Stakeholders' Interpretations of Data for Equitable Computing Education

Benjamin Xie

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Committee:

Amy J. Ko, Chair

Jason C. Yip

Jevin West

Jennifer Mankoff

Program Authorized to Offer Degree:

Information Science

University of Washington

Abstract

Stakeholders' Interpretations of Data for Equitable Computing Education

Benjamin Xie

Chair of the Supervisory Committee:
Professor Amy J. Ko
The Information School

Computing education has growing inclusion and equity challenges (e.g. exclusionary online learning experiences, biased assessments, inadequate student feedback mechanisms). Many groups experience minoritization in computing education, including students who are Black, Indigenous, and people of color (BIPOC), women, non-binary students, transfer students, international students, first-generation students, and students with disabilities. To ensure diverse students can realize their potential to participate in and challenge computing communities, we must enable stakeholders (e.g. students, teachers, curriculum designers) to take informed, timely, and equitable actions. This dissertation explores how to design interactions with data to inform stakeholders in support of such actions.

While data often perpetuates and exacerbates inclusion and equity challenges when improperly used, it can also support equity-oriented goals if properly contextualized for interpretation by stakeholders. I explored how stakeholders interpreted data in three contexts: 1) informing students of what to learn next in an adaptive, self-directed online learning experience; 2) informing curriculum designers with empirical evidence of assessment bias; 3) and informing teaching teams of inequities using contextualized student feedback. Through these studies, I identified how stakeholders' relationships with educational structures and systems impacted their interpretations of data for equity-oriented goals. These factors have implications to the research and practice of learning at scale, computing education, and human-computer interaction. Therefore, I claim the following

thesis statement:

Interactions with data that consider prior knowledge, perceptions of power relationships, and cultural competency can enable computing education stakeholders to connect their interpretations of data with their domain expertise in service of equity-oriented goals.

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

AGENCY: A learner's capacity to define and pursue learning goals [14]. Learners may exert agency by choosing to use one tool to learn over another, to study a specific topic, to try a certain exercise, to review some instruction, or to decide they know enough and quit. Exerting agency is required to make decisions.

BIPOC (BLACK, INDIGENOUS, AND PEOPLE OF COLOR): A commonly used phrase and acronym that refer to ethnic groups that computing communities and societal structures often bias or disadvantage against. While popular in usage, this terminology is not without problems. In particular terms such as *people of color* can be considered harmful [301] and often ambiguous. For example, it is ambiguous whether BIPOC includes people who are Asian and Pacific-Islander. I use BIPOC to prioritize common understanding and familiarity without further explanation (e.g. when showing teaching teams demographic data with StudentAmp in Chapter 5). But I prefer to use more specific language when there are opportunities for stakeholders to become familiar with the term. For example, I refer to African/Black, Hispanic/Latinx, Native/Indigenous, and Pacific-Islander/Alaskan Native (AHNP) to refer to ethnic groups that are often minoritized when working directly with curriculum designers to interpret data on DIF in Chapter 4.

BIAS: Systematic favoring of certain (often dominant) groups over other (often minoritized) in a socially situated context. Types of bias include *preexisting biases* that have roots in social institutions, practices, and attitudes; *technical biases* which arise from technical constraints or design decisions; and *emergent biases* which arise in a context of use [113].

CRITICAL: An stance that seeks to understand how dominant ideologies become infused in social norms, ultimately allowing systems and ideologies of oppression to occur [273].

COMPUTING EDUCATION RESEARCH: How people learn and teaching computing, broadly construed [162]. Within this dissertation, I investigate include data science within the domain of computing education and investigate it at the K-12 (secondary) and university (post-secondary) level.

DATA: Artificial constructs that reflect decisions and biases of people who created and use them [305]. Within this dissertation, data I consider include log data of learners' past actions on

an online learning system, performance on assessments, student demographics, challenges students report, and students' perceptions of classmates' challenges.

DISPARITY: Unjust or unfair differences, typically implying the need to address these differences. This dissertation considers educational disparities, which may result from differential or biased treatment of students from minoritized groups, differences in preparatory privilege, and different responses to educational systems or different sets of educational needs [234].

DIVERSITY: Which demographic groups are and are not represented or included in various spaces and practices [172]

EQUITY: Improving access and supporting successful participation and achievement of diverse students learning computing [172]. Whereas equality might require equal access, equity requires equalizing outcomes [274]. The may necessitate unequal inputs to addressing disparities that arise from structures and norms failing to include or serve students of minoritized groups.

INTERPRETATION (OF DATA): Making sense of data through a complex intersection of implicit beliefs that reflect broader social discourses [23]. In this dissertation, I consider how prior knowledge, perceived power relationships, and cultural competence affect the formation of stakeholders' beliefs and experiences in the process of interpreting and making sense of data to consider equity-oriented actions.

LEARNING: A process in which learners can realize their human dignity and potential (to participate in existing systems, challenge oppressions) without enduring a process of dehumanization. Learning computing provides learners with an opportunity to engage with powerful tools that they can use to unseat individual and collective social oppressions. Adapted from [282].

MINORITIZED: Groups that are not positively privileged or favored and often stigmatized. In computing education, they include students who are women, non-binary, African-American/Black, Hispanic/Latinx, Native American/Indigenous, Pacific Islander, transfer students, not fluent in English, and/or first-generation, as well as students who have disabilities and/or have financial or familial responsibilities. Systemic cultures and norms tend to favor dominant groups and disadvantage minoritized groups.

STAKEHOLDER: Any group or individual who can affect or is affected by the achievement of objectives related to computing education. Stakeholders are defined by and understood in relationship to their interaction with a technology or sociotechnical system [112]. Adopted from the framing of stakeholders from Value Sensitive Design [90, 112].

VALIDITY: An evaluation of coherence and completeness of an argument for a given stakeholder to interpret and use data within a given sociopolitical context. This is a situated framing of Kane's framing of validity in psychometrics literature [148, 149, 150].

ACKNOWLEDGMENTS

Firstly, I acknowledge the Coast Salish peoples of this land, the land which touches the shared waters of all tribes and bands within the Duwamish, Puyallup, Suquamish, Tulalip and Muckleshoot nations.

Thank you to my girlfriend, Nicole. I'm difficult. Our PhDs are challenging. And it's frustrating because I want to follow my instinct of working and making myself more work more. But you bring me back to a reality where my life is more than what I do in front of a keyboard. So thanks for leaning in every time life became overwhelming. Growing together during our PhDs was enriching, and the next chapters are going to be a thrill. And thanks for convincing me to adopt our rescue dog Curie. She is definitely the "cutest, sweetest, smartest, spunkiest" dog ever.

Thank you to my mom and dad. I don't think my research is exactly what any of us expected me to do as a career. But I am grateful your support as I ponder who I am and who I want to be. And to the rest of my family: Thanks for showing me since childhood that humans just being humans is good enough.

Infinite gratitude to my advisor Dr. Amy Ko. I decided to do a PhD because I thought you would support me as a human being. I just didn't realize how one human could provide such insight, mentorship, and support for me, our labmates, and so many others. All the while demonstrating how to be true to yourself. Academia often trends towards being inhumane and exploitative, but you have challenged the norms by fostering a student-centric community to make my PhD experience fulfilling both professionally and personally. It's been a privilege to spend these past six(!) years learning from you and honing the "what would Amy do?" voice in my head. #amykoeffect.

And to my labmates: I got paid to collaborate and learn with you all, and that's special. The PhD is characterized by struggles, triumphs, and bewilderment. And I got to share all that through

our problematization of every institution we're aware of, birth of new ideas in whiteboard sessions, and Slack backchanneling. In a PhD that became more distant due to pandemics, you all were consistent community I felt I always belonged with.

Thank you to the iSchool and DUB communities for not only accepting me and my amoeba-shaped research ideas, but also putting perhaps blind faith in me to organize and run events. Deep down, I just like connecting people so we can have shared experiences. So thank you for letting me do that.

To the mentors I've had along the way: Annie Ross, David Weintrop, Eric Klopfer, Hal Abelson, Jason Yip, Jevin West, Justin Reich, Ken Holstein, Motahhare Eslami, Mark Guzdial, Philip Guo, R. Ben Shapiro, Ravi Karkar, Sayamindu Dasgupta, and so many others. Thanks for not only sharing knowledge with me, but also sharing such infectious excitement, all the while keeping things candid and real.

To my friends: Thank you for persisting despite me dodging messages and bailing on hangouts. Your never-ending nudges to have me join your adventures always bring joy and presence to my life.

And thank you to everyone who does the supporting work that I rely on. People like Dr. Andrea Salazar for years of conversations that helped me discover myself and handle trauma from the past and present; Dora Tkach for magically replenishing my bank account after conference expenses; folks at iSchool IT for getting me another laptop after my first one mysteriously disappeared; people in the IRB office who ensure my research does no harm.

Finally, I acknowledge the countless that who have been harmed by academia because of rampant abuse and sexual harassment. Quoting Aida Behmard [18]:

> *Why do we bemoan the exit of sexual harassers, and those who otherwise harm their colleagues, from the scientific community? Instead, we should mourn the loss of all the promising scientists that they forced out, whose contributions will never be known.*

Learn more about sexual harassment in academia: `https://www.belowthewaterline.org/`.

DEDICATION

To Nicole, my adventure buddy.

Chapter 1

INTRODUCTION: OPPORTUNITIES AND RISKS OF WIELDING DATA

The representations we make up often take on realities of their own. In the 1930s, Otto G. Lindberg and Ernest Alpers of General Drafting Co. were creating a road map of New York state [213]. To prevent competing companies from copying their maps, they created the fictitious place of "Agloe," as shown in Figure 1.1A. The idea was that if anybody else produced a map with Agloe on it, Lindberg could sue them for copying their map. Fast-forward two decades and sure enough, the map company Rand McNally produced a New York state map that included Agloe. But when Lindberg tried to sue for copyright infringement, Rand McNally lawyers defended themselves by saying that Agloe actually *did* exist. Because somebody had seen Agloe on a map, realized nothing was there, and built the Agloe General Store (Fig 1.1B). And while nothing exists at that location after the Agloe General Store closed decades ago, Algoe appeared on road maps as recently as the 1990s, and on the United States Geological Survey (USGS) Geographic Names Information System and Google Maps in 2014. The made-up data that was Agloe, NY took on a reality of its own.

Data are powerful not just because they are abstract representations of reality, but also because they take on realities of their own. The fake location of Agloe is an innocuous example of this phenomenon. But when data relate to people and their wellbeing, the stakes are higher. We have seen that the decisions we make when we produce, sample, analyze, model, interpret, and use data lead to a "coded gaze" where the views of the select few who have the power to develop systems propagate throughout society [37]. As a result, many groups find themselves being excluded in a data-defined society. We have already seen examples of data exclusion and the consequences: The 2020 US census only asks about biological gender, excluding non-binary, trans, and gender non-conforming people from being considered in government decision-making; facial recognition

Figure 1.1: How a representation becomes a reality of its own: The fictitious town of Agloe, NY was originally created to protect a map from copyright infringement (A). But then it became a reality (B). ©Booklist/ American Library, Joyce Conroy

datasets are predominantly of white men, resulting in diminished classification accuracy for darker skin and the false arrest of a Black man in Detroit [217, 242]; comparisons of academic performance by race (typically with white, non-Hispanic students as the baseline) cannot consider contextual factors that impact achievement, resulting in a deficit framing for students who are Black, Indigenous, and People of Color (BIPOC).

1.1 Context of study: equity in computing education

Within the context of computing education, a boom in interest and enrollment resulted in the use of more scalable data-driven technologies to support learning experiences. Examples include online learning platforms to make remote learning more feasible [298], intelligent tutoring systems that use data from other students' performance to personalize and adapt learning experiences [304, 144], and auto-graders to make evaluating assessments more efficient [73]. But these learning experiences are often either standardized to serve the majority or trained on data from students of dominant identity groups (e.g. able white and Asian men). As a result, these experiences will typically fail to serve and even harm students in minoritized groups, groups that have

been excluded or isolated because of societal structures (e.g. systemic racism, exclusions, oppression) [238, 239]. Examples of minoritized groups include Black, Indigenous, and People of Color (BIPOC), LGBTQ+, and people with physical, cognitive, and social disabilities. So while using data-driven tools can help scale learning and engage a broader audience, it can also exclude, harm, and oppress learners in hidden ways.

Formal education in computing is the primary pathway for participation in the computing community, yet formal computing courses face persistent diversity, equity, and inclusion issues [143, 296]. In part due to the growing demand for computing skills in the workforce, enrollment numbers in computing majors have surged recently, straining the capacity of instructors to scale teaching [209]. Despite their popularity, computing courses still face challenges retaining and supporting diverse students in both high school [94] and college [278, 296, 319]. This results in the loss of diverse potential contributors to the computing field [209] and raises social justice issues around who can access and engage with computing communities [163, 257, 258].

Improving equity would require not just improving access to computing education, but also supporting successful participation and achievement by diverse students learning computing [172, 105]. Structural and systemic inequities embedded in and around computing courses can manifest as barriers to participation (e.g. unconscious bias of instructors excluding students of color from successful participation [244]), affect students' sense of belonging and identity (e.g. instructional materials promoting gender bias [197]), and exacerbate existing disparities in privilege (e.g. students cannot synchronously engage with instructors and other classmates because of timezone differences, work commitments, or familial responsibilities [311, 204]) [172]. Inequities arise when structures and norms fail to include or serve students of minoritized groups, such as students who are African-American/Black, Hispanic/Latinx, Indigenous/Native American, Pacific Islander, women, non-binary, first-generation, transfers, and/or are part of other groups that not dominant within computing communities.

This challenge of broadening participation in computing education requires informed, timely, and equitable action. Stakeholders such as students, instructors, and curriculum designers must have training in culturally responsive practices that connect diverse learners' identities to learning

experiences by framing them as assets [244, 257, 289]. And with this development of cultural competence [297] must come changes to practices at a systemic level (e.g. changes to student outreach and admissions, recruiting of instructors [187]) as well as at class and individual levels (e.g. instructors diversifying or debiasing educational materials they use [197], instructors and students considering how unconscious bias affects their actions [244]). Given the booming enrollment and interest in computing, especially with a growing public awareness of how computing disproportionately harms minoritized groups (e.g. [134, 5]), there is a need to take informed and timely action to make learning computing a more equitable experience.

Data can help inform and enable timely action by substantiating nuanced patterns related to disparities and bias so we can consider how to appropriately adjust resources [231]. While equity is a goal that is typically too complex and situated to easily measure with data [305], it cannot exist without first analyzing data to understand existing disparities and biases [231]. Data can measure the existence and extent of disparities and biases, but judgement is required to interpret and act on data-related findings [313]. So enabling stakeholders with domain expertise to interpret and use data can help them take more informed, timely, and equitable action.

1.2 Research studies: Design explorations into how stakeholders interpret data

This dissertation explores how to design experiences that enable stakeholders (students, teachers, curriculum designers) to interpret and use data to support equitable learning experiences in computing education. I explore this within three domains:

1. Affording and informing agency in *self-directed online learning environments* to support learners of varying levels of self-efficacy. (Chapter 3)

2. How contextualizing data on *assessment* bias with curriculum designers' domain expertise can support equitable assessment and curriculum improvements. (Chapter 4)

3. How contextualizing *student feedback* with demographic information and peer perspectives can help instructors of large computing courses become aware of challenges that students

from minoritized groups face while preserving student privacy and sense of belonging. (Chapter 5)

1.3 Implications & contributions to learning analytics and computing education research communities

While data-driven techniques (e.g. statistical and machine learning) benefit from clearly defined dichotomies and independent features which purport to reflect all relevant information, stakeholders' identities are fluid, intertwined in intersectionality, and responsive to an everchanging context. That is to say that the quantitiative methods rely on static, isolated, and easily measureable features, but learners and their environments are dynamic, situated, and not easy to quantify [115]. This lack of alignment suggests that to support equity-oriented goals, we must design ways to enable direct stakeholders to interpret data by connecting it to their domain expertise. This has implications to learning analytics/learning at scale and computing education research communities.

- Learning analytics/learning at scale: Traditionally the learning analytics and neighboring research communities (e.g. Learning Analytics and Knowledge, Learning at Scale) has wielded quantitative methods in their use of data to understand and model learning experiences. But they have struggled to support equitable learning experiences [239] and "close the loop" by making data-driven insights actionable [51]. This dissertation makes steps towards using data for equitable action by contributing a framework and design explorations with empirical evaluations that inform the design of interactions with data that enable stakeholders to interpret data in support of equity-oriented goal.

- Computing education: Computing education research has used data to identify unequal disparities in participation and learning outcomes (e.g. [248]) and conducted investigations of equity where researchers were the interpreters of data (e.g. [237]). But it is still an open question how data may inform equity-oriented goals, which are more more socially situated. This dissertation contributes a framework and design explorations that consider a new approach to supporting equity-oriented goals in computing education that positions situated

stakeholders as interpreters of data.

My dissertation explores how to design interactions with data to support equity-oriented goals by bridging the gap between data and domain expertise and having stakeholders (e.g. students, teachers, curriculum designers) be the interpreters of data. Using data for equity-oriented goals requires us to connect interpretations of data with people's prior beliefs and experiences. Considering data for equity-oriented goals is challenging because it often requires stakeholders to consider data that deviates from their perceived norms. So we must consider these perceptions when we design interactions with data. Otherwise, stakeholders will tend to disregard the data.

Across my three design explorations, I have identified three factors that we must consider when designing interactions with data for equity-oriented goals: *relevant prior knowledge*, *perceptions of power relationships*, and *cultural competence* [72, 297].

This dissertation demonstrates the following thesis statement:

Interactions with data that consider prior knowledge, perceptions of power relationships, and cultural competency can enable computing education stakeholders to connect their interpretations of data with their domain expertise in service of equity-oriented goals.

1.4 Dissertation outline

In Chapter 2, I provide background information to situate my dissertation. This includes sharing my theoretical foundation of critically probelmatizing existing norms, describing the context of minoritized groups and equity in computing education, describing the expanding use of data in educational contexts through learning analytics, and finally sharing my unifying framework of factors to consider when designing for interpretations of data for equitable computing education. In Chapter 3, I describe a design exploration to support learners of varying levels of self-efficacy in self-directed online learning by affording and informing agency. In Chapter 4, I describe a design exploration to address assessment bias by having curriculum designers interpret empirical evidence of potential test bias (Differential Item Functioning) in a collaborative workshop. In Chapter 5, I describe a design exploration to understand how teaching teams identified inequities in their large,

remote, computing courses by interpreting contextualized student feedback. Finally in Chapter 6, I interpret findings from my study as informing future work on the design for interpretations of data for equitable computing education, findings with implications to the learning analytics and computing education research communities.

1.5 Positionality statement

The research in my dissertation required a reduction of people to the responses they were willing to share, so I acknowledge my assumptions and values in this section. By doing so, I follow critical approaches to quantitative methods which require researchers "to engage in critical self-reflexivity as a necessary first step for the long journey of deracializing statistics" [115]. As part of this process, I define assumptions and commitments that were the foundation of this dissertation.

Firstly, I attempted to approach this research with humility due to my positionality within many dominant groups of computing and with preparatory privilege. As an English-speaking, able-bodied, Chinese-American man with advanced degrees in computing from top research universities, I had been privileged within computing communities [183, 319, 94, 124]. Therefore, I could not assume that my experiences align with others, especially of those from minoritized groups. Put simply, I could not design for myself.

I also recognized the power structures and heterogeneity of people within different roles. Direct stakeholders in this research included teaching teams (including faculty members leading the teaching of a course and teaching assistants (TAs) supporting the teaching), university students, and K-12 curriculum designers. But even within these groups, there were differences. Faculty members leading instruction range from tenured research-track faculty who worked for years at an institution to newer teaching-track lecturers. And TAs were all undergraduate students, but their experience with the courses they taught ranged from never having taught or taken the course to having years of previous experience taking and teaching the course. University students ranged from full-time students who may or may not have been accepted to their major (admissions to computing majors is very competitive and not guaranteed as part of admission to the university). Most were enrolled to take computing courses, but some may have had listener status where they were not

taking it for official credit. Some transferred from other higher education institutions (e.g. two-year institutions) with different norms, while others came directly from high school. Curriculum designers had varying backgrounds and experiences within their organization. A common theme across all stakeholders: The data we collect is a partial and biased lens into their experiences in a select few courses as part of a a much larger educational experience.

I also acknowledged the tensions between representing people as data and labels, the *intersectionality* of people's identities (students in particular), and ensuring privacy. Intersectionality denotes the various ways in which ethnicity and gender (and other demographic labels) interact to shape the peoples' lived experiences [70]. Prior work has found that simplistic labeling of people can harm minoritized groups in particular. Labels of demographics (e.g. ethnicity, gender) academic experience (e.g. year in school, major, transfer or not), and lived experience (e.g. disabilities, familial language) are overly-simplistic. Furthermore, I needed to balance the nuance of the labels I select between how representative they were to diverse individuals and how anonymous they were such that others (e.g. instructors) could not map responses back to individuals or small groups of students. Despite these risks, I believed that stakeholders could still use these labels in such a way to help stakeholders contextualize data to support equity-oriented goals. Our perspectives align with the notion that "race is a measure of a relationship – not an inalterable trait" [316].

Finally, I acknowledged that models are always wrong in that they never fully reflect the complex phenomena I want them to represent, but they can be designed such that they are useful in informing stakeholders of hidden challenges. I framed the work I did as producing simplified models of complex phenomena such as learning computing, bias in measuring learning, and inequities in classes. I did not believe that in itself models will help, but they could support conversations, interactions, and interventions that address the systemic issues I sought to bring to light [59]. The objective of this dissertation was to explore designs that enabled stakeholders to interpret data for equity-oriented goals, and that is a first of many steps in making learning experiences more equitable and just.

Chapter 2

BACKGROUND

To understand how stakeholders' interaction with data can support equity-oriented goals, I must first provide a framng for equity. To do so, I draw upon the fields of critical data studies and learning sciences to adopt a critical framing. I apply this critical framing to the context of computing education, where many groups are minoritized. This critical framing enables me to contribute to discourse on computing education research by considering a role of data beyond identifying inequalities and disparities. It also enables me to contribute to discourse on learning analytics by developing a framework to inform the design of equitable learning experiences at scale.

2.1 Theoretical Foundation: Critically problematizing dominant structures, systems, and discourse

To understand how to support equity, we must first understand how inequities exist. To do this, I take a critical stance by stepping back to *problematize* dominant historical and cultural norms in computing education. This follows Michel Foucault's tradition of problematics [107], which has since been adopted into the method of *problematization* by James D. Marshall [189].

Problematization frames knowledge within malleable historical and cultural norms that we should first investigate by "stepping back" [189, 107]. Foundational to this methodology is the relationship between individuals (self) and knowledge. Traditions such as structuralism subscribe to a notion shared objective knowledge across individuals, but fall short of conceptualizing differences of individual lived experiences. In contrast, other traditions such as those started by Descrates emphasized "self-knowledge," or knowledge of one's own sensations, thoughts, beliefs, and other mental states [116], but fall short of investigating relationships across people in a society. Foucault

framed individuals as within historical and cultural norms that can change.

The objective of problematization is to understand thought by stepping away from prior beliefs and exercising freedom [189]. This method "enables the freedom to detach oneself from what one does...to establish it as an *object of thought* and to reflect upon it as a problem." It asks what political, social, and economic contexts have to say about the "problem" with which they are confronted. Examples of problems include the formation of a scientific domain, political structure, and moral practice. I use problematization as a framing to consider and critically question dominant systems, discourses, and actions to identify more equitable alternatives. By doing so, we can consider how to use data to inform stakeholders on who is or is not being served by existing systems so that they can imagine more equitable alternatives.

2.2 Context: Minoritized groups learning computing face systemic challenges

Within this critical framing, my dissertation explores equity within contexts related to computing education, the study of how people learn and teach computing, broadly construed [162]. It includes teaching programming skills for software development or data analysis, learning how to use computing for creative expression, and how to be conversationally "fluent" within computing communities [47]. In K-12 settings, computing is often taught in dedicated computer science classes, but there are also efforts to connect computing education with other classes as well [57]. In two and four year colleges and universities in the United States, computing is not only taught within computer science departments, but also in information science, math, engineering, communications, and social science departments [101].

While computing is taught in many contexts, there is a prevalence of minoritized groups that are not typically not dominant within computing communities in the United States (US). Dominant groups are positively privileged [299], unstigmatized [247], and generally favored by the institutions of society [186], particularly within social, economic, political, and educational systems [88]. For my dissertation, I characterized dominant groups as including white and Asian men who started or will start college at a four year institution shortly after high school (not transfer students), do not have disabilities, have little or no financial or familial responsibilities, have English fluency,

and have at least one parent who completed a four year college degree. In contrast, minoritized groups are groups that are not positively privileged or favored and often stigmatized. They include students who are women, non-binary, African-American/Black, Hispanic/Latinx, Native American/Indigenous, Pacific Islander, transfer students, not fluent in English, and/or first-generation, as well as students who have disabilities and/or have financial or familial responsibilities. Systemic cultures and norms in computing education tend to favor dominant groups and disadvantage minoritized groups.

2.2.1 Factors contributing to minoritization in computing education

Computing Education Research (CER) has identified the many social and environmental factors that contribute to the exclusion and minoritization of non-dominant groups. The following factors most directly relate to the design explorations in this dissertation:

- **stereotypes**: A dissociation between self-perception and perceptions of members of computing communities can lead to a reluctance to pursue computing education [86]. A Google-Gallup survey found that students and parents perceived TV and film media to present a stereotypical image of people in computer science as being white and Asian men wearing glasses [117]. These stereotypes can negatively influence girls, Black, Hispanic, and other students from minoritized groups from engaging and can contribute to stereotype threat [272, 300]. Nuancing considerations of student identity to go beyond stereotypes is a focal point in Chapter 5 and the design of *StudentAmp*.

- **self-efficacy**: An individual's belief in their ability to succeed in a particular domain [13] affects their educational and career experiences [25]. Prior work has found many factors that correlate with self-efficacy, including prior programming experiences [235, 158], gender [177], self-assessment [119], metacognitive strategies [177], understanding of programming concepts [235], and sense of belonging [288, 297]. Strategies to improve self-efficacy include peer instruction [315] and culturally responsive teaching [96]. Self-efficacy as the

foundation of agency relates to the design of the *Codeitz* online learning system in Chapter 3.

- **college experiences**: College experiences such as those at predominantly white institutions can result in students of minoritized groups (e.g. Black women [237], transfer students [168]) feeling isolated or inadequate [101]. Chapter 5 in particular considers how differences in experiences of different students at the same institution.

- **implicit bias**: Implicit or unconscious bias can cause harms to minoritized groups based on negative stereotypes [122]. Implicit bias can come from instructors (e.g. teachers not selecting women), peers (e.g. students not wanting to work with an older transfer student), or even curriculum (e.g. gender bias in computing textbooks [197]) [172]. Approaches to mitigate implicit bias include online practice spaces as part of professional learning [244]. Consideration of implicit bias comes up in workshops with curriculum designers in Chapter 4 as well as collaborative interpretations of data with teaching teams in Chapter 5.

2.2.2 Framing equity relative to equality, diversity, inclusion, & justice in computing education

Improving equity would require not just improving access to computing education, but also supporting successful participation and achievement by diverse students learning computing [172, 105]. Structural and systemic inequities embedded in and around computing courses can manifest as barriers to participation (e.g. unconscious bias of instructors excluding students of color from successful participation [244]), affect students' sense of belonging and identity (e.g. instructional materials promoting gender bias [197]), and exacerbate existing disparities in privilege (e.g. students cannot synchronously engage with instructors and other classmates because of timezone differences, work commitments, or familial responsibilities) [172]. Inequities arise when structures and norms fail to include or serve students of minoritized groups, such as students who are African-American/Black, Hispanic/Latinx, Indigenous/Native American, Pacific Islander, women, non-binary, first-generation, transfers, and/or are part of other groups that not dominant within

computing communities.

Critical to the concept of equity is a consideration of historical injustices. This differentiates equity from related concepts of *equality* and *inclusion*. Understanding equality or lack thereof is often a necessary precursor to supporting equity [231]. But efforts to support inclusion may actually be detrimental to efforts to support equity and *justice* because they distract from a need for broader systemic change [137].

Equity goes beyond the related concept of equality. Equality is rooted in the concepts of equal access and equal opportunities [274]. But equality is limited and insufficient because there is a lack of consideration for historical minoritization. For example, equality might mean students are admitted into a computing major based on their grades and test scores. But this does not account for factors contributing to grades and test scores, such as differences in preparatory privilege (e.g. students with more financial resources can afford test preparatory services). Equity goes beyond equality's framing of equal access and equal opportunity by focusing on equalizing outcomes. Doing so often necessitates unequal corrective measures to adjust for historical social inequalities [231]. So while equity goes beyond equality, it often cannot exist without first assessing inequalities to consider how to appropriately adjust resources [231].

Computing education research has used data to identify inequalities in access and achievement. To measure progress (or lack thereof) towards equal access and achievement, prior work has looked to measure access and performance by across demographic groups (e.g. gender, race), economic groups (e.g. by socio-economic status), and regional groups (e.g. comparing with census data). This requires data from a broadly deployed instrument, such as a survey or standardized test. For example, the Computing Research Association (CRA) has conducted large scale surveys to identify gender differences related to joining and persisting in computing majors [278]. Dr. Barbara Ericson has used data from the College Board and US Census to explore differences in access and achievement on AP Computer Science exams by demographics and gender [94]. She and her colleagues found that participation amongst reported female, Black, and Hispanic students is low but pass rate is slightly higher than AP Calculus exams, suggesting limited access but relatively strong achievement. Identifying disparities can help substantiate the existence of inequalities, but these

analyses typically do not provide insight as to what actions to take to address these disparities.

Equity also goes beyond the concept of *inclusion* by challenging systemic structures. Inclusion can be problematic as it typically demands that minoritized groups assimilate to existing dominant norms [93]. As Dr. Ruha Benjamin stated [20]:

> *...those who design the world according to their own values and biases are employing the rhetoric of "inclusion" as progressive veneer for deeply discriminatory practices.*

Minoritized groups in STEM have experiences that can vastly differ from those of dominant groups [172], so merely acknowledging cultural differences is often insufficient because it does nothing to seriously challenge dominant norms [17]. Within computing education, Rankin, Thomas, & Erete identified problems with inclusion as they identified typical experiences in traditional K-12 classrooms, predominantly white higher education institutions, and industry internships as sites of violence for Black women learning computing [237]. If not also accompanied by restructuring of existing dominant power structures, discourse around inclusion can fall short and even distract from equity-oriented goals [137].

Justice is a concept that often goes beyond equity by focusing on systemic changes at the broader scale of educational reform. Justice typically requires more continuous and systemic changes that consider how social and political power relationships at individual, classroom, university/organizational, and social levels affect students' lived experiences [68]. Advancing social justice often involves reforming educational policy [3]. Within computing education, Sepehr Vakil proposed a justice-centric framework that aimed to reform the context of the curriculum by considering ethics, incorporating the role of intersectional identity into the design of learning environments [70], and incorporating a political vision rooted in commitments towards transforming and empowering communities into the purpose of computing education [284, 283].

Whereas justice-centric efforts typically require systemic reform beyond the scope of individuals or computing courses, I frame equity-oriented goals as changes that can occur within the design of computing courses while largely working within constraints of existing institutional norms. Examples of equity-oriented goals within computing education include improving equity in activities

such as pair programming [171], addressing gender bias in instructional materials [197], targeted support for Hispanic students transitioning to remote learning [280], and teacher training to help address unconscious bias [244].

The work in my dissertation focuses on equity-oriented goals at the scope of individuals learning computing and achievement or experiences of students within computing courses. While my findings can inform a broader discourse related to justice-centric efforts, these justice-centric efforts are out of the immediate scope of investigation for this dissertation.

There is a plurality of conceptualizations relating to equality, inclusion, equity, and justice, with different framings serving different groups and contexts. To summarize, here is a concrete example on changes to enrollment policies that differentiates my conceptualizations of equality, inclusion, equity, and justice:

- Equality would advocate for a merit-based enrollment policy that prioritized consideration of seemingly objective measures of academic achievement (e.g. test scores, grade point average) without consideration of students' positionality relative to dominant structures.

- Inclusion would advocate for more enrollment ("increasing the pipeline") of minoritized groups in computing majors to train them to be software engineers at large technology companies.

- Equity would advocate for culturally responsive changes to practices and policies at universities and technology companies to ensure minoritized groups could recognize their dignity and potential without enduring dehumanization.

- Justice would advocate for situating computing education in a sociopolitical context and teaching minoritized groups with goals of justly serving society, its peoples, and their needs as they saw fit.

2.3 Learning analytics considers expanding uses of data

Research communities such as ACM Learning @ Scale (L@S) and ACM Learning Analytics and Knowledge (LAK) have investigated the expanding role of the data to improve learning experiences and outcomes. I will collectively refer to these efforts to use data analytics to improve human[1] learning experiences as *learning analytics*.

Learning analytics consists of the design and evaluation of data-driven applications that help learners improve their behaviors or help institutions manage information regarding learners' progress [11]. It was originally designed to exploit the growth in the collection of multiple data sources and typically collects data related to demographics (e.g. self-reported demographics from surveys), behavior (e.g. digital trace data of student actions), and performance (e.g. responses to tasks or assessments) to provide targeted interventions [109]. At a high level, learning analytics consists of a cyclical process of of 1) collecting data related to learning experiences, 2) generating data, 3) using it produce metrics, analytics, or visualizations, and 4) "closing the loop" by using findings to intervene or change a learning experience [51]. A common critique of learning analytics research is too great of an emphasis on the first three steps of creating robust data-driven models and tools, and a lack of emphasis on closing the loop with interventions that support learners [51, 12].

A promising future for learning analytics is to support reporting of information for human intervention [261]. But there are multiple ways to scale reach and empowerment, as designing for learning at scale could involve scaling through *efficiency* by enabling the same number of instructors help a larger set of learners (e.g. student feedback tools, automated graders) or by scaling through *empowerment* by enabling a larger number of people assist learners effectively (e.g. learnersourcing to facilitate scaled peer assessment [74]) [166].

Learning analytics lacks design principles for equitable use and impact. Data-rich technologies in learning experiences do not inherently disrupt systems of inequity; they often perpetuate them [109]. And access to learning is necessary but not sufficient to making learning more equitable.

[1]The machine learning techniques often uses learning to refer to training a model on data [205]. While learning analytics often uses machine learning techniques, mentions of "learning" in this section and the dissertation will refer to human beings learning, unless otherwise specified.

That is because social and cultural forms of exclusions are powerful and they are hard to understand and address [238]. Reich & Ito identified four themes related to equitable design of learning technologies: 1) unite stakeholders around shared purposes; 2) align home, school, and community; 3) connect to the interests and the identities of culturally diverse students; and 4) measure and target the needs of subgroups [239].

2.4 Role of data in equity-oriented goals

Many quantitative techniques exist to use data to identify issues with equality and/or fairness in learning experiences. Data can help us quantify disparities (differences in achievement by demographic group [296]) and/or biases (how systems disadvantage students of certain groups in disproportionate ways [83]). But to address these disparities and biases as part of equity-oriented actions, we must go beyond quantifying them to actually make judgement calls [313].

This dissertation explores how to use data for equity-oriented purposes by engaging the domain-expertise of stakeholders. Improving equity in computing would require access to learning computing as well as successful participation and achievement by diverse students [172]. Structural and systemic inequities embedded in computing courses can manifest as barriers to participation (e.g. unconscious bias of instructors excluding students of color from successful participation [244], affect students' sense of belonging and identity (e.g. instructional materials promoting gender bias [197]), and exacerbate existing disparities in privilege (e.g. students cannot synchronously engage with instructors and other classmates because of timezone differences, work commitments, or familial responsibilities). Because stakeholders are situated in contexts that are too complex to easily quantify, I believe they have the domain expertise to make the judgement calls about data to support equity-oriented goals.

2.4.1 Opportunities and associated risk of using data for equity-oriented goals

As stated previously, data can support equity-oriented goals by identifying nuanced patterns of disparities or bias to focus equitable interventions. But how to interpret and use data comes with

it associated risks that designers of learning analytics tools and experiences must be aware of. We cannot assume that findings from data alone can explain the culturally situated phenomena of inequity. Equity requires a consideration of historical structures of domination, so we must consider contextualize findings in our data within a broader context of the domain. Failing to do so risks using data to perpetuate inequities.

A commonly raised concern about risks of data about learners is violations of privacy [109, 2, 232, 222, 263]. Privacy involves consideration of trust, transparency, student agency/control over data, security, and accountability/reasonable care of data [222, 238]. But more than that, there must be a *sociocritical* perspective on the use of learning analytics that considers the role of power, impact of surveillance, and identity of students as transient, temporal, and context-bound constructs [263]. A critical approach to privacy could involve a lens of *vulnerability* where designers consider the needs of minoritized groups rather than those of dominant groups [194]. But as Prinsloo & Slade state [232], addressing privacy concerns ultimately involves a context-rich information that considers asymmetrical power relationships:

> *While learning analytics can and should play an important role in students' self-awareness, self-knowledge, self-efficacy, and healthy loci of controls, a lack of specific context can result in limited or even faulty assumptions. In the current collection and use of student data, students often have no insight into the data collected by their [higher education institutions] and so there is no possibility that data can be verified or any context provided. Considering the asymmetrical relationship of students and their institutions, students potentially then become quantified selves based on, for example, the number of log-ins, clicks, downloads, or time-on-task. It is important to allow opportunities for context-rich information so that institutions and students may better understand the complexities and interdependencies in the nexus between students, institutions, and the impacts of socioeconomic, technological, environmental, political, and legal contexts.*

Another concern about the use of data is reduction of nuanced students and their identities to

low dimensional data. I summarized these risks in the story of Agloe in the introduction (Section 1) and Figure 1.1. The metaphor "The map is not the territory" also summarizes these kinds of risks. Mayer-Schoenberger & Cukier have coined this "transformation of social action into online quantified data... allowing for real-time tracking and predictive analysis" as *datafication* [191]. This postivistic framing assumes an objective aggregation and quantification of data can predict individual behavior to sufficient accuracy [286]. But this assumption that aggregate behavior reflects individuals is contrary to goals of equity because it purports that behaviors of minoritized groups reflect that or dominant groups.

Of particular concern with reduction of people and their experiences through datafication is the reduction of identities into demographic data. Intersectionality denotes the various ways in which ethnicity and gender (and other demographic labels) interact with social structures to shape the peoples' lived experiences [70]. Prior work in computing education has found that simplistic labeling of attributes can harm minoritized groups in particular. For example, Dr. Christina Convertino identified how oversimplifying the narrative that women were an underrepresented, invisible monolith is a unproductive reduction that fails to allow space for nuance, such as how BIPOC women in CS push back on the dominant discourse of underrepresentation [58]. Ross et al. conducted a more intersectional analysis of survey data, comparing experiences of computing students who were Black women, non-Black women, and Black men to nuance the intersection of being Black and being a women [248]. Pushing back on datafication by providing more nuanced and contextual data can come in tension with privacy-related goals though, a tension I investigate in StudentAmp (Chapter 5).

To summarize, I believe that data can support equity-oriented goals by showing nuanced data of students with intersectional identities that also preserves students' privacy within a sociopolticial learning context. Within my dissertation, I explore this by having stakeholders with domain expertise be the interpreters of data.

2.4.2 Unifying framework: How beliefs and exeperiences affect data sensemaking

As stated previously, data can show evidence of disparities, bias, or inequities, but it alone does not translate to equitable action. Because equity-related data can be unfamiliar, uncomfortable, or conflict with expectations [216, 209], we must carefully design interactions with data. To support equitable actions, this dissertation explores how to enable stakeholders to connect their interpretations of data to their domain expertise in three contexts:

1. Recommending future actions to students using self-directed online learning tool: How can we use data on students' prior actions and performance to support a self-directed online learning experience that is equitable to learners with varying self-efficacy? (Chapter 3)

2. Showing curriculum designers evidence of potential test bias to inform their practice: How can empirical evidence of potential bias by gender and race help curriculum designers develop more inclusive learning experiences? (Chapter 4)

3. Showing teaching teams contextualized student feedback to inform them of inequities in their large, remote courses: How can contextualizing challenges students report with demographic information about students inform teaching teams' of inequities in their classes? (Chapter 5)

To understand the process of interpreting data for equity in educational contexts, I draw upon a prior framework of teachers' sensemaking of data for equity by Bertrand & Marsh [23]. This framework identified how teachers explain the causes of outcomes in observed data, as shown in Figure 2.1. Connecting attribution theory to sensemaking, this theoretical framework of teachers' data use considered interactions between beliefs, past experiences, current circumstances and social context when teachers explained causes observed in data. Past experiences and beliefs influence how teachers make sense of data, but this framework did not elaborate on factors that contributed to their development.

Building off the framework developed by Bertrand & Marsh [23] (shown in Figure 2.1), I consider three factors that contribute to beliefs and experiences when interpreting data for equity:

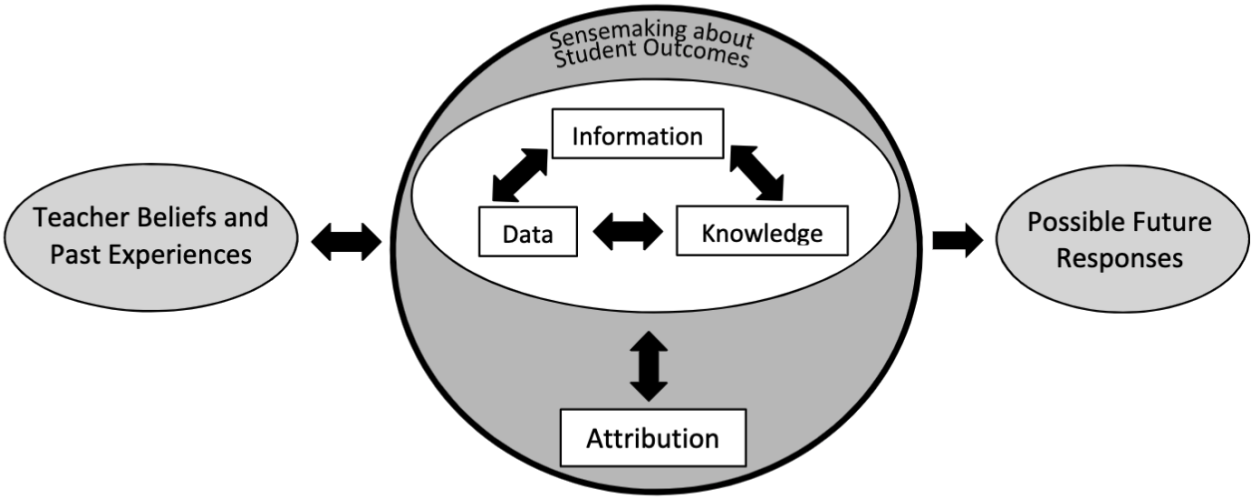


Figure 2.1: Bertrand & Marsh's theoretical framework of teachers' sensemaking of data (Fig. 1 of [23]). The framework I define (Fig. 2.2) expands upon this one by identifying factors the affect the formation of beliefs and past experiences.

prior knowledge, perceptions of power relationships, and cultural competence. This is summarized in Figure 2.2 and described in the following sections.

Relevant prior knowledge

People interpret data relative to the prior knowledge they deem relevant. As consistent with theoretical framings from learning sciences [216, 209], I frame the act of interpreting data as connecting new knowledge to existing knowledge frameworks.

While impactful, the effect of prior knowledge on interpreting new information varies. If data appears to conflict with prior knowledge, then people may dismiss the new information in the data and instead rely on their existing knowledge to come to conclusions [268, 209]. Or perhaps prior knowledge biases people to narrowly interpret data in a way that does not conflict with their prior knowledge [209].

Ideally, people use prior knowledge to provide additional information and richness without limiting their their interpretations of data. Doing so requires requires conscious effort though [209]. Engaging with prior knowledge is difficult without considering how the data exists in a

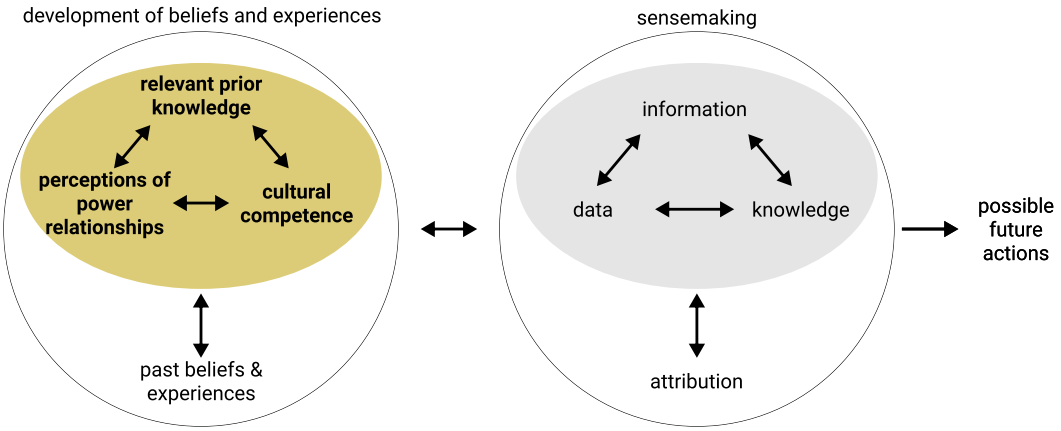


Figure 2.2: Unifying framework of my dissertation: The development of beliefs and experiences requires consideration of relevant prior knowledge, perceptions of power relationships, and cultural competence. Adapted from Fig. 1 of [23].

situated context. In my dissertation, I explored how stakeholders interpreted data to understand how prior knowledge from their domain expertise and lived experiences may provide richness to interpretations.

Cultural competence

Cultural competence is a model to guide actions taken at individual, organizational, and systemic levels to meet the needs of culturally and racially diverse groups in a culturally appropriate way [72]. Cross et al. [72] defined cultural competence as follows:

> (A) set of congruent behaviors, attitudes, and policies that come together in a system, agency, or among professionals and enable that system, agency, or those professionals to work effectively in cross-cultural situations. The word *culture* is used because it implies the integrated pattern of human behavior that includes thoughts, communications, actions, customs, beliefs, values, and institutions of a racial, ethnic, religious, or social group. The word *competence* is used because it implies having the capacity to function effectively.

Originally created for use in social work with minoritized groups [72], cultural competence was recently adopted for use in computing education [297]. Developing cultural competence of individuals as well as organizations can support more diverse, equitable, and inclusive environments and technologies [297].

There are four core components of cultural competence: attitude, awareness, skills, and knowledge [297]. *Attitude* relates to valuing diversity. It is not just recognizing that diversity includes many factors, but also appreciating and valuing how all factors are critical for inclusive environment. *Awareness* relates to cultural self-assessment. It is a recognition of own beliefs and positionality and how they interacts with others' beliefs and positions in a given context. *Skills* relate to understanding the historical impact of certain actions, words, and beliefs on people and adapting to better meet needs of people from diverse backgrounds. Finally, *knowledge* relates to institutionalized cultural knowledge across all organization levels.

The development of these four components of cultural competence occur through the following six stages [72, 297]:

1. cultural destructiveness: attitudes, policies, and practices are destructive to cultures and individuals within culture

2. cultural incapacity: not intentionally destructive, but lack capacity to support marginalized groups (e.g. ignorance)

3. cultural blindness: ignoring cultural strengths and encouraging assimilation. At this stage, "success" is seeing marginalized groups approximate middle-class.

4. cultural pre-competence: realization of deficits and attempt to improve through active and intentional efforts. At this stage, concerns of false sense of accomplishment or tokenism of marginalized groups.

5. cultural competence: clear, intentional, working examples of all elements of cultural competence. Understanding of effects of policy on practice and actively working to ensure enacted

policies support diverse and inclusive environment.

6. cultural proficiency: highest level. Valuing culture in highest regards and constantly searching to add to their knowledge through active research, development of new strategies, dissemination of results. Cultural competence is clearly understood and constantly demonstrated.

In this dissertation, I consider cultural competence as a factor that affects how stakeholders interpret data. I posit that more developed cultural competence can support deeper engagement and interpretations of data. More cultural competence in interpreting data can consider the data as a signal of disparities or bias that arise from exclusive or harmful systemic policies and practice.

Perceptions of power relationships

At the foundation of my approach for supporting equity-oriented goals is a commitment to a critical perspective, which requires a consideration of power relationships [273]. As mentioned in Section 2.1, critical perspectives require a consideration of power relationships in systems, ideologies, and institutions in a given contexts [273]. Within the United States, education often has a neoliberal framing and is a private good that benefits individuals by improving their economic status or providing a larger, more skilled pool of laborers for corporate interests [220, 129, 28, 91]. Within computing education, we see this neoliberal framing through efforts to improve and diversify "pipelines" to train individuals to earn high paying jobs in large technology corporations (e.g. [118]). Within this framing, power relationships take on a monetary value, incentivizing those dominating these relationships to oppress others to preserve this societal advantage.

Considering power relationships at a systemic level is an exploration into how dominant systems came to be and how they operate to oppress minoritized groups. The key commitment that enables this analysis of power relationships at a systemic level (compared to individual level) is a dialectical relationships where individuals influence social contexts and social contexts influence individuals in heterogenous ways [153, 273]. That is to say that individuals exist in and are constantly influencing social contexts and norms that are constantly changing. And those social

contexts and norms are shape individuals and their knowledge. Supporting equity-oriented goals involves understanding how power relationships shape dominant social systems and ideologies.

But individual's relationships with social systems are constructed, and so too are their *perceptions* of power relationships. Individuals perceive power relationships relative to their own positionality [195]. Learning in formal contexts is a process mediated by networks power relationships between students and their peers, instructors, and the subject matter itself [67]. For example, formal education of computing often involves learning computing in a traditional K-12 classrooms, taking courses at a predominantly white institutions, and perhaps participating in industry internships. While this is norm may benefit students of dominant groups, Rankin et al. identified how Black women experienced "sites of violence" in each of these contexts [237]. So rather than attempt to take an objective or global view to power relationships, we must instead consider *perceptions* of power relationships.

Framework resembles one for K-12 science argumentation

The resulting framework most resembles frameworks for culturally responsive argumentation and sensemaking for K-12 science classrooms. Levine et al. synthesized a hybrid model of representation to represent the "dialogic, multi-layered, and culturally situated" nature of argumentation [170]. They framed argumentation as enacting situated roles (similar to my considerations of power relationships), drawing upon cultural belief systems (parallel to my consideration of cultural competence), and situating argumentation in an ongoing dialogue (similar to my consideration of prior knowledge). The context of developing scientific argumentation skills amongst K-12 students and having adult stakeholders interpret data for equity-oriented goals is quite different. Nevertheless, both frameworks emphasize the situated nature of stakeholders and their beliefs affect interpretations of data (or argumentation) that build off of existing discourses. Frameworks like this can guide the design of experiences that critically stakeholders in new kinds of interpretations as they use them to inform their own practices.

Chapter 3

CODEITZ: ADAPTIVE RECOMMENDATIONS FOR SELF-DIRECTED ONLINE LEARNING TO SUPPORT LEARNERS OF VARYING SELF-EFFICACY

In this chapter, I describe a study that investigated how to support the equity-oriented goal of equitable self-directed online learning experiences. I built *Codeitz*, a self-directed online learning environment that used data on students' prior actions to recommend future actions.

The goal of the recommendations was to support learners of varying levels of self-efficacy by affording and informing *agency*, a learner's capacity to define and pursue learning goals [14]. The main hypothesis was that learners with high self-efficacy could guide their own learning experience, whereas learners with low self-efficacy could follow the recommendations. From a between-subjects evaluation of Codeitz, I found that affording and informing agency affected engagement, but had no detectable difference in learning outcomes. Post-survey responses suggested that the affordances of the design could not overcome the prior knowledge and perceptions of power relationships that participants had. Most participants were university students who had limited agency within their formal learning experiences. So a potential explanation was that agency was unfamiliar and the design affordances could not overcome this unfamiliarity. This substantiates the need to design interactions with data that consider how prior knowledge and perceptions of power relationships affect self-directed online learners' interpretations of recommendations.

The work in this chapter was conducted in collaboration with Dr. Greg L. Nelson, Matt J. Davidson, Harshitha Akkaraju, William Kwok, and Dr. Amy J. Ko and published to 2020 ACM

Learning @ Scale [309]. [1] [2]

3.1 Introduction: Design space for agency

Agency, or the sense we are in control of our actions and their effects [202, 200], is important to learning. Agency can make students contributors to their learning experience rather than just products of them [15]. In classroom settings, teachers overwhelmingly believed that affording students agency improves motivation and learning outcomes, while also recognizing that limits to agency were necessary [106]. This recognition that both freedom to make choices as well as the scaffolding to limit these choices suggests that designing to promote effective learner agency is important, but nuanced.

The need to balance freedom and guidance is especially true in self-directed learning settings, such as online tutorials and educational games, where designers create the entire instructional experience and no human teacher is available to provide assistance. In these experiences, agency can be framed as a phenomenon involving both a learner and their learning environment, in which the actions that learners desire are among those they can actually take [295]. The goal of having learners exert agency is to have learners make informed choices to support their engagement [40, 250], motivation [66], and learning [265, 277, 250, 66]. Agency might manifest as a learner deciding that an exercise is too easy, choosing to jump ahead to a more difficult exercise, or realizing that they lack some understanding, and reviewing some prior instruction. These decisions emerge from a learner having a goal, taking an available action to support their goal, and then reflecting on the result of their action [201].

Elements of self-directed learning environments will always influence learner agency, but not always in ways that benefit learning. To exert agency, learners must first perceive that they can

---

[1]This study was pre-approved with exempt status by the UW Institutional Review Board (IRB) as STUDY00005282. This material is based upon work supported by the National Science Foundation under Grant No. 1735123, 1539179, 1703304, 1836813, 1450681, and 12566082 and unrestricted gifts from Microsoft, Adobe, and Google. We thank Alannah Oleson for the Gender Mag walk-through [39], and our Code & Cognition lab mates for their constructive critique. Supplemental material (code, surveys, post-test) can be found at `https://github.com/codeandcognition/archive-2020las-xie`.

[2]I will use "we" instead of "I" in this chapter to acknowledge the shared contributions of all authors.

do so [200, 63]. Learners rely on their perceptions of the environment to develop their sense of agency [295], so the design of the learning environment is impactful to their agency. Designers exert *indirect control* [254] over learner actions. These elements of indirect control can inadvertently result in learners following similar paths for no reason beneficial to learning, therefore unnecessarily limiting their agency. This was the case in a computer-based math game, where learners were afforded the agency to play mini-games in any order but instead tended to follow a dotted line which visually connected mini-games in a somewhat arbitrary order [127, 212], resulting in no difference in learning outcomes between high- and low-agency variations of this game. Therefore, designers must effectively scaffold a self-directed learning experience to ensure learners exert agency by making informed choices that benefit their learning.

Prior work on agency in self-directed learning environments has primarily explored the effect of more or less agency on learning. For example, studies of the self-directed educational game Crystal Island [250] have found that limiting available actions in the virtual environment led to better learning gains when compared to a high-agency condition, but limiting options also led to an increased propensity for guessing [279, 253]. However, prior work has also found that too much agency can also be detrimental [9]. This was the case in Chen et al. 2019, which found that learners with more prior knowledge in a high-agency condition (where they could choose their own preparation tasks) exhibited similarly unproductive behavior such as guessing [46]. These findings suggest that designing for agency means finding a "sweet spot" that brings the benefits of choice, while preventing learners from being overwhelmed [265, 87, 63].

While prior work on self-directed learning has explored varying levels of agency [46, 181, 182, 246] and agency over different aspects of learning [64, 63, 62], it has not jointly explored varying levels of *information* to support agency. And information is key: there is a difference between giving a learner a choice about what to do next, and giving them carefully designed information about the risks and opportunities of those choices. Prior theoretical work calls this *proximal action-related information* [201], which aims to help learners determine their 1) capacity to act (e.g. empty check boxes indicating practice that has not been completed, showing available mini-games and hiding previously completed ones), 2) current ability to do so (e.g. skill bars showing estimated

knowledge in an open learner model [142], earned badges to reflect accomplishments), and 3) the predicted result of taking an action (e.g. an adaptive recommendation denoting that a specific practice question can serve as review). This framework suggests how systems might provide such information, but prior work provides no design guidance on the effects of varying levels of action-related information on agency.

To contribute to this design guidance, we built a self-directed learning environment for learning Python programming, varying both the amount of agency afforded and the amount of information provided to support learning decisions. We specifically studied three design variations: 1) informed (high-information) high-agency, 2) uninformed (low-information) high-agency, and 3) informed low-agency. With these alternatives, we then conducted a between-subjects experiment to investigate the effects of these design choices on 1) learners' experiences, and 2) learning outcomes. Participants in the study engaged in self-directed learning for a week, then took a survey and post-test measuring learning gains. In the rest of this paper, we discuss the design alternatives and our study design in detail, then present our results and their implications on designing for agency.

3.2 Theoretical Background on Agency

Before discussing our design alternatives and study design, we discuss the theoretical views that inform both. In particular, while there are many definitions of *agency*, in this paper we frame it as occurring when a learner can take actions that align with their learning-related goals [295]. Within this framing, we position Bandura's notion of *self-efficacy* as the primary individual factor that influences both learning and the use of proximal action-related information found in a learning environment [15, 14]. From this view, learners must believe in their abilities to organize and execute a course of action as well as process information from the environment regarding potential actions to take and their implications.

While agency is dependent on self-efficacy, acting upon self-efficacy requires information from a learning environment. We specifically draw upon frameworks of *proximal action-related information*, which positions information that is situated near and related to a decision [201] as critical to agency. Examples of such information include skill bars indicating the current state of understand-

ing, check boxes indicating what a learner has or has not completed, or adaptive recommendations suggesting a next topic to learn.

Finally, we also draw upon the *Preference Construction* (PC) model of decision-making to explain the importance of proximal action-related information to agency. This model is commonly used in explaining economic decision-making and frames preferences as a contextually developed construct [24, 175]. PC draws upon Herbert Simon's notion of *bounded rationality*, which states that the complexity of a decision task, limitations of cognitive resources and knowledge of people, and the tendency to reduce decision effort lead to a limited rationality [262]. This implies a trade-off between decision-making effort and the accuracy of the decision outcome [228] and that because PC is contextual, it is susceptible to different kinds of biases. Within the context of recommender systems, influences such as context effects, primary/recency effects, framing effects, and anchoring effects may bias how people make decisions [185]. PC states that humans do not have a clear preference in the very beginning, but rather develop preferences within the context of a decision process. Therefore, proximal information is critical for exerting agency.

An aspect of agency that is beyond the scope of this paper is metacognition, one's ability to monitor and regulate their own cognitive processes, behavior, and affect [209]. Metacognitive skills can support agency [211, 201], but vary amongst novice programmers [180]. We attempted to remove this confound through random assignment in our study, detailed later in the paper.

3.3 Three designs to explore agency

Given these theoretical foundations, we considered three variations on degrees on agency and proximal information: an Informed, High-Agency (*IH*) design that gave learners agency and information; an Informed, Low-Agency (*IL*) design that gave learners information but little choice; and Uninformed, High-Agency (*UH*), which gave less informed choices. In this section, we describe the learning domain, how we provided proximal information, and our three designs.

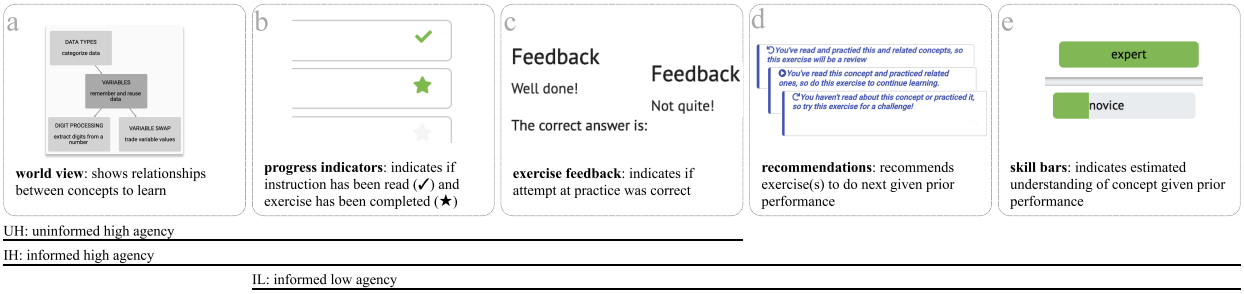


Figure 3.1: Features of Codeitz designed to provide learners with proximal action-related information for deciding what to learn next. Variations of the environment exposed learners to different subsets of the features (see lines at bottom of figure).

3.3.1 Learning Domain: Self-directed intro to Python

To explore agency and proximal information, we selected the domain of learning to program. As a domain, programming has many attractive features: at this point in history, many people want to learn it; there are many examples of self-directed learning environments for learning to code online; and the domain itself has concepts with relatively clear inter-dependencies that are amenable to learner modeling.

To support our investigation, we developed *Codeitz*, a new self-directed learning environment to teach Python programming (see Fig. 3.1). All variations of Codeitz shared the same introductory curriculum adapted from the materials defined in [308]. This curriculum was designed to assume no prior programming knowledge and cover introductory Python concepts including basic constructs (data types, operators, variables, print statements, conditionals) and templates demonstrating ways to use what was learned (variable swap, float comparison, find min/max, digit processing).

While the original learning materials were created for more linear learning, we relaxed this constraint to make learning through exploration more feasible. Following semantic dependencies defined by the Python programming language and extending that pattern of hard dependencies to templates [308], we developed a concept hierarchy that learners could use to decide what to learn next (shown in Fig. 3.2 and described below). To adapt the learning materials to match the

concept hierarchy we defined, we adjusted instructional content to assume learners only visited parent concepts and created additional exercises to practice which relied on fewer other concepts. We kept some examples and exercises which relied on concept dependencies not reflected in our hierarchy, so this adaptation was not complete.

From an instructional design perspective, we designed Codeitz to be a self-contained learning environment. To learn a concept, learners could read instruction to develop conceptual understanding of an aspect of Python and then attempt practice exercises where they received feedback related to correctness from the system. Practice exercises included multiple choice, short answer, filling in Memory Tables [310] to trace program state, and writing code. To support a formative experience, learners were able to retry practice exercises and see the answer whenever they wanted. Each page of instruction or exercise mapped to exactly one concept.

3.3.2 Three Codeitz Designs Varying Agency, Information

All three variations of Codeitz had the same instructional material and included conventional feedback on learning progress (Fig. 3.1b) and exercise correctness (Fig. 3.1c) common to online learning tools such as online courses (e.g. edX, Coursera) and learning platforms (e.g. Khan Academy, Codecademy). However, the designs varied in the amount of agency and predictive information afforded to learners.

We specifically focused on supporting learners' decision of *what to learn next* by varying the presence or absence of three features that either afforded agency or offered proximal information to support learning decisions. One feature was a **world view** showing Python concepts and their dependencies (Figures 3.1a, 3.2). We designed the world view to be as nonlinear as possible so as to encourage learners to exert agency and explore different concepts while having an understanding of their underlying relationships. Learners could use the world view to explore concepts as they relate to other concepts they may have already learned. Another feature was **recommendations** of what to learn next (Fig. 3.1d). These were based on the estimated difficulty of the exercise relative to learners' current levels of understanding for a concept. Recommendations supported the goals of *reviewing* (exercise involves a concept learner is knowledgeable with), *continuing*

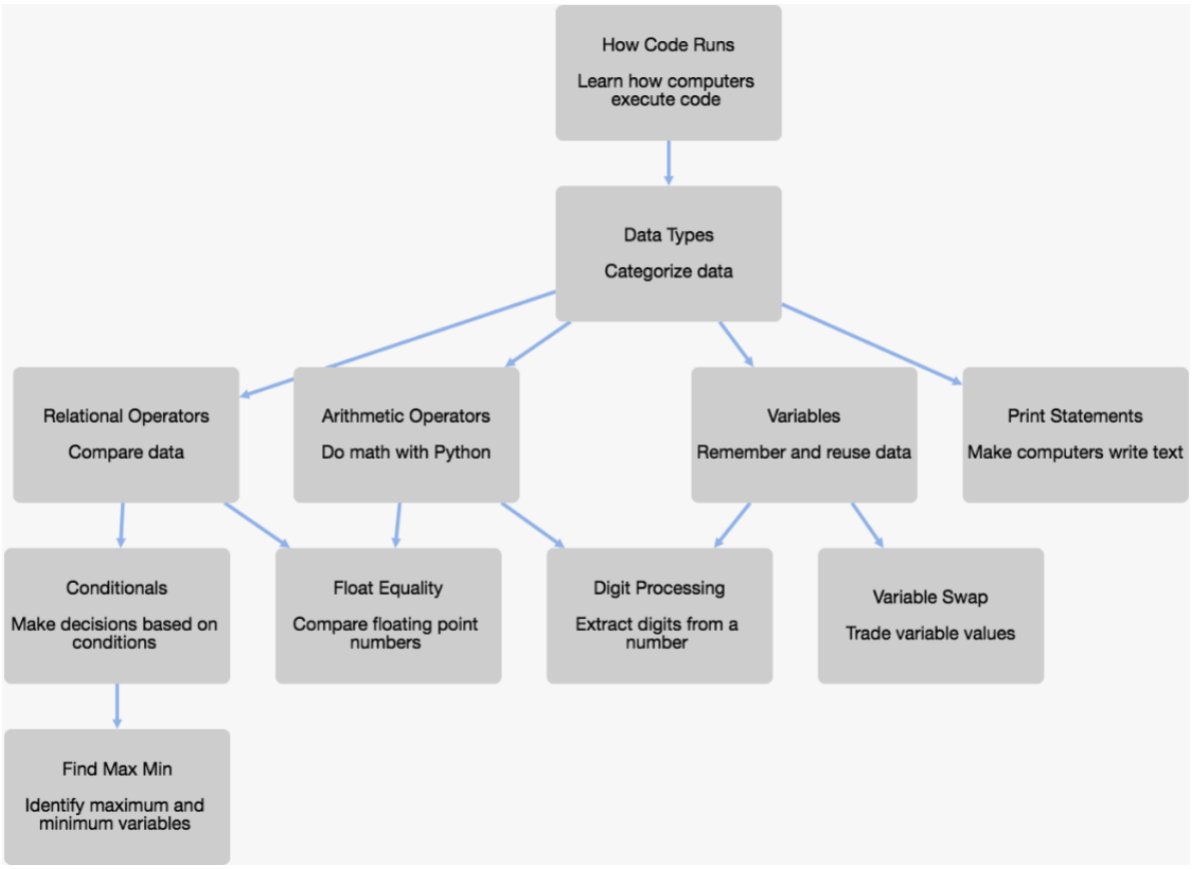
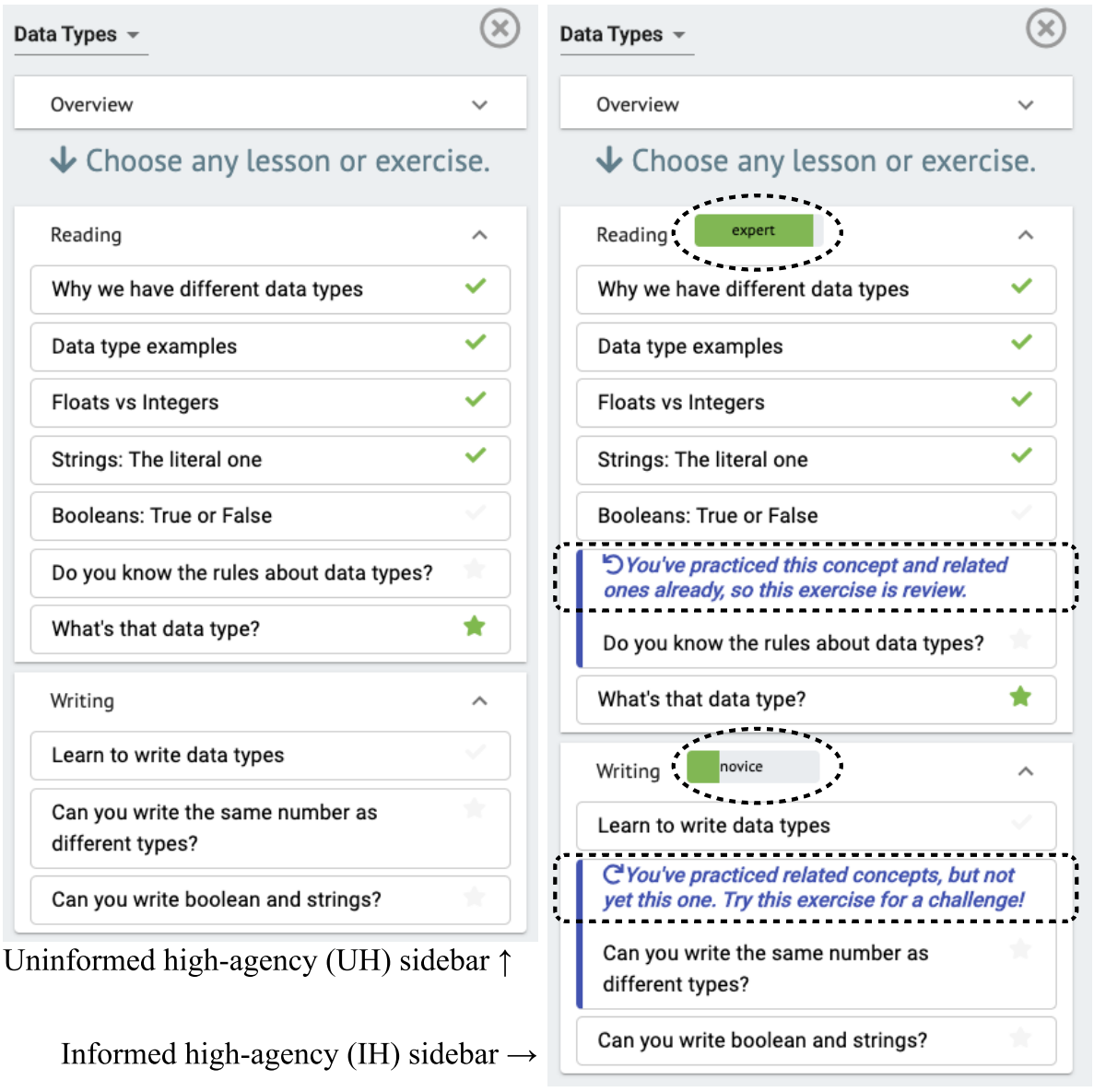


Figure 3.2: The world view, showing Python concepts taught and major dependencies between them.

(exercise involves concept learner has made progress with), or *challenging* (exercise involves a concept a learner has little experience with). Learners can use recommendations to judge how certain exercises may support current goals. And finally, **skill bars** provided estimated levels of mastery for a concept (Fig. 3.1e), to help learners determine if they needed to complete all of the instruction and exercises or whether they could move on to another concept. Learners can use skill bars to judge how well they have mastered a specific skill. Our three designs offered unique combinations of these three features.



Uninformed high-agency (UH) sidebar ↑

Informed high-agency (IH) sidebar →

Figure 3.3: Sidebars for the uninformed high-agency (UH) and informed high-agency (IH) variations of Codeitz. The UH version (left) only shows what instruction and exercises a learner has completed (using check marks and stars). The IH version (right) includes skill bars (dotted ovals) to denote estimated mastery and blue goal-oriented recommendations for next exercises to consider (dotted rectangles).

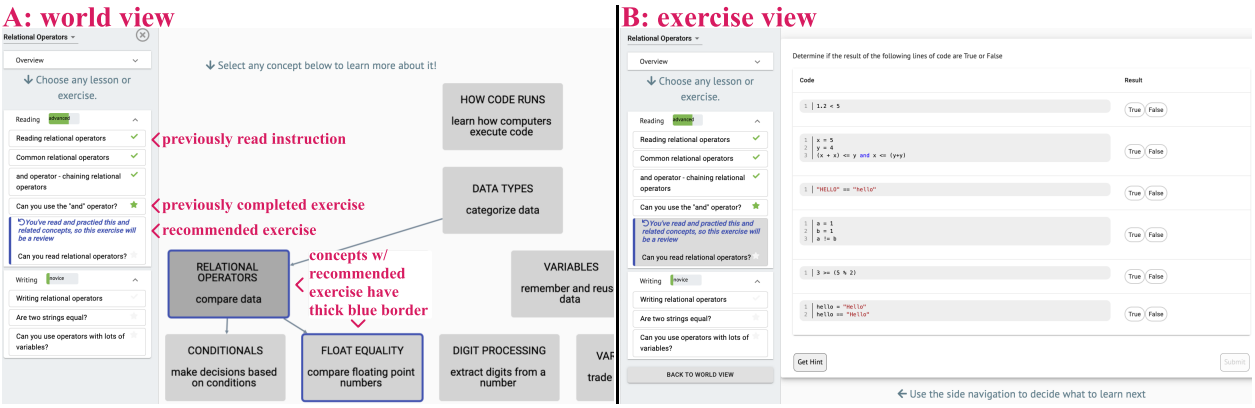


Figure 3.4: Two primary views of Codeitz for informed high-agency (IH) condition. **A**: The learner followed a the recommendations and selected the *Relational Operators* concept and is able to view the instruction and exercises for that concept in the sidebar. **B**: After clicking on the recommended exercise (*Can you read relational operators?*), the learner is then taken to the exercise view where they can attempt the exercise as practice. In the uninformed high-agency (UH) condition, there are no blue recommendations. In the informed low-agency (IL) condition, there is no world view (A) and learners must instead follow system recommendations.

UH: The uninformed high-agency version lacked recommendations & skill bars, but still required learners to exert agency.

We intended for this version to reflect an open online course (e.g. a MOOC) in the information provided to a learner as well as its availability of content. In this design, learners were uninformed of system predictions from their prior responses. They used information about the knowledge domain, progress they made, and exercise feedback (Fig. 3.1a-c) to guide their learning experiences. They would select a concept from the world view (Fig. 3.2), then use the sidebar as shown on the left of Figure 3.3 to look at what instruction and exercises they had/ had not completed. With this information, learners using the UH version of Codeitz had the freedom to explore any instructional material in any concept.

IH: The informed high-agency version provided recommendations and skill bars while requiring learners to exert agency.

Recommendations highlighted specific concepts in the world view and certain exercises in the side bar. The right side of Figure 3.3 shows the sidebar for the IH version with skill bars to show estimated mastery of a concept and recommendations to show recommended exercises which may support different goals (e.g. review, challenge). We intended for the IH version of Codeitz to reflect a recommender system where learners could follow recommendations but could also choose to deviate from them at no penalty. Figure 3.4 shows the interactions of the IH condition.

IL: The informed low-agency version provided recommendation and skill bars, but limited choices to a single next recommendation or prior exercises.

We intended for the IL version of Codeitz to reflect a Computerized Adaptive Test (CAT) [43, 44] or basic Intelligent Tutoring System (ITS) [304] where the system decided the next exercise for learners. So rather than being free to choose a concept and then an exercise as high-agency conditions did, learners using the IL version clicked a "next" button and the system selected the concept of the top recommended exercise. From there, they could choose to 1) do the exercise, 2) read related instruction, or 3) review any prior concepts. Only after they attempted the exercise would they be provided with a new one.

3.3.3 Adaptivity with Bayesian Knowledge Tracing (BKT)

To estimate learners' knowledge and recommend/select exercises for IH and IL designs, we implemented a modified version of the Bayesian Knowledge Tracing (BKT) algorithm [65]. BKT is a Hidden Markov Model that has the key assumption that learners can undergo a one-way transition from the *unlearned* to *learned* state for each concept, after which there is a change in the probability they will get an exercise correct [230, 151, 156]. While BKT typically assumes items to be equal, we used the Knowledge Tracing Item Difficulty Effect Model (KT-IDEM, [223]) to encode exercise difficulty.

Our model had two parameters at the concept level and two at the exercise level. The concept-level parameters were $P(L_0)$, the probability a learner already knew a concept before attempting an exercise, and $P(T)$, the probability of a learner transitioning from an unlearned to learned state after an exercise attempt. The exercise-level parameters are $P(S_m)$, the probability of a learner who had learned a concept *slipping* and getting an exercise m wrong, and $P(G_m)$, the probability of a learner who had not learned a concept *guessing* and getting m correct. A more difficult exercise would have a higher slip probability and a lower guess probability. We fitted these model parameters using expert review [178] based on ≈15 responses and exercise properties (e.g. closed or open form, perceived difficulty of exercises), and knowledge domain; all parameters ranged from 0.01 to 0.25.

Put together, we used this modified version of BKT to estimate the probability of getting an exercise correct. The estimated probability a learner will get a given exercise m correct at the n-th attempt is $P(correct_n|M_n = m) = P(L_n)(1-P(S_m)) + (1-P(L_n)(P(G_m))$, or the sum of the probability of getting the exercise correct in the learned and unlearned states. We used this to incrementally update the probability a learner was in the learned state after the n-th opportunity as follows: $P(L_n|correct_{n,m}) = P(L_n)(1-P(S_m))/P(correct_{n,m})$. We used this probability of being in a learned state as an estimate of a learner's understanding of that concept.

To select exercises, we used the BKT-Sequence Algorithm [75] which orders exercises based on a minimum difference between predicted difficulty and desired difficulty based on current learner understanding. After each exercise attempt, the probability a concept is learned ($P(L_n)$) was updated. We then updated the sequence of recommended exercises:

1. Calculate $MinScore$ and $MaxScore$, the minimum and maximum $P(correct_m)$ for all incomplete exercises.

2. For all incomplete exercises, calculate $WantedScore_m = (MaxScore - MinScore) \cdot (1 - P(L_n))$ where n is the concept corresponding to each exercise.

3. Calculate $diff_m = WantedScore_m - P(correct_m)$.

4. Order exercises in ascending order by $|diff_m|$.

We then selected the top two exercises, as well as the top two exercises from current, parent, or child concepts.

3.4 Study: Agency on Experiences, Learning

To understand the effects of varying information and agency afforded in our three versions of Codeitz on engagement and learning, we conduct a between-subjects study with 79 novice programmers. We sought to be ecologically consistent with discretionary use tools to support novice programmers learning in formal learning environments (e.g. an online practice tool used by students in an introductory CS course).

The study included novice programmers who were primarily university and community college students near an industrialized urban center of the United States. We recruited participants through flyers placed throughout a university and surrounding area, pitches to computing-related courses, and posts to closed social media groups. Our inclusion criteria specified participants had to be at least 18 years old, never learned or used Python, completed at most one non-Python programming course prior (although 9 participants violated this criteria: UH:2, IH:5, IL:2), have access to a computer with internet, and be fluent in English. Participants' self-reported ethnicities were Asian (52%), Caucasian (27%), Hispanic/Latinx (9%), mixed race (6%), and Black/African (3%), with 4% choosing not to disclose. Genders of participants were men (51%), women (44%), and non-binary (1%), with 4% not disclosing. Most (84%) reported working towards one of 40+ different degrees (roughly, physical sciences: 23% of all participants, computer science & informatics: 19%, engineering: 16%, humanities, arts, social sciences: 10%, business & finance: 8%, math: 1%, undeclared: 3%).

Participation in the study began with participants creating an online acount and then getting randomly assigned to one of three conditions. They then completed a pre-survey which asked questions relating to programming self-efficacy (as measured by a programming self-efficacy survey [236]), mindset [89], and motivation for participating in the study. They then used Codeitz

across the span of a week and then when they felt ready, took a post-survey and post-test. We compensated participants with a $50 gift card upon completion of an exercise in most concepts and the post-survey.

The post-survey asked learners about their experience using Codeitz (which also served as a distractor task [111]), then administered the hour-long post-test, then measured their programming self-efficacy again, then mindset, and finally asked about demographic information. Demographic information was not asked until the end to avoid stereotype threat [255]. The post-test measured learning outcomes for basic Python knowledge taught in Codeitz, adapting questions from [307, 224, 49].

For questions relating to learner experience, we focused on how learners decided what to learn next and how important different features of Codeitz were in their decision.

We used the following questions to analyze experiences:

1. After you were done with a lesson or exercise in Codeitz, how did you decide what to do next? (open response).

2. Think back to when you finished an exercise. How important were the following parts of Codeitz in deciding what to do next? (Likert-type, shown in Fig. 3.5).

3. Were there other parts of Codeitz that you considered when deciding what to do next? If so, please describe them and how important they were. (open).

4. If you remember seeing the blue recommendation text (pictured below), how did you use it to decide what to do next? (open).

5. What about using Codeitz caused you to feel frustrated, if anything? (open).

6. What about using Codeitz was helpful to you, if anything? (open).

Table 3.1: Data by condition. Sample size (n) includes number of low (↓) and high (↑) performers on post-test. Histograms of post-test score (max: 39.5), number of Codeitz exercises attempted (max: 44), and number completed (max: 43) shown with median ($\tilde{x}$) and interquartile range (iqr) (approx. to histogram bin).

cond.	n	test score	# attempted	# completed
IH	25 ↓: 7. ↑:9	 $\tilde{x}$:23.5; iqr:12	 $\tilde{x}$:43; iqr:1	 $\tilde{x}$: 43; iqr: 3
IL	31 ↓:12. ↑:12	 $\tilde{x}$:21.8; iqr:17	 $\tilde{x}$:29; iqr:34	 $\tilde{x}$:29; iqr:34
UH	23 ↓: 7. ↑:5	 $\tilde{x}$:23.0; iqr:14	 $\tilde{x}$:43; iqr:0	 $\tilde{x}$:43; iqr:2.5

3.5 Results: Experiences, Learning

To answer our research question of the effect of varying levels of agency and information to support agency on engagement and learning outcomes, we analyzed two aspects: 1) learners' experience in the three designs and 2) the outcomes of these experiences on learning.

3.5.1 Experiences varied by condition, performance

To analyze learners' experiences, we took two perspectives, first analyzing post-survey responses and log data on the use and perception of Codeitz features, and then analyzing learners' experiences between the three design alternatives.

Use and Perception of Agency Information

Figure 3.5 shows participants' ratings of the importance of design features in Codeitz across conditions (as described in Fig. 3.1). They rated these features on a five point Likert-type scale from "Not at all important" to "Extremely important." This scale also had a sixth "Not applicable" (N/A) option because some features were not present in some versions of Codeitz.

Qualitative and Likert-type survey responses suggested that the features available in all three conditions were generally viewed as valuable to learning. Participants in all conditions found the **progress indicators** (check marks and stars) denoting instruction and exercise completion to be

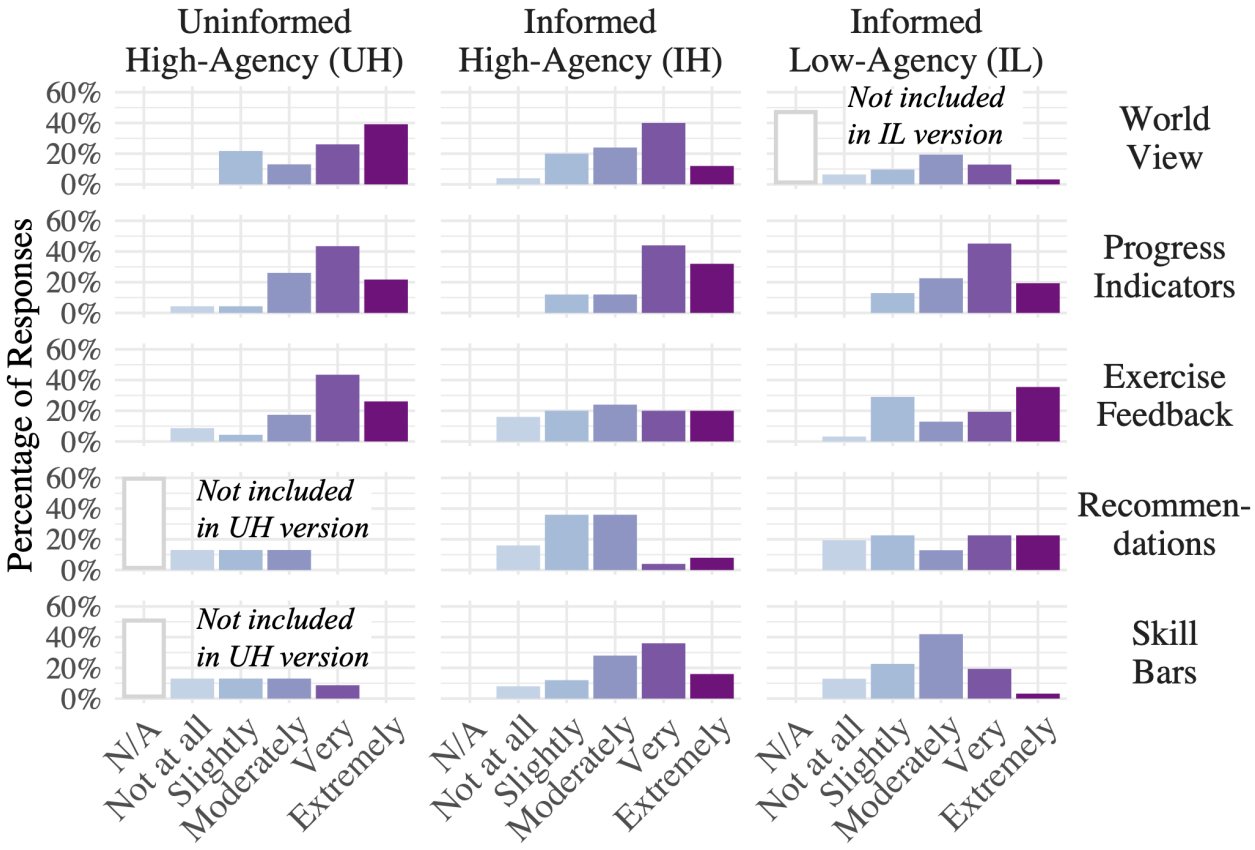


Figure 3.5: Importance of different features of Codeitz by condition. Not all features were present in each version of Codeitz (see Fig. 3.1).

helpful: of the 79 participants across all conditions, 89% found the progress indicators at least moderately helpful and 68% found them very or extremely helpful (see second row of Fig. 3.5). Few, however, reported how they used them. Despite **exercise feedback** in Codeitz consisting only of binary correctness feedback for at least 85% of exercises, participants still found the feedback valuable (third row Figure 3.5). But participants across all conditions tended to want more intermediate feedback on their exercises to help them persevere after getting incorrect exercise attempts, something that less than 15% of exercises had. P68 in the UH condition reflected this tendency to want intermediate feedback: "*hints and better feedback when you get an answer incorrect... would help me feel more confident about completing the task*."

IH and UH participants who had the **world view** reported it as important to guiding their learning: 22 of 23 UH participants and 11 of 25 IH participants reported using the world view to decide what to learn next. Most participants in both high-agency conditions (UH: 65%, IH: 52%) reported finding the world view very or extremely important. Many valued the world view for how it explicitly revealed dependencies ("*I like seeing the connection between concepts in the world view. That was very helpful to see how concepts fit together.*", P18, UH). Others noted that "*hidden dependencies*" between concepts caused confusion (e.g., "*the practice in the [Conditionals] has questions that need you to actually read Arithmetic Operators first before you can solve it... I need[ed] to go back to the other concepts before solving the exercise*", P1, IH).

Participants viewed the information provided only to the informed conditions (BKT-based recommendations and skill bars) as less valuable. Only 53% of the 56 (IH & IL) participants who saw them found **recommendations** at least moderately important to deciding what to learn next (Fig. 3.5). A common complaint was that recommendations tended to "*jump around*" (P72, IL) or "*jump too far*" (P51, IH), suggesting that recommendation behavior was unpredictable. One reason may be because of *cold-start* in which the recommender system initially had no information about a new user and therefore was more prone to making poor recommendations [176]. Recommendations did seem to improve as participants used Codeitz more and the system collected more data (e.g., "*...at first I would just click on any blue exercise without realizing they could require applying concepts that were entirely unfamiliar (this lead to some frustration), but eventually after solely using the*

[next] button to proceed I would only encounter exercises after I was sure I could complete them.", P74, IL).

Compared to recommendations, participants found **skill bars** as more valuable, with 63% of participants reporting them to be least moderately important. An IH high-performer reported using skill bars to determine what to learn next: "*I considered when the green skill bars would become 'advanced' which helped me know whether or not I should move on to the next topic.*" (P45, IH). Still, some participants were skeptical of the skill bar ratings: "*The skills levels, 'novice, expert, etc' varied depending on which section I was completing. For example, Even though the end section showed my reading skill as 'expert', if I clicked back to the initial section it showed it as a 'novice'. This made progress tracking feel a little empty, as it appeared to be simply used as visual feedback, not actual tracking*" (P55, IL). This unexpected behavior was because estimates of knowledge were localized to specific concepts and had no relationship to other concepts, even if those other concepts were dependencies (elaborated on later).

Learning Experiences by Condition and Outcomes

To understand how learners' experiences varied by conditions, we analyzed open-ended responses within each condition and compared the responses of participants who scored in the top 1/3 (*high performers*, $> 28.25/39.5$) and bottom 1/3 (*low performers*, < 18.5) on the post-test across all study participants; following the *contrasting groups* method of psychometrics [178, 179]. We conducted an open coding and thematic analysis [1] of post-survey responses from these contrasting groups, seeking to understand how they used features of Codeitz to decide what to learn next and what factors may help explain their performance on the post-test. We used log data (e.g. number of exercises attempted) to triangulate our findings.

Participants in the **Uninformed High-Agency (UH)** condition (N=23) tried to define their own learning trajectory to varying success. Without recommendations or skill bars, UH participants had to decide what to learn next using the world view showing them the concept hierarchy, progress indicators showing them what instructions and exercises they've completed (check marks for read instructions, stars for correct exercises), and exercise correctness feedback. Almost all (22) of the

23 UH participants explicitly mentioned using the world view to decide "*I looked at the flowchart and picked a connecting branch.*" (P57). Overall, participants in the UH condition tended to attempt all the exercises, with 78% attempting all the exercises and 65% getting them all correct (all participants were allowed unlimited retries).

High performers in the UH condition all reported exerting agency. While the UH condition did not have recommendations or skill bars to inform learners, some of the 5 high-performers in this condition noted still being able to deviate from the world view's explicit paths in ways that benefited their learning. For example, 2 high-performers noted trying exercises and then reviewing lessons if they were unfamiliar with the exercise. Another skipped to trying exercises first but jumped back to reading lessons when necessary: "*I mostly skipped around to the exercises, if I felt like I could understand what was going on in the lesson, and then moved back through the lesson if I couldn't do an exercise.*" (P5). The remaining 3 high-performers reported finding Codeitz the structure and presentation of the curriculum helpful, the design of the curriculum as intuitive: "*The lessons were easy to understand and exercises helped cement the knowledge*" (P21).

In contrast to UH high performers, the 7 **UH low performers** noted struggling to navigate their learning experience. For example, P39 was frustrated because there was *"no proper path"* and P63 felt "*the order [of concepts] did not seem intuitive.*" P18, who reported minimal programming self-efficacy prior to the study, noted how his confidence affected what he chose to learn next: "*To decide which lesson I would try next from a lower level, particularly the second level, I looked at which concept I felt most confident taking on first.* (P18). Low-performing UH participants also noted wanting additional instructional content, such as "*video explanation*" (P3) and more feedback in code writing exercises (P9).

Participants in the **Informed High-Agency (IH)** condition (N=25) reported similar experiences to the UH condition, with additional comments about the recommendations and skill bars. To decide what to learn next, 11 of the 25 IH participants reported using the world view, 10 reported following the recommendations, and 2 reported trying the recommendations but then abandoning them. So whereas almost all UH participants reported following the world view, less than half of the IH participants reported doing so, with 28% reporting at least trying to follow the recommen-

dations. Overall, participants in the IH condition also tended to attempt all the exercises, with 64% attempting all the exercises and 56% of IH participants getting them all the correct.

IH high performers reported evolving interpretations of the recommendations as they deviated from the world view's prescribed paths. Three of the 9 high-performing IH participants reported using the recommendations; of those 3, 2 of them reporting trying the recommendations at first, but then abandoning them because they led them to exercises that were too advanced: "*At first, I looked at the blue highlighted boxes. However, I felt like it made me jump too far. For example, one lesson had started talking about if statements but I hadn't learned the syntax for those yet. So then I just followed the tree from top down, left to right.*" (P25). Three high-performing IH participants ignored the recommendations because they wanted to complete all of the curriculum. P1 noted how he ignored the recommendation but how he could see its benefit for less motivated learners: "*I did not really pay attention to the blue recommendation because I was motivated to do all the lessons and the practices to receive maximum knowledge. I think this blue recommendation might be useful for people who have low motivation to do more exercises.*" Log data confirmed tendency to do all exercises, as all high-performing IH participants completed at least 40/43 of the exercises (93%), although this complete coverage of exercises was consistent across the entire IH condition.

In contrast to the high-performing IH participants, the 7 **low-performing IH** participants tended to ignore or misinterpret the recommendations and followed a perceived intended path. Three participants ignored the recommendations, such as P24 who "*usually just did the problem even though it was review.*" Another participant was confused by the recommendations updating: "*... sometimes [a recommendation] would be there and sometimes it would not. Typically I would see this after I finished in exercise.*" (P38). Of the two participants who reported using the recommendations, P43 reported that he "*did what [the recommendation] said*" while P23 used the recommendation to estimate how much time an exercise would take: "*While having limited free time, it was helpful to see a note indicating that the next tab was a review exercise, meaning it would likely be quick to complete.*" (P23). Three low performing IH participants struggled with having to choose their own trajectory. P38 reflected this in her description of how she got lost and found choosing what to learn next frustrating: "*I found the layout of what lessons to take were confusing. I went from*

top to bottom and left to right, however, during the exercises I would find myself lost on multiple occasions. This would be due to either skipping sections but having to use it before I learned the material. Instead of having the choice of choosing what lesson to take next, it would have been more helpful if it was just given (P38).

Participants in the third and final **Informed Low-Agency (IL)** condition (N=25) only had three choices at any given time: completing the given exercise, reviewing instruction related to the exercise, or reviewing prior lessons and exercises. Three participants found this lack of overview made it challenging to keep track of how much they had completed, how much remained, and how the concepts related to each other. P72 reported his challenges of keeping track of where he was in his learning process: "*Everything seemed to jump around and it was hard to keep track of what I was on or what I was supposed to do next.*" (P72). Overall, participants in the IL condition tended to attempt fewer exercises than the high-agency conditions: whereas both high-agency conditions had most (78% for UH, 64% for IH) participants attempt all 43 exercises, only 41% of participants in the IL condition attempted all the exercises. All 41% of those participants did get all exercises correct, though.

The 12 **high-performing IL** participants varied in how they interpreted the next exercise presented to them. Half (6) reported viewing recommendations as indicators of an exercise or a required next step: *When I saw this blue recommendation, I would make sure to click it in order to complete it as it seemed to mean 'required'* (P42). Three others reported using the recommendation text as informative in deciding whether to attempt an exercise, read instruction, or go back to a previous exercise: *I only used it to see how difficult the exercise was. I would still go straight to the exercise even it told me I hadn't learned about the concept, and I will come back to it later if I didn't figure it out. If it told me it's something I had already learned, I wouldn't leave the exercise until I figured it out.* (P35). The remaining 3 reported not using recommendations (1) or did not comment on their usage (2). Multiple participants reported a desire to have an overview of all concepts and explore concepts more freely: "I didn't know how many topic there are in total and could only view them after doing the previous topic and unlocking it. I feel it would be better if I can see how much I am completing and how much still has to be done" (P75).

Table 3.2: Coefficients of linear regression to model learning outcomes (post-test scores). *** indicates $p < 0.001$, * that $p < 0.05$, & . that $p < 0.10$.

coefficient	estimate (std. err.)	*t*	Pr(>|t|)
(Intercept)	14.41 (2.72)	5.301	0.000 ***
condition: IL	0.70 (2.58)	0.270	0.787
condition: IH	0.90 (2.74)	0.328	0.744
self-efficacy (pre)	2.43 (1.02)	2.387	0.020 *
taken CS course	4.62 (2.52)	1.836	0.070 .

In contrast to the high-performing IL participants, the 12 **low-performing IL** participants reported relying much less on the recommendations. Three reported not even seeing or noticing the recommendations. Of the 4 low-performing IL participants that mentioned using recommendations, two saw them as indicators of "*a signal that the selected block [was] an exercise*" (P61). One participant "*used [the recommendation] as an indicator for a concept practice/challenge*" and that "*the practice challenges were very helpful... in learning python*" (P61). Low-performing IL participants also completed fewer exercises: Only 25% (4) of these participants attempted more than half of the exercises; in contrast, 81% (9 of 11) high-performing IL participants completed more than half of the exercises.

3.5.2 Learning & Exercise Completion by Condition

Condition, Self-efficacy, & Prior Knowledge on Post-Test

To understand how the varying designs of Codeitz conditions affected learning, we used a linear regression to model post-test scores. In addition to passing into the regression the condition participants were in (UH, IH, IL), we also considered self-efficacy prior to using Codeitz (range: 1-7) and whether a participant reported taking a prior CS/programming course (true/false), as both self-efficacy and prior knowledge are important to learning [101]. We found no violations of linear regression assumptions: normality (Shapiro-Wilk, p=0.23), homoscedacity (spread-location plot), and linearity [35, 145, 243]. Table 3.2 shows the coefficients of the linear regression.

Table 3.3 shows the results of a linear regression model analysis of variance (ANOVA). The

Table 3.3: ANOVA results and effect sizes for linear regression of post-test scores. ε denotes a small positive value ($0.001 - 0.004$). * indicates that $p < 0.05$.

variable (df)	SE	F	Pr(>F)	η^2 [95% C.I.]
condition (2)	45	0.3	0.776	0.006 [0, 0.08]
self-eff, pre (1)	499	5.7	0.020 *	0.066 [ε, 0.21]
taken CS course (1)	583	6.7	0.012 *	0.077 [ε, 0.23]
residuals (74)	6483			

ANOVA indicated a statistically significant effect on post-test scores of prior self-efficacy ($F(1,74) = 5.7, p < 0.05$). Whether a participant had previously taken a programming course was also had a statistically significant effect ($F(1,74) = 6.7, p < 0.05$). Both significant factors had medium effect sizes ($\eta^2 > 0.06$) with large confidence intervals which did not include zero. The condition participants were in did not have a statistically significant effect ($F(2,74) = 0.3, n.s.$).

We conducted non-parametric post-hoc analyses to understand how prior self-efficacy and programming course experience affected post-test score. The median post-test score of participants who had taken a prior programming course was 29.12 (IQR = 13.4) and of participants who had not was 21.50 (IQR = 15.3). This difference was statistically significant according to a Mann-Whitney U test ($U = 364.5$, $p = 0.012 < 0.05$). We interpreted the medium Vargha and Delaney A effect size to state that there is 69.1% chance a post-test score for a random participant who has taken a programming course will be greater than a score for a random participant who has not [287]. For self-efficacy, we calculated Kendall's non-parametric rank correlation [154]. We found a significant correlation ($\tau = 0.25, p = 0.0014 < 0.01$) between prior programming self-efficacy and post-test score. We convert τ to $r = 0.38$ [291, 155] and identified a medium effect size between prior self-efficacy and post-test score [243].

Number of Exercises Completed by Condition

To check for a difference in the number of exercises completed by condition, we conducted a Kruskal-Wallis test [243]. We decided on this non-parametric test because the data was not normal (Shaprio-Wilks: $p < 0.05$). Table 3.1 shows the distribution, median, and IQR for the number

of completed questions by condition. We found statistically significant differences in number of completed exercises between conditions ($\chi^2(2, N = 79) = 11.33, p = 0.003 < 0.01$.

We conducted a pairwise post-hoc analysis with Mann-Whitney U tests with Holm correction. We found that a statistically significant difference between the IL condition and the other two conditions (Mann-Whitney U for IL/UH: $U = 504, p = 0.014 < 0.05$; IL/IH: $U = 225, p = 0.014 < 0.05$). We can interpret the medium Vargha and Delaney A effect sizes to say that there is a 71% chance that a random UH participant completed more Codeitz exercises than a random IL participant, and that there is a 71% a random IH participant completed more exercises than a random IL participant.

3.6 Discussion: Interpretations & Implications

The objective of this study was to jointly understand how affording and informing agency affected engagement and learning outcomes. We did so by designing three variations of a self-directed online learning environment that varied the amount of agency or information afforded to participants as they learned introductory Python.

We found that the specific features offered in these three conditions led to very different learning experiences and degrees of engagement, but that these differing experiences led to no detectable effect on learning outcomes. We also found that low-agency (IL) participants completed significantly fewer exercises than high-agency (IH, IL) ones.

There are multiple ways to interpret these findings related to learner experience and learning outcomes. One interpretation is that our recommendations were not "intelligent" enough to be helpful. Our BKT implementation faced challenges such as parameter tuning [128] and cold-start [176], as consistent with most statistical models. While we did our best to fit parameters according to best practices and given the response data we had available, we also recognized that better parameters could improve the performance of BKT. But Codeitz is representative of a discretionary use self-directed online learning environment in that recommendations and item selection will never be perfect or optimal for all learners, especially early on before we have a large corpus of response data. So understanding how to design information such as adaptive recommendations to

affect agency and learning also requires understanding how learners interpret and use information that come with the inevitable imperfections and inaccuracies of data-driven adaptation. And despite many participants feeling like the recommendations "jumped around" and were not always accurate, participants in both high and low agency conditions still found ways to use them to inform their decision-making process and learning. So while our BKT model had identified problems, participants could still use information from it. And participants' experiences and reports can help better inform how we design adaptive online environments that promote learner agency.

Another interpretation is that other confounds made our post-test an invalid or unreliable measure of learning. Because we wanted to investigate the design of self-directed online learning, we set up our study such that it could emulate this discretionary, informal learning. We did so by having participants learn on their own time across the span of a week and then take the post-test whenever they felt ready. While this experimental design introduced confounds including variation in amount of time spent learning and an uncontrolled test-taking environment, they were externally valid to many online learning environments (e.g. MOOCs, online coding platforms, remote/hybrid courses). Such confounds were also distributed across the conditions. Furthermore, post-test items came from concept inventories [224] or were piloted with representative users with think-aloud [95].

3.6.1 Design Considerations and Future Work

A third interpretation of our findings is that designing for agency is nuanced and requires careful design considerations we are only beginning to understand. While prior work investigated varying agency to measure its effect on learning, we designed and varied the information and agency afforded to learners. Our results suggest possible explanations and design considerations to explore in future work:

The value of agency may be dependent on the structure of knowledge to learn. Relating to programming, this study used learning materials with concepts that have rigid hierarchical relationships. This knowledge domain may lend itself to more linear instructional content. Agency to support learning here may be in the form of jumping ahead for a challenge or back for review. In

contrast, learning to use programming for expression (e.g. with Scratch) may lend itself more to non-linear instruction. Agency in this case may be in the form or exploration of one of many paths. Therefore, how to afford agency may be dependent on the structure of the knowledge domains and learning objectives.

Agency may be valuable to more than just learning outcomes. Our findings suggest that agency may support motivation to continue learning. Prior work has generally found more agency relating to increased motivation (e.g. [251, 66, 250, 211, 114]). Our findings suggest that there was a 71% chance a random high-agency participant completed more exercises than a low-agency one. This suggested that affording learners the agency to see everything there was to learn (with the World View) and choose for themselves may have had a motivational benefit to help learners continue to engage.

Recommendations may have different roles to different learners. In Codeitz, we intended for the recommendations (Fig. 3.1d) to be cues to exert agency. While learners in the informed high-agency condition tried to use the recommendations to guide them, many treated the recommendations not as cues as to what to expect from a given exercise, but simply as indicators of an incomplete exercise. It may be important to consider not only the intended role of cues to inform agency, but also to consider alternatives ways learners may interpret them initially as well as after some interactions.

Consider learners' prior experiences with related tools. Just as learners come with prior perceptions related to what they are learning, they also come with prior perceptions related to how to interact with learning environments. While we designed Codeitz to not have an apparent order in high-agency conditions, we found that a majority of participants (in the UH condition especially) reported following or trying to follow an intended order. Such behavior might prevent any potential benefits from exercising agency from materializing.

Overviews, while valuable, may indirectly constrain learners' decisions. Learners found value in Codeitz's world view, suggesting it provided an integrated view of concepts to be learned. But the overview might have also acted as an *indirect control* [254], limiting agency. Designs may need to consider the unintended side effects of offering conceptual overviews on how learners choose to

sequence their learning.

Our evidence, and these possible interpretations, suggest that designing for agency, and in particular, designing information that encourages agency, is far from straightforward. Just as offering choice is not consistently beneficial to learners, offering information to support those choices is not consistently beneficial either. Future work should explore with more granularity the interaction between self-directed learning environments, learners' interpretation of what the environment provides, and learning outcomes. And designers should be wary about the benefits of learner agency, and pay close attention to the specific domain of learning and the specific unintended side effects of how learners use the affordances in a self-directed learning environment.

3.7 Conclusion: Consideration of prior knowledge and power relationships required

Connecting the findings of this design exploration back to the framework I defined in Fig. 2.2, this study suggested that the design of Codeitz did not effectively consider students' *relevant prior knowledge* and *perceptions of power relationships* in regards to exercising agency in a learning experience. Most participants were familiar with formal higher education learning experiences, where instructional design often left little room for self-directed learning. So participants were likely unfamiliar with exercising agency, resulting in them disregarding data that could have informed their self-directed learning.

An alternative design that better engaged participants' prior knowledge and perceptions of power relationships could have better support the exercising of agency. Such designs could have prompted participants on the benefits of exercising agency and provide opportunities to practice exercising agency. This design exploration suggested that designing interactions with data that engage with students' prior knowledge and perceptions of power relationships could better support students in their use of the data, especially to take unfamiliar actions.

Chapter 4

DIFFERENTIAL ITEM FUNCTIONING (DIF) TO DETECT POTENTIAL BIAS IN TEST QUESTIONS

In this chapter, I describe a study that investigated how to support the equity-oriented goal of addressing biases in test scores. I used demographic and psychometric data on students' gender, ethnicity, and responses to assessments from the online Code.org middle school computing curriculum to identify *Differential Item Functioning* (DIF), or empirical evidence of potential test question bias (e.g. by gender or race) [290, 138, 99, 190].

I then conducted a workshop with a group of curriculum designers to understand how they interpreted DIF data. From their discussions, I discovered how their prior training and lived experiences affected their prior knowledge and cultural competence. I also discovered how their primary roles as curriculum designers affected their perceptions of power relationships. Publicly with colleagues, designers tended to interpret DIF data within the controlable context of what changes to curriculum and test design they could make. But more privately, designers considered how DIF may just be evidence of systemic injustices that go beyond the curriculum and test design that they can control.

The work in this chapter was conducted in collaboration with Matt J. Davidson, Baker Franke, Emily McLeod, Dr. Min Li, and Dr. Amy J. Ko and published to 2021 ACM Learning @ Scale [306]. [1] [2]

---

[1]This study was pre-approved with exempt status by the UW Institutional Review Board (IRB) as STUDY00010647. Code.org affiliates reviewed this publication and proposed changes prior to public release. This material is based upon work supported by the National Science Foundation under Grant No. 1735123, 12566082, 2031265, 1703304, 1539179, and unrestricted gifts from Microsoft, Adobe, and Google. Supplemental material can be found at `https://github.com/codeandcognition/archive-2021las-xie`.

[2]I will use "we" instead of "I" in this chapter to acknowledge the shared contributions of all authors.

4.1 Introduction: How DIF can improve equity

Successful learning requires equity, which can be viewed as access to and successful participation in education within economic, social, cultural, and political contexts of a given time and place [103]. Equity also implies a goal of implementing corrective measures to adjust for aggregate harm from historic social inequalities [231]. This might mean providing additional and personalized support to students from minoritized groups ([19, 305, 143]), as they face unique challenges that if left unaddressed could pose serious impediments to science and technology learning [69]. Previous efforts to design equitable learning experiences include designing adaptive and personalized online environments [33, 309], adjusting environments to support inclusion [160], and enabling broader access [165].

However, achieving equity is rarely straightforward: inequities in learning stem from a complex interplay between multiple structures and interactions [259]. Student achievement is not a static construct that we can measure in isolation, but rather impacted by characteristics of and interactions between students, classrooms, and school contexts [206, 67, 100].

Because of the complexity of context, even gathering information about the presence of inequities is hard. It requires understanding where needs and gaps exist to target support [98, 104]. Students alone cannot be responsible for identifying equity issues because their focus is on learning [203, 147] and self-advocacy may bring about burden and risks to minoritized groups including stereotype threat [300, 86] and social-desirability biases [123, 108]. Teachers have a significant role in addressing inequities [259], but they would need information that is understandable and actionable [26, 104] and often work within the constraints of pre-defined learning objectives and materials.

While equality and equity are different concepts, equity cannot exist without first assessing inequality to consider how to appropriately adjust resources [231]. Data can support equity by enabling rapid improvement of practices through experimentation and measurement of change that is understandable and actionable [26]. Large-scale analysis can reveal patterns not easily seen at a micro-level by individuals [102], such as through analysis of intersectional identities [248]. Data

can provide evidence to support disruption of the status quo [104].

Connecting data on inequalities with domain experts' contextual knowledge to identify equity issues can help, but current methods to do so have slow feedback cycles, require custom testing infrastructure, or rely on metrics that are difficult to interpret in the context of learning. Participatory approaches such as action research and design-based research can help deeply understand a phenomena, but they are costly in time and resources to conduct [45, 152, 282]. Quantitative approaches such as data mining techniques require technical infrastructure to set up and rely on tracking specific metrics that may lead to ignoring broader and potentially more important patterns that cannot be measured [102]. Improving equity in learning experiences is a complex and iterative process that requires a multitude of methods and stakeholders' expertise [274].

One way to measure inequality in a learning experience is by using *differential item functioning (DIF)* to measure how students of diverse identity groups perform on formative assessments. A common formative use of assessments by teachers and students is to measure student understanding by re-exposing them to key content [225, 209]. A fair assessment would measure differences in students knowledge/understanding of a domain (e.g. computing and programming) without being affected by differences in identity-based factors (e.g. gender, race). DIF methods determine the fairness of test questions by determining to what extent test-takers with similar knowledge levels but differing group membership (e.g. different genders) perform similarly on questions [190, 78].

DIF analyses suggests potential bias, but judgement is required to interpret and act on DIF results [313]. And because domain experts such as curriculum designers have contextual expertise, they are well-positioned to interpret and use DIF analysis to enact change by revising instructional materials that support more equitable learning. So we explored how DIF results on potential item bias by gender and race in formative assessments could inform domain-experts, thereby connecting quantitative DIF data with contextualized knowledge of domain-experts to support equitable curriculum changes.

In this study, we investigated how domain experts (curriculum designers) interpreted quantitative measures of bias in formative assessment items to augment their existing contextual knowledge and support understandings of equity challenges in an online CS curriculum. To do so, we worked

with response data and curriculum designers from Code.org, a nonprofit dedicated to expanding access to computer science and increasing participation of young women and students from other underrepresented groups. We explored the following question: **How do curriculum designers interpret and use data on potential assessment item bias by gender and race in the context of designing equitable learning experiences?** To answer this question, we conducted a quantitative analysis of assessment bias that was among the first to include students who reported as non-binary and also the largest (by sample size) in the computing education domain. We analyzed 139,097 responses to Code.org's globally deployed online Computer Science Discoveries (CSD) curriculum to identify potential assessment bias (dis)advantaging students of different genders and races even after matching them by knowledge level. We then partnered with Code.org curriculum designers to understand how they interpreted the data within their domain expertise. We discussed our findings in the context of a broader vision of using data to augment domain experts' capabilities to design equitable learning experiences.

4.2 Background: Overview of DIF methods

DIF methods were originally developed to address concern that ability tests were unfair to minority test-takers, and has become a standard part of operational screening at testing companies [292]. They measure the fairness of a test question by determining to what extent test-takers with similar knowledge levels but different groups (e.g. different binary genders [76, 190, 80]) perform similarly on a given test question. Developers of concept inventories (tests used as a standardized measure of conceptual understanding [174]) often use DIF methods to understand how fair an inventory is (e.g. [190, 80]). Developers of high-stakes large-scale tests use them to identify items that exhibit DIF and remove these items from a pool of potential items to avoid test scores advantaging certain groups [314, 84]. Researchers have used DIF to detect potential bias in summative evaluations (e.g. final exams) in large courses, such as introductory CS [76]. Rather than addressing equity more broadly, DIF is a narrow analysis of potential bias in item performance; therefore, DIF can help detect potential bias in test questions, but does not provide much insight on the causes of such unfairness [84]. In all these cases, psychometricians with expertise in educational

and psychological measurement typically conduct the analysis and interpret the findings.

DIF occurs when people of approximately equal knowledge from different identity groups perform in substantially different ways on an item. DIF methods provide information on measurement invariance, allowing one to judge whether items (and ultimately a test as a whole) are functioning in the same manner for different groups of test-takers [317]. DIF methods work by 1) designating a reference group and a focal group, 2) matching test-takers of similar knowledge and skill from different groups, and then 3) measuring DIF between groups of test-takers for each item in a test. DIF is often used to compare between test-takers of different genders (e.g. women as focal group, men as reference group) and races [317].

4.2.1 Three classes of DIF methods

DIF methods differ in how they model item responses and match test-takers of different groups. We can use DIF methods to detect *uniform DIF* in which an item disadvantages a group of students uniformally across all knowledge levels as well as *nonuniform DIF* in which the DIF interacts with the knowledge levels of students and the groups they are in. At least three frameworks for investigating DIