetect and BloomMatrix effectively generate insights and improved agency and awareness for many students. Many of these students described insights that they claimed that they would not have been able to identify without the RSTs. BloomMatrix provided information about the cognition of other students in the class which would be difficult for students to obtain and reflect on. In most cases, this increased awareness led to insights, and students appreciated this new perspective. However, some students questioned the value of seeing other students' perspectives, saying that it had little impact on how they think. For students who valued this expanded awareness, many described improvements in their motivation and intentions to make changes to their learning. This increased agency is a strong motivator for continuing to develop and deploy RSTs in the classroom.

In spite of these successes for most students, the question remains of how to motivate students who do not care about their peers' perspectives. Students generally expressed an interest and curiosity about the data presented to them, and that may be one inroad to engage the disinterested students. At the same time, motivation has always been difficult to foster in classrooms, and these interventions already appear to be effectively engaging and motivating most students.

7.1.4 Automatically Captured Data was More Trustworthy

IneqDetect captured information about students' experiences automatically using microphones. BloomMatrix crowdsourced students' manual responses about their cognition. What was interesting about these two approaches is that students were very confident about the data that was automatically captured, but less confident about the manually collected data. Many students speculated that the results from IneqDetect were 90% accurate. Although BloomMatrix students were not asked about their perceived accuracy of the heatmaps, some students were skeptical that the results accurately measured cognition. Some students indicated that they did not think other students accurately responded to the heatmap. More broadly, students questioned whether aggregating other students' cognitive processes accurately mapped to their own cognition. Students who used IneqDetect also questioned what the accuracy meant, but they did not question the integrity of the data itself. For instance, students had questions about whether talk time was an accurate measure of a conversation. There appears to be a bias from students toward trusting data that is captured automatically. This is an open question, because many of these students were CS majors. Students did not appreciate the extra effort that they personally put into collecting the data for BloomMatrix. This suggests that for RSTs, manual data collection increases effort for students without a clear additional benefit.

7.1.5 Reflection as a Conversation

An interesting insight from the reflective writing assignments was that many students appeared to write their reflections for the reader more than for themselves. There were few instances where students changed their opinion or questioned their perspective midway through their writing. Instead, many students described reflective insights that appeared to have already occurred, they used words like 'you', they used arguments to support their ideas, and provided additional context that would only be valuable for the reader. In a few cases, students were more explicitly writing for the reader saying things like *"as I said in the last question...'* Across these many examples, some students saw reflection as a conversation between themselves and the reader. For this reason, reflection needs to be differentiated from reporting about reflective insights. It may also be that students see reflection as a conversation between themselves and the teaching staff, an opportunity to express themselves and their ideas. That is valuable, but it is not reflection in the classical sense. In parallel, a study is needed to understand how students reflect when no one is watching. Additionally, affordances need to be created that allow students to share their thoughts and express themselves to their peers and the teaching staff.

7.1.6 Data as a Shared Representation

Students using IneqDetect did not always describe a shared group experience the same way as their group members. Some students observed and noticed specific changes while others were not aware of those changes. In Chapter 5, I discuss some reasons for this, including non-linear perceptions of conversation. Students appeared to perceive time and conversation differently depending on whether they were listening or talking. They also perceived high quality contributions to the conversation as having taken more time.

Based on these observations, it is possible that data and visualizations can serve as common ground for students in a group or class. In a few instances, students described how they did or would use the data to convince their team members of their point. Having this data appeared to be an objective measure of an experience and students felt less intimidated to address problems highlighted by the data, instead of problems that they identified on their own.

7.1.7 Strategies for Evaluating Reflections

This work made many assumptions about reflective practice and about how to study reflection and RSTs. This was necessary because few studies exist that attempt to capture and measure reflection. Like creativity, reflection is a process that is complex, ephemeral, and difficult-to-define. Unlike creativity, there do not exist any standardized instruments for measuring it. Understandably, many of the assumptions were not supported by the data, and as a result, this dissertation has many recommendations for future studies that evaluate reflection.

The assumptions made include:

- In-the-moment reflections can be captured at a single point in time through reflective writing, interviews, and RSTs.
- In-the-moment reflection is a good proxy for a students' reflective practice and their capacity for reflection.
- Reflective prompts would have a small effect on students' reflective responses.
- Students will reflect primarily for themselves.

7.1.7.1 What is Reflective Practice? When does it happen?

As mentioned, some students' reflections appeared to describe insights that had happened earlier. Even when students work on reflective writing assignments, it is possible that they reflect on the prompt and share that reflection in the writing. In an interview, students may describe reflective insights that happened previously, or they may be actively developing those reflective insights in real-time. This suggests that in-the-moment reflection may not always be captured in real-time. It also suggests that in-the-moment reflection is not necessarily indicative of a broader reflective practice. Students may identify insights and have momentary intentions to change, but these may not last beyond the immediate context. Alternatively, students can have a high capacity for reflection in some aspects of their lives, but that may not translate to in-the-moment reflections in other aspects of their lives. More work is needed to understand reflective practice, to understand its relationship to in-the-moment reflection, and to understand how to measure and capture both. Reflection encompasses many aspects, such as self-efficacy, world-view, and metacognition. Our understanding of these concepts is developing, and so it will be an ongoing effort to continue to operationalize reflection.

7.1.7.2 Triangulation is Necessary

Given the complexity and that reflective practice may not be fully captured in a single source, it may be necessary to consider multiple ways to capture different aspects of students' reflective practice. Reflective writing may capture one aspect of students' reflections, but interviews capture another aspect and may actually lead to more in-the-moment reflection. Students in the interviews frequently made a comment, reflected on the comment, and then revised or developed their thought further. In this way, the reflective process is at least to some extent externalized. Self-reported insights from using RSTs and examples of behavioral change are another way to capture some aspects of reflection. It is possible that these data sources show the results of reflection more than reflection itself. Through triangulation, it is possible to capture the externalized process and the outcomes of reflection. This triangulation is also more likely to be representative of a student's reflective practice.

7.1.7.3 Expect the Unexpected

Using RSTs and engaging in reflection are nebulous processes. The experiences can vary widely across students in the class. This makes generalization very difficult. It also makes it difficult to form hypotheses, because the experiences are often very open-ended. Scaling down the experiment to mitigate the number of confounds and to control for specific variables can reduce some of the complexity. However, removing these aspects also influences students' reflections. This can lead to an overconfidence in trends and patterns that change when the confounds are re-introduced in real-world settings. For instance, in these studies, prompts had a significant effect on students responses, but this was only apparent because students answered 45 different prompts across four classes. Asking students the same three questions each week throughout the semester would have been more interpretable but would have missed this important aspect. For this reason, it is important to expect the unexpected and embrace the complexity of reflection. There were many interesting surprises in these studies. For instance, in one case IneqDetect, which was designed to improve conversational equality, led to competitive, anti-equitable behavior. It also led another team to switch leaders and led the former leader to talk less as a result. These surprises are valuable for understanding reflection and how to build tools that support students effectively.

7.1.7.4 Responding to Reflection May Bias Results

In these studies, I responded to students' reflections with brief encouragements. This was done to improve students' motivation and to increase their response rates to the reflection prompts. It appears to have been effective; however, it may also have caused them to write for my benefit, rather then for their own benefit. I presented examples where students provided additional context and used words such as ‘you’ as if they were writing text that was intended to be read. Additional studies could investigate this important aspect of reflection.

7.2 Design Implications for RSTs

For those who want to build RSTs that address students’ needs, I have identified the following best practices to guide the design of RSTs.

7.2.1 Ease of Use and the Ratio of Value to Effort

A theme that repeatedly emerged throughout this work was that students valued *Ease of Use*. This addresses one of Christopher Day’s primary reasons for abandoning reflection, *a lack of time* [44]. Students indicated repeatedly that they preferred one form of reflection to another because it was faster or it was easier. Students have many priorities competing for their time, and it is understandable that reflection can be seen by students as just another demand on their time. In surveys, students repeatedly described ease of use as the reason that they preferred reflective writing or one of the RSTs. This is another reason why automated data collection methods may be preferred for reflection. BloomMatrix required students to manually enter data and was cognitively demanding to use. These factors were cited by students as reasons for why they preferred reflective writing to BloomMatrix. At the same time, some students complained about the additional work, but acknowledged that they thought the reflections were valuable. They lauded the insights that they obtained and knew that it was worth their time. This suggests that although ease of use is paramount, students will make the extra effort if the perceived value of the insights is high.

7.2.2 RSTs should be Appropriable

Research on gamification has shown that competition can be engaging for some people, but it can cause other people to discontinue participation [51]. Similarly, some students appeared to become highly engaged by some reflective writing prompts, while others were completely disengaged. It is likely that RSTs will be used differently by each student. For example, IneqDetect was designed to improve conversational equality within groups. However, IneqDetect was used by one student to compete with his teammates. This use was contradictory to the design goals of IneqDetect, but it provided direct benefits to that student. Appropriation is a natural part of most successful sociotechnical systems. Ignoring this possibility limits the potential effectiveness of the tool. Embracing this possibility ensures that RSTs will meet the various reflective needs of students. For this reason, ecologies of RSTs should be made available to students when possible. Most students preferred IneqDetect, but some students also preferred reflective writing or BloomMatrix. Providing students with options ensures that they will be able to explore their experiences more holistically, likely leading to deeper, more critical reflections. Consequently, success should be defined in part by a tool's ability to support insights in a variety of ways.

7.2.3 Task-based Reflective Activities

I have already shown how RSTs addressed one of Christopher Day's primary reasons for abandoning reflection, *a lack of time*. Here I discuss his other primary reason for abandoning reflection, *a lack of structure* [44]. The studies in this dissertation were designed to provide students with the minimal amount of guidance when using the RSTs. I adopted this approach to understand students' existing reflective process and to understand each RST's ability to be appropriated. This was important to do because there is not a lot of existing research about how students use RSTs or how they interpret their own data. I was also interested to see whether and how they form and test their own hypotheses with data. Providing students with specific tasks or hypotheses would have embedded my own biases and expectations into the study. It would have also prevented me from seeing whether and how students form and test hypotheses with their own data.

This open-ended approach was reasonable for a first step and led to many insights about how RSTs can be appropriated for new uses. However, more work is needed to understand whether and how tasks impact students' reflections. Based on the results of this dissertation, it appears that open-ended tasks are better for tool appropriation and for students to develop their own questions and hypotheses. But some students indicated that they wanted more guidance and specific tasks to complete. Students liked this about reflective writing, it gave them a prompt to follow. These students need to know what to track and what patterns to look for. In the IneqDetect study, students often asked me what to look for in the visualizations, or they asked if their results were *good*. Students seldom think about their group dynamics or their cognition. It is reasonable that not every student had a frame of reference to interpret the results. Incorporating tasks is a good area for future work. However, I think that these tasks should be offered only after students reflected on the results on their own. Reflection is open-ended, and overly specific tasks can help, but they should be used with some caution.

7.2.4 Provide Data that is Unavailable Otherwise

Data-driven RSTs are uniquely positioned to provide students with information about themselves and the class that they would not be able to obtain otherwise. For instance, an RST can capture and visualize emotional prosody or slight facial expressions that are relatively undetectable for students. Alternatively, RSTs can collect and summarize the behaviors of many students in the class. This data may be more valuable than data that students can obtain directly themselves. At the same time, this data may be less interpretable than conversational equality or cognitive processes; these two data types were already difficult for students to understand and operationalize in their own contexts during these dissertation studies. It may be most valuable to provide multiple different types of data. Data that is unattainable directly can be contextualized by familiar measures of collaboration such as talk time. As an example, students' speaking patterns can be visualized temporally, with emotional prosody, gesture events, and intonation plotted above it. Students can then more easily find reference points and contextualize the unfamiliar with aspects that are familiar and observable, such as regions where they were speaking.

7.3 Future Work

This dissertation has prompted many new research questions which were outlined in previous chapters and earlier in this chapter. Based on those insights, I propose the following future research directions.

7.3.1 A Reflective Ecology

I intend to continue working with RSTs in the classroom. The goal of this work is to continue building and deploying RSTs into the classroom to form an ecology of RSTs. These tools would capture students' behaviors across different learning tasks. Guided by the New DEAL Model, students could reflect on their behavior across time and learning activities. Tools could provide information about students group dynamics, their affect, cognition, and evolving understanding of course concepts. Meta-level RSTs could even orchestrate students' use of these various different RSTs.

7.3.2 RSTs to Scaffold Reflective Practice

In addition to supporting more comprehensive reflection along multiple learning dimensions, I would also like to develop tools that help students improve their capacity for reflection and foster a reflective practice. Currently, RSTs focus on in-the-moment reflections that may or may not extend beyond the immediate context. They are specialized to support reflection about a specific aspect. Future RSTs could be developed to support reflection itself. In this case, reflection is the goal not the means to improving learning. These RSTs could help students identify when they are being biased. They could encourage students to consider alternative perspectives. They could help students know when to reflect and what aspect to reflect on. Improving students' reflective practice would also likely help students to use existing RSTs.

7.3.3 Dialogical Reflection with Chatbots

One way to help students develop their reflection skills is to use dialogical reflection. In students' reflective writings, I often saw potential for reflective insights, but students did not explore their thoughts or experiences deeply enough. Dialogical reflection could help guide students to reflect more deeply with adaptive prompts or chatbots. For instance, existing stage-based models of reflection start with a description and progress through analysis, forming relationships, and questioning assumptions. A chatbot could talk students through these stages by asking questions iteratively based on students' reflective responses. Eventually, it is possible that students would go beyond descriptions without the help of the chatbot.

7.3.4 Developing an Understanding of Reflective Practice

One of the most salient findings from this dissertation is that we still do not have a good understanding of reflection. Reflection is a nebulous, ephemeral, and personal process that is difficult to define, as shown in Chapter 2. Our understanding of cognition and learning has evolved by leaps and bounds in the last century, but we still do not know how to reliably measure and evaluate it. Reflection is similarly complicated and has received much less attention.

This dissertation begins to shed light on the nature of reflection and reflective practice. Previous tools have focused exclusively on in-the-moment reflection and insights, but have seldom investigated how RSTs affect students' reflective practice or their capacity for reflection. Future work needs to investigate the relationships between short-term reflective experiences and long-term reflective practice, and between reflective insights and reflective thinking. It is also important to understand reflective thinking and related skills. Similarly to the way that coding and computational thinking are not the same, reflection and reflective thinking may not be the same. The priming and framing effects of RSTs need to be explored more deeply. This dissertation contributes in this area, and it adds to our evolving understanding of priming effects in reflection [34]. Finally, it is important to consider how to evaluate reflection in light of the findings in this dissertation.

7.3.5 Reflective Pedagogy

In addition to developing our understanding of reflection and reflective practice, it is important to develop pedagogies that help students learn to reflect on their experiences. This aspect is not novel and many instructors are integrating reflection activities into their classes. The vast majority of the reflective tasks are centered around reflective writing. Based on the findings from this dissertation, I'm not convinced this is an effective approach to developing students' reflective practices. I identified many challenges related to supporting reflection in the classroom. First, many students do the minimum required work to get the reflection over with. Second, there were instances of reflections that were for the benefit of the reader and not true reflections. Third, reflection does not appear to engage all students equally. There are many assumptions currently made about reflection and few studies and activities are grounded in theory.

To develop a reflective pedagogy, more work is needed to understand the nature of reflection. Reflection should not be taught as an activity but as a way of thinking. Better measures are needed to track and evaluate reflective practice. Tools need to be constructed to triangulate students' experiences and to allow them to reflect in various ways. It is also important to better understand the long-term effects of engaging in reflection activities.

7.3.6 Reflection Thinking as a Digital Literacy

Along with reflective pedagogies, the purpose of reflection should also be considered critically. Reflection is an essential aspect of a modern data-driven society. Making sense of complex and conflicting data is a skill required by many jobs. Being able to question biases and assumptions is an important skill that is required for dealing with many modern problems, like fake news and propaganda. In these ways, reflective thinking goes beyond in-the-moment insights. Reflective thinking should be considered more broadly than just reflective activities and reflective experiences. In the same way that is possible to code without engaging in computational thinking or prototyping without engaging in design thinking, it is also likely possible to use RSTs without engaging in reflective thinking. Students should not be expected to reflect on demand, but should instead develop a critical, reflective way of interacting with the world around them.

Some possible ways to encourage reflective thinking are to incorporate design fictions into the classroom or to encourage students to imagine design futures related to their course work. Another way to encourage reflective thinking is to have students take cross-disciplinary courses, like philosophy, which challenge their biases and preconceived notions about the world. Computational thinking and design thinking will not be sufficient in a world that is struggling with problems like misinformation, climate change, and other wicked problems that design thinking is not fully equipped to address on its own. Students need to be capable of thinking for themselves and challenging the status quo to create change in the world.

7.4 Conclusion

In this work, I have demonstrated that data-driven RSTs are an effective way of supporting students in the classroom. Students obtained insights, they increased their awareness, and they made changes that resulted from these insights. Despite these numerous successes, there were many surprising findings that have shed light on how students reflect on their learning, and how students use RSTs. Based on these insights, there is clearly a need for more research that disentangles reflective experiences and reflective practice. Reflection needs to be further operationalized and existing measures and pedagogies that feature reflection should be reconsidered in light of the findings in this dissertation.

REFERENCES

[1] S. Alaoutinen and K. Smolander. Student self-assessment in a programming course using bloom's revised taxonomy. In *Proceedings of the Fifteenth Annual Conference on Innovation and Technology in Computer Science Education*, ITiCSE '10, pages 155–159, New York, NY, USA, 2010. ACM.

[2] A. Alexiou and F. Paraskeva. Enhancing self-regulated learning skills through the implementation of an e-portfolio tool. *Procedia-Social and Behavioral Sciences*, 2(2):3048–3054, 2010.

[3] T. Amiel and T. C. Reeves. Design-based research and educational technology: Rethinking technology and the research agenda. *Journal of educational technology & society*, 11(4):29, 2008.

[4] L. W. Anderson, D. R. Krathwohl, P. Airasian, K. Cruikshank, R. Mayer, P. Pintrich, J. Raths, and M. Wittrock. A taxonomy for learning, teaching and assessing: A revision of bloom's taxonomy. *Artz, AF, & Armour-Thomas, E.(1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. Cognition and Instruction*, 9(2):137–175, 2001.

[5] P. André, M. C. Schraefel, A. Dix, and R. W. White. Expressing well-being online: Towards self-reflection and social awareness. In *Proceedings of the 2011 iConference*, iConference '11, pages 114–121, New York, NY, USA, 2011. ACM.

[6] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, 2012.

[7] C. Argyris. Double-loop learning, teaching, and research. *Academy of Management Learning & Education*, 1(2):206–218, 2002.

[8] C. Argyris and D. A. Schön. *Theory in practice: Increasing professional effectiveness.* Jossey-Bass, 1974.

[9] S. L. Ash and P. H. Clayton. The articulated learning: An approach to guided reflection and assessment. *Innovative Higher Education*, 29(2):137–154, 2004.

[10] A. Ayobi, T. Sonne, P. Marshall, and A. L. Cox. Flexible and mindful self-tracking: Design implications from paper bullet journals. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 28:1–28:14, New York, NY, USA, 2018. ACM.

[11] M. Balaam, G. Fitzpatrick, J. Good, and R. Luckin. Exploring affective technologies for the classroom with the subtle stone. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1623–1632, New York, NY, USA, 2010. ACM.

[12] S. Barab and K. Squire. Design-based research: Putting a stake in the ground. *The journal of the learning sciences*, 13(1):1–14, 2004.

[13] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1505–1512, 2006.

[14] E. P. Baumer. Reflective informatics: Conceptual dimensions for designing technologies of reflection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 585–594, New York, NY, USA, 2015. ACM.

[15] E. P. Baumer, V. Khovanskaya, M. Matthews, L. Reynolds, V. Schwanda Sosik, and G. Gay. Reviewing reflection: on the use of reflection in interactive system design. In *Proceedings of the 2014 conference on Designing interactive systems*, pages 93–102. ACM, 2014.

[16] C. M. Beard and J. P. Wilson. *Experiential learning: A best practice handbook for educators and trainers.* Kogan Page Publishers, 2006.

[17] T. Bergstrom and K. Karahalios. Conversation clock: Visualizing audio patterns in co-located groups. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS 2007)*, pages 78–78. IEEE, 2007.

[18] M. Bhattacharya and M. Hartnett. E-portfolio assessment in higher education. In *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual*, pages T1G–19. IEEE, 2007.

[19] R. Bjork, J. Metcalfe, and A. Shimamura. *Metacognition: Knowing about knowing.* MIT Press, 1994.

[20] B. Bloom, M. Englehart, E. Furst, W. Hill, and D. Krathwohl. Taxonomy of educational objectives: The classification of educational goals. *Handbook I: Cognitive Domain. New York: David McKay*, 1956.

[21] B. Bloom, D. Krathwohl, and B. Masia. *Bloom taxonomy of educational objectives.* Allyn and Bacon, Boston, MA, USA, 1984.

[22] G. M. Bodner. Constructivism: A theory of knowledge. *Journal of chemical education*, 63(10):873, 1986.

[23] T. Borton. *Reach, touch, and teach: Student concerns and process education.* McGraw-Hill, New York, 1970.

[24] D. Boud, R. Keogh, and D. Walker. Promoting reflection in learning: A model. *Reflection: Turning experience into learning*, pages 18–40, 1985.

[25] T. Bourner. Assessing reflective learning. *Education+ training*, 45(5):267–272, 2003.

[26] D. Boyd, H.-Y. Lee, D. Ramage, and J. Donath. Developing legible visualizations for online social spaces. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS 2002)*, pages 1060–1069. IEEE, Jan 2002.

[27] V. Braun and V. Clarke. *Successful qualitative research: A practical guide for beginners*. sage, 2013.

[28] A. K. Brooks. Critical reflection as a response to organizational disruption. *Advances in developing human resources*, 1(3):66–79, 1999.

[29] R. Carol. Defining reflection: Another look at john dewey and reflective thinking. *Teachers College Record*, 104(4):842–866, June 2002.

[30] L. Carter. Why students with an apparent aptitude for computer science don't choose to major in computer science. pages 27–31, 2006.

[31] S. Chandrasegaran, S. K. Badam, L. Kisselburgh, K. Peppler, N. Elmqvist, and K. Ramani. Vizscribe: A visual analytics approach to understand designer behavior. *International Journal of Human-Computer Studies*, 100:66 – 80, 2017.

[32] E. K. Choe. *Designing self-monitoring technology to promote data capture and reflection*. PhD thesis, 2014.

[33] E. K. Choe, B. Lee, M. Kay, W. Pratt, and J. A. Kientz. Sleeptight: Low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 121–132, New York, NY, USA, 2015. ACM.

[34] E. K. Choe, B. Lee, S. Munson, W. Pratt, and J. A. Kientz. Persuasive performance feedback: The effect of framing on self-efficacy. In *AMIA Annual Symposium Proceedings*, volume 2013, page 825. American Medical Informatics Association, 2013.

[35] E. K. Choe, B. Lee, H. Zhu, N. H. Riche, and D. Baur. Understanding self-reflection: How people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth '17, pages 173–182, New York, NY, USA, 2017. ACM.

[36] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[37] J. Cohoon and W. Aspray. *A critical review of the research on women's participation in postsecondary computing education*. Mit Press, 2006.

[38] A. Collins, D. Joseph, and K. Bielaczyc. Design research: Theoretical and methodological issues. *The Journal of the learning sciences*, 13(1):15–42, 2004.

[39] J. Condell, J. Wade, L. Galway, M. McBride, P. Gormley, J. Brennan, and T. Somasundram. Problem solving techniques in cognitive science. *Artificial Intelligence Review*, 34(3):221–234, 2010.

[40] J. O. Cooper, T. E. Heron, W. L. Heward, et al. Applied behavior analysis. 2007.

[41] M. M. Crossan, H. W. Lane, and R. E. White. An organizational learning framework: From intuition to institution. *Academy of management review*, 24(3):522–537, 1999.

[42] C. H. Crouch, J. Watkins, A. P. Fagen, and E. Mazur. Peer instruction: Engaging students one-on-one, all at once. *Research-Based Reform of University Physics*, 1(1):40–95, 2007.

[43] P. Dalsgaard. Experimental systems in research through design. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4991–4996, New York, NY, USA, 2016. ACM.

[44] C. Day. Professional development and reflective practice: purposes, processes and partnerships. *Pedagogy, Culture & Society*, 7(2):221–233, 1999.

[45] N. Dehbozorgi, S. MacNeil, M. L. Maher, and M. Dorodchi. A comparison of lecture-based and active learning design patterns in cs education. In *2018 IEEE Frontiers in Education Conference*, FIE '18. IEEE, Oct 2018.

[46] E. Deitrick, M. H. Wilkerson, and E. Simoneau. Understanding student collaboration in interdisciplinary computing activities. In *Proceedings of the 2017 ACM Conference on International Computing Education Research*, pages 118–126. ACM, 2017.

[47] J. Dewey. How we think. a restatement of the relation of reflective thinking on the educative practice. *Lexington, MA: Heath*, 1933.

[48] J. Donath. A semantic approach to visualizing online conversations. *Commun. ACM*, 45(4):45–49, Apr. 2002.

[49] S. H. Edwards. Using software testing to move students from trial-and-error to reflection-in-action. *SIGCSE Bull.*, 36(1):26–30, Mar. 2004.

[50] P. A. Ertmer and T. J. Newby. Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance Improvement Quarterly*, 26(2):43–71, 2013.

[51] A. Eveleigh, C. Jennett, S. Lynn, and A. L. Cox. I want to be a captain! i want to be a captain! gamification in the old weather citizen science project. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, Gamification '13, pages 79–82, New York, NY, USA, 2013. ACM.

[52] H. Faste and H. Lin. The untapped promise of digital mind maps. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1017–1026. ACM, 2012.

[53] A. Fekete, J. Kay, J. Kingston, and K. Wimalaratne. Supporting reflection in introductory computer science. In *Proceedings of the Thirty-first SIGCSE Technical Symposium on Computer Science Education*, SIGCSE '00, pages 144–148, New York, NY, USA, 2000. ACM.

[54] A. Fessl, O. Blunk, M. Prilla, and V. Pammer. The known universe of reflection guidance: a literature review. *International Journal of Technology Enhanced Learning*, 9(2-3):103–125, 2017.

[55] L. D. Fink. A self-directed guide to designing courses for significant learning. *University of Oklahoma*, 27:p11, 2003.

[56] J. H. Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906, 1979.

[57] R. Fleck and G. Fitzpatrick. Teachers' and tutors' social reflection around sensecam images. *International Journal of Human-Computer Studies*, 67(12):1024–1036, Dec 2009.

[58] R. Fleck and G. Fitzpatrick. Reflecting on reflection: Framing a design landscape. In *Proceedings of the 22Nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*, OZCHI '10, pages 216–223, New York, NY, USA, 2010. ACM.

[59] J. L. Fleiss, J. Cohen, and B. S. Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323, 1969.

[60] J. Fook, S. White, F. Gardner, et al. Critical reflection: a review of contemporary literature and understandings. *Critical Reflection in Health and Social Care*, 3:20, 2006.

[61] C. Frayling. Research in art and design. *Royal College of Art Research Papers*, pages 1–5, 1993.

[62] P. Freire and D. Macedo. *Pedagogy of the Oppressed: 50th Anniversary Edition.* Bloomsbury Publishing, 2018. Reprint of 1970 Version.

[63] U. Fuller, C. G. Johnson, T. Ahoniemi, D. Cukierman, I. Hernán-Losada, J. Jackova, E. Lahtinen, T. L. Lewis, D. M. Thompson, C. Riedesel, and E. Thompson. Developing a computer science-specific learning taxonomy. In *Working Group Reports on ITiCSE on Innovation and Technology in Computer Science Education*, ITiCSE-WGR '07, pages 152–170, New York, NY, USA, 2007. ACM.

[64] H. E. Gardner. *Multiple intelligences: New horizons in theory and practice.* Basic books, 2008.

[65] W. W. Gaver, A. Boucher, S. Pennington, and B. Walker. Cultural probes and the value of uncertainty. *interactions*, 11(5):53–56, Sept. 2004.

[66] S. E. George. *Learning and the Reflective Journal in Computer Science*, volume 24. IEEE Computer Society Press, Los Alamitos, CA, USA, Jan. 2002.

[67] M. N. Giannakos, J. Krogstie, and N. Chrisochoides. Reviewing the flipped classroom research: Reflections for computer science education. In *Proceedings of the Computer Science Education Research Conference*, CSERC '14, pages 23–29, New York, NY, USA, 2014. ACM.

[68] G. Gibbs. *Learning by doing: A guide to teaching and learning methods.* Oxford Centre for Staff and Learning Development, Oxford Brookes University, 1988.

[69] J. Goode. If you build teachers, will students come? the role of teachers in broadening computer science learning for urban youth. *Journal of Educational Computing Research*, 36(1):65–88, 2007.

[70] P. Goodyear. Educational design and networked learning: Patterns, pattern languages and design practice. *Australasian journal of educational technology*, 21(1):82–101, 2005.

[71] S. Govaerts, K. Verbert, E. Duval, and A. Pardo. The student activity meter for awareness and self-reflection. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, pages 869–884, New York, NY, USA, 2012. ACM.

[72] A. Hadwin and M. Oshige. Self-regulation, coregulation, and socially shared regulation: Exploring perspectives of social in self-regulated learning theory. *Teachers College Record*, 113(2):240–264, 2011.

[73] J. H. Hansen and T. Hasan. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal processing magazine*, 32(6):74–99, 2015.

[74] E. Hargittai and A. Hinnant. Digital inequality differences in young adults' use of the internet. *Communication Research*, 35(5):602–621, 2008.

[75] E. Heinrich, M. Bhattacharya, and R. Rayudu. Preparation for lifelong learning using eportfolios. *European Journal of Engineering Education*, 32(6):653–663, 2007.

[76] P. Heslop, A. Preston, A. Kharrufa, M. Balaam, D. Leat, and P. Olivier. Evaluating digital tabletop collaborative writing in the classroom. In *Human-Computer Interaction*, pages 531–548. Springer, 2015.

[77] C. P. Hoadley. Creating context: Design-based research in creating and understanding cscl. In *Proceedings of the conference on computer support for collaborative learning: Foundations for a CSCL community*, pages 453–462. International Society of the Learning Sciences, 2002.

[78] V. Hobbs. Faking it or hating it: can reflective practice be forced? *Reflective practice*, 8(3):405–417, 2007.

[79] D. Holman, K. Pavlica, and R. Thorpe. Rethinking kolb's theory of experiential learning in management education: The contribution of social constructionism and activity theory. *Management Learning*, 28(2):135–148, 1997.

[80] T. A. Jenkin. Extending the 4i organizational learning model: information sources, foraging processes and tools. *Administrative Sciences*, 3(3):96–109, 2013.

[81] C. G. Johnson and U. Fuller. Is bloom's taxonomy appropriate for computer science? In *Proceedings of the 6th Baltic Sea conference on Computing education research: Koli Calling 2006*, pages 120–123. ACM, 2006.

[82] S. D. Kauer, S. C. Reid, A. H. D. Crooke, A. Khor, S. J. C. Hearps, A. F. Jorm, L. Sanci, and G. Patton. Self-monitoring using mobile phones in the early stages of adolescent depression: randomized controlled trial. *Journal of medical Internet research*, 14(3), 2012.

[83] S. Kennedy-Clark. Research by design: Design-based research and the higher degree research student. *Journal of Learning Design*, 6(2):26–32, 2013.

[84] Y.-H. Kim, J. H. Jeon, E. K. Choe, B. Lee, K. Kim, and J. Seo. Timeaware: Leveraging framing effects to enhance personal productivity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 272–283, New York, NY, USA, 2016. ACM.

[85] P. M. King and K. S. Kitchener. *Developing Reflective Judgment: Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults. Jossey-Bass Higher and Adult Education Series and Jossey-Bass Social and Behavioral Science Series.* ERIC, 1994.

[86] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li. Voice activity detection using mfcc features and support vector machine. In *Int. Conf. on*

Speech and Computer (SPECOM07), Moscow, Russia, volume 2, pages 556–561, 2007.

[87] D. A. Kolb. *Experiential learning: Experience as the source of learning and development*. FT press, 2014.

[88] D. A. Kolb and M. S. Plovnick. The experiential learning theory of career development. 1974.

[89] J. Kopp. Self-monitoring: A literature review of research and practice. In *Social Work Research and Abstracts*, volume 24, pages 8–20. Oxford University Press, 1988.

[90] D. R. Krathwohl. A revision of blooms̀ taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

[91] B. R. Krogstie, M. Prilla, and V. Pammer. Understanding and supporting reflective learning processes in the workplace: The csrl model. In *European conference on technology enhanced learning*, pages 151–164. Springer, 2013.

[92] B. R. Krogstie, M. Prilla, D. Wessel, K. Knipfer, and V. Pammer. Computer support for reflective learning in the workplace: A model. In *2012 IEEE 12th International Conference on Advanced Learning Technologies*, pages 151–153, July 2012.

[93] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[94] C. Latulipe, N. B. Long, and C. E. Seminario. Structuring flipped classes with lightweight teams and gamification. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, pages 392–397. ACM, 2015.

[95] C. Latulipe, S. MacNeil, and B. Thompson. Evolving a data structures class toward inclusive success. In *2018 IEEE Frontiers in Education Conference (FIE 2018)*, San Jose, USA, Oct. 2018.

[96] V. R. Lee. The quantified self (qs) movement and some emerging opportunities for the educational technology field. *Educational Technology*, pages 39–42, 2013.

[97] H. M. Lefcourt. Internal versus external control of reinforcement: A review. *Psychological bulletin*, 65(4):206, 1966.

[98] K. Lewin et al. Field theory in social science. 1951.

[99] C. M. Lewis, R. E. Anderson, and K. Yasuhara. I don't code all day: Fitting in computer science when the stereotypes don't fit. In *Proceedings of the 2016 ACM conference on international computing education research*, pages 23–32. ACM, 2016.

[100] C. M. Lewis and N. Shah. How equity and inequity can emerge in pair programming. In *Proceedings of the Eleventh Annual International Conference on International Computing Education Research*, ICER '15, pages 41–50, New York, NY, USA, 2015. ACM.

[101] I. Li, A. Dey, and J. Forlizzi. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 557–566, New York, NY, USA, 2010. ACM.

[102] I. Li, A. Dey, J. Forlizzi, K. Höök, and Y. Medynskiy. Personal informatics and hci: Design, theory, and social implications. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 2417–2420, New York, NY, USA, 2011. ACM.

[103] S. MacNeil, M. Dorodchi, and N. Deborghzi. Using spectrums and dependency graphs to model progressions from introductory to capstone courses. In *2017 IEEE Frontiers in Education Conference*, FIE '17. IEEE, 2017.

[104] S. MacNeil, K. Kiefer, D. Takle, B. Thompson, and C. Latulipe. Ineqdetect: Visualizing students' conversations to increase awareness and support reflection. In *Proceedings of the Global Computing Education Conference*, CompEd 2019, New York, NY, USA, 2019. ACM.

[105] S. MacNeil, C. Latulipe, B. Long, and A. Yadav. Exploring lightweight teams in a distributed learning environment. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, SIGCSE '16, pages 193–198, New York, NY, USA, 2016. ACM.

[106] S. MacNeil, C. Latulipe, and A. Yadav. Learning in distributed low-stakes teams. In *Proceedings of the Eleventh Annual International Conference on International Computing Education Research*, ICER '15, pages 227–236, New York, NY, USA, 2015. ACM.

[107] S. MacNeil, J. Okerlund, and C. Latulipe. Dimensional reasoning and research design spaces. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, C&C '17, pages 367–379, New York, NY, USA, 2017. ACM.

[108] S. Mann. Lifeglogging cameras, Feb 2013.

[109] J. Margolis, R. Estrella, J. Goode, J. J. Holme, and K. Nao. *Stuck in the shallow end: Education, race, and computing.* MIT Press, 2010.

[110] R. Martínez, A. Collins, J. Kay, and K. Yacef. Who did what? who said that?: Collaid: An environment for capturing traces of collaborative learning at the tabletop. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*, ITS '11, pages 172–181, New York, NY, USA, 2011. ACM.

[111] J. E. McGrath. *Groups: Interaction and performance*, volume 14. Prentice-Hall Englewood Cliffs, NJ, 1984.

[112] P. Mendels, J. Frens, and K. Overbeeke. Freed: A system for creating multiple views of a digital collection during the design process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1481–1490, New York, NY, USA, 2011. ACM.

[113] J. Metcalfe. Metacognitive processes. In *Memory*, pages 381–407. Elsevier, 1996.

[114] J. Mezirow. *Transformative dimensions of adult learning.* ERIC, 1991.

[115] J. Mezirow. On critical reflection. *Adult education quarterly*, 48(3):185–198, 1998.

[116] M. M. Mukaka. A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3):69–71, 2012.

[117] R. O. Nelson and S. C. Hayes. Theoretical explanations for reactivity in self-monitoring. *Behavior Modification*, 5(1):3–14, 1981.

[118] B. Niemierko. Taxonomies of educational goals as a lead into creative teacher training. *Polish Journal of Social Science*, 4(1):93–106, 2009.

[119] J. W. Norman. *A comparison of tendencies in secondary education in England and the United States.* Number 119. Teachers College, Columbia University, 1922.

[120] M. T. Orne. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist*, 17(11):776, 1962.

[121] M. T. Orne. Demand characteristics and the concept of quasi-controls. *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnows Classic Books*, 110:110–137, 2009.

[122] R. O. Otienoh. Reflective practice: The challenge of journal writing. *Reflective practice*, 10(4):477–489, 2009.

[123] E. Panadero. A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8:422, 2017.

[124] E. Patitsas, M. Craig, and S. Easterbrook. A historical examination of the social factors affecting female participation in computing. In *Proceedings of the 2014 Conference on Innovation & Technology in Computer Science Education*, ITiCSE '14, pages 111–116, New York, NY, USA, 2014. ACM.

[125] R. L. Pecheone, M. J. Pigg, R. R. Chung, and R. J. Souviney. Performance assessment and electronic portfolios: Their effect on teacher learning and education. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 78(4):164–176, 2005.

[126] J. Piaget and M. Cook. *The origins of intelligence in children*, volume 8. International Universities Press New York, 1952.

[127] J. Pirker, M. Riffnaller-Schiefer, and C. Gütl. Motivational active learning: Engaging university students in computer science education. In *Proceedings of the 2014 Conference on Innovation & Technology in Computer Science Education*, ITiCSE '14, pages 297–302, New York, NY, USA, 2014. ACM.

[128] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, volume 5, pages 2–4, 2005.

[129] A. Rapp and M. Tirassa. Know thyself: a theory of the self for personal informatics. *Human–Computer Interaction*, 32(5-6):335–380, 2017.

[130] T. Reeves. Design research from a technology perspective. In *Educational design research*, pages 64–78. Routledge, 2006.

[131] T. C. Reeves. Pseudoscience in computer-based instruction: The case of learner control research. *Journal of computer-based instruction*, 20(2):39–46, 1993.

[132] T. W. Rinker. *sentimentr: Calculate Text Polarity Sentiment.* Buffalo, New York, 2018. version 2.6.1.

[133] H. W. Rittel and M. M. Webber. Dilemmas in a general theory of planning. *Policy sciences*, 4(2):155–169, Jun 1973.

[134] M. J. Rodríguez-Triana, L. P. Prieto, A. Vozniuk, M. S. Boroujeni, B. A. Schwendimann, A. Holzer, and D. Gillet. Monitoring, awareness and reflection in blended technology enhanced learning: a systematic review. *International Journal of Technology Enhanced Learning*, 9(2-3):126–150, 2017.

[135] G. Rolfe, D. Freshwater, and M. Jasper. *Critical reflection for nursing and the helping professions: A user's guide.* Palgrave Basingstoke, 2001.

[136] J. B. Rotter. Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied*, 80(1):1, 1966.

[137] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 269–276. ACM, 1993.

[138] J. Saldaña. *The coding manual for qualitative researchers.* Sage, 2015.

[139] K. Sanders, J. Boustedt, A. Eckerdal, R. McCartney, and C. Zander. Folk pedagogy: Nobody doesn't like active learning. In *Proceedings of the 2017 ACM Conference on International Computing Education Research*, ICER '17, pages 145–154, New York, NY, USA, 2017. ACM.

[140] B. Schmitz and B. S. Wiese. New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. *Contemporary educational psychology*, 31(1):64–96, 2006.

[141] D. A. Schön. *The reflective practitioner: How professionals think in action*, volume 5126. Basic books, 1983.

[142] G. Schraw and D. Moshman. Metacognitive theories. *Educational psychology review*, 7(4):351–371, 1995.

[143] C. Schulte and M. Knobelsdorf. Attitudes towards computer science-computing experiences as a starting point and barrier to computer science. In *Proceedings of the Third International Workshop on Computing Education Research*, ICER '07, pages 27–38, New York, NY, USA, 2007. ACM.

[144] D. H. Schunk and B. J. Zimmerman. Social origins of self-regulatory competence. *Educational psychologist*, 32(4):195–208, 1997.

[145] L. A. Sepp, M. Orand, J. A. Turns, L. D. Thomas, B. Sattler, and C. J. Atman. On an upward trend. In *2015 122nd ASEE Annual Conference and Exposition*. American Society for Engineering Education, 2015.

[146] O. Shaer, M. Strait, C. Valdes, T. Feng, M. Lintz, and H. Wang. Enhancing genomic learning through tabletop interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2817–2826, New York, NY, USA, 2011. ACM.

[147] M. Sharmin and B. P. Bailey. Reflectionspace: An interactive visualization tool for supporting reflection-on-action in design. In *Proceedings of the 9th ACM Conference on Creativity & Cognition*, C&C '13, pages 83–92, New York, NY, USA, 2013. ACM.

[148] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–, Washington, DC, USA, 1996. IEEE Computer Society.

[149] J. Silge and D. Robinson. tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, 1(3), 2016.

[150] S. Skach, P. G. Healey, and R. Stewart. Talking through your arse: Sensing conversation with seat covers. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society (CogSci 2017)*, pages 3186–3190, 2017.

[151] L. Soler, S. Zwart, M. Lynch, and V. Israel-Jost. *Science after the practice turn in the philosophy, history, and social studies of science*, volume 14. Routledge, Mar 2014.

[152] D. S. Spiegel. *Coterie: A visualization of the conversational dynamics within IRC*. PhD thesis, Massachusetts Institute of Technology, 2001.

[153] A. Stankiewicz and C. Kulkarni. $1 conversational turn detector: Measuring how video conversations affect student learning in online classes. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, L@S '16, pages 81–88, New York, NY, USA, 2016. ACM.

[154] P. Stappers and E. Giaccardi. Research through design. *The Encyclopedia of Human-Computer Interaction, 2nd Edition*, pages 1–94, 2017.

[155] J. A. Stone. Using reflective blogs for pedagogical feedback in cs1. In *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education*, SIGCSE '12, pages 259–264, New York, NY, USA, 2012. ACM.

[156] J. Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[157] E. Thompson, A. Luxton-Reilly, J. L. Whalley, M. Hu, and P. Robbins. Bloom's taxonomy for cs assessment. In *Proceedings of the tenth conference on Australasian computing education-Volume 78*, pages 155–161. Australian Computer Society, Inc., 2008.

[158] R. Tong, B. Ma, K.-A. Lee, C. You, D. Zhu, T. Kinnunen, H. Sun, M. Dong, E. S. Chng, and H. Li. The iir nist 2006 speaker recognition system: Fusion of acoustic and tokenization features. In *presentation in 5th Int. Symp. on Chinese Spoken Language Processing, ISCSLP*, 2006.

[159] M. Van Woerkom. The concept of critical reflection and its implications for human resource development. *Advances in developing human resources*, 6(2):178–192, 2004.

[160] L. Vygotsky. Interaction between learning and development. *Readings on the development of children*, 23(3):34–41, 1978.

[161] B. J. Wadsworth. *Piaget's theory of cognitive and affective development: Foundations of constructivism*. Longman Publishing, 1996.

[162] L. L. Werner, B. Hanks, and C. McDowell. Pair-programming helps female computer science students. *Journal on Educational Resources in Computing (JERIC)*, 4(1):4, 2004.

[163] G. Wiggins and J. McTighe. *Understanding by Design*. ASCD, 2005.

[164] J. M. Wing and M. Guzdial. Cs woes: Deadline-driven research, academic inequality. *Commun. ACM*, 52(12):8–9, Dec. 2009.

[165] R. Xiong and J. Donath. Peoplegarden: Creating data portraits for users. In *Proceedings of the 12th Annual ACM Symposium on User Interface Software and Technology*, UIST '99, pages 37–44, New York, NY, USA, 1999. ACM.

[166] T. Yamada, S. Nakamura, and K. Shikano. Robust speech recognition with speaker localization by a microphone array. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1317–1320. IEEE, 1996.

[167] Y.-C. G. Yen, S. P. Dow, E. Gerber, and B. P. Bailey. Listen to others, listen to yourself: Combining feedback review and reflection to improve iterative design. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, C&C '17, pages 158–170, New York, NY, USA, 2017. ACM.

[168] B. J. Zimmerman and D. H. Schunk. *Self-regulated learning and academic achievement: Theory, research, and practice.* Springer Science & Business Media, 2014. Reprint of 1989 Version.

[169] J. Zimmerman, J. Forlizzi, and S. Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 493–502, New York, NY, USA, 2007. ACM.

APPENDIX A: REFLECTION PROMPTS BY SESSION FOR BLOOMMATRIX

Table 12: An overview of when each reflection prompt was given in each of the BloomMatrix classes. Standard reflection prompts are highlighted in bold. The number of non-blank reflections submitted for each prompt is shown in the last column.

Session	**Reflection Prompt**	**n**
HCI-1	Reflection in design is often associated with thinking about products..	18
HCI-1	Reflection is a part of everyday life for designers. It is a way to impr..	19
HCI-2	Take a few minutes to look at indeed com for your dream job What ...	28
HCI-2	This summer course is condensed into 5 weeks. It is going to be a ...	27
HCI-2	What are you hoping to learn in this class? How do you hope to grow ...	27
HCI-3	How did the interactions with your team members go today? Did you ...	28
HCI-3	If you think of your brain as a person, what was it doing during the ...	27
HCI-3	What did you learn from this experience?	28
HCI-3	**What did you do today? What was the purpose?**	**28**
HCI-3	**What went well? What can be improved?**	**28**
HC-6	What did you learn from this experience?	26
HC-6	What have you learned this semester?	27
HCI-6	**What did you do today? What was the purpose?**	**28**
HCI-6	**What went well? What can be improved?**	**28**
AI-1	How do you imagine society might improve because of AI?	24
AI-1	In what ways do you think you will be able to incorporate AI or ...	24
AI-1	This class is condensed into 5 weeks What are some strategies you ...	24
AI-1	What aspect of your own life do you currently reflect on (Religion, ...	24
AI-1	What negative effects do you imagine AI might have on society ...	24
AI-2	Describe some ways in which your brain is similar to an AI agent ...	21
AI-2	How do you remember things? Why do you think that you forget?	21
AI-2	How does your brain work? What happens when you're thinking?	21
AI-2	What do these similarities and differences cue you to think about?	21
AI-3	How can you better help your brain to learn more effectively?	22
AI-3	If you think of your brain as a person what was it doing during the ...	22
AI-3	What did you learn from this experience?	22
AI-3	**What did you do today? What was the purpose?**	**22**
AI-6	What did you learn from this experience?	21
AI-6	What have you learned this semester?	21
AI-6	**What did you do today? What was the purpose?**	**18**
AI-6	What went well? What can be improved?	20

APPENDIX B: REFLECTION PROMPTS BY SESSION FOR INEQDETECT

Table 13: An overview of when each reflection prompt was given in the GDD class that used IneqDetect. Standard reflection prompts are highlighted in bold. The number of non-blank reflections submitted for each prompt is shown in the last column.

Session	Reflection Prompts	n
GD-1	Designers reflect on users to understand users needs and ...	6
GD-2	Take a few minutes to think about your dream job. What ...	10
GD-2	This summer course is condensed into 5 weeks. It is going ...	10
GD-2	What are you hoping to learn in this class? How do you hope ...	10
GD-2	What games do you like? What about those games makes you ..	10
GD-3	How did the interactions with your team members go today? ...	9
GD-3	If you think of your brain as a person, what was it doing ...	9
GD-3	What did you learn from this experience?	9
GD-3	**What did you do today? What was the purpose?**	**9**
GD-3	**What went well? What can be improved?**	**9**
GD-4	How did the conversation with your team members go today? ...	14
GD-4	What did you do today in your group?	14
GD-4	What went well in your group? What can be improved?	14
GD-5	How did the interactions with your team members go today? ...	7
GD-5	If you think of your brain as a person, what was it doing ...	7
GD-5	What did you learn from this experience?	7
GD-5	**What did you do today? What was the purpose?**	**7**
GD-5	**What went well? What can be improved?**	**7**
GD-6	What did you learn from this experience?	8
GD-6	What have you learned this semester?	8
GD-6	**What did you do today? What was the purpose?**	**8**
GD-6	**What went well? What can be improved?**	**8**

Table 14: An overview of when each reflection prompt was given in the HCI class that used IneqDetect. Standard reflection prompts are highlighted in bold. The number of non-blank reflections submitted for each prompt is shown in the last column.

Session	**Reflection Prompts**	**n**
HCI-1	Designers often reflect on existing products and designs ...	10
HCI-1	Designers reflect on users to understand users needs and ...	10
HCI-1	Last week we discussed color as an example of something ...	10
HCI-1	Why have you chosen to take this class? What do you hope ...	10
HCI-2	Describe the roles that you typically fulfill within a ...	10
HCI-2	Describe your communication style and how you ...	10
HCI-2	What is the most important part of a successful team? ...	10
HCI-2	What would you change about the way that you ...	10
HCI-2	Why do you think that you communicate the way that you do?	10
HCI-3	Describe the interactions with your team members today? ..	11
HCI-3	How could your team interaction improve going forward? ...	11
HCI-3	In other groups do you talk more or less? Why do you ...	11
HCI-3	What did you learn from this experience?	9
HCI-3	**What did you do today? What was the purpose?**	**11**
HCI-3	**What went well in your group? What can be improved?**	**7**
HCI-4	Do you communicate differently in this group than in ...	7
HCI-4	How did the conversation with your team members go today? ...	7
HCI-5	Describe the interactions with your team members today...	8
HCI-5	How could your team interaction improve going ...	8
HCI-5	In other groups do you talk more or less? Why do ...	8
HCI-5	What did you learn from this experience?	8
HCI-5	**What did you do today? What was the purpose?**	**8**
HCI-6	What have you learned this semester?	7
HCI-6	**What did you do today? What was the purpose?**	**7**
HCI-6	**What went well What can be improved?**	**6**
HCI-6	What did you learn from this experience?	6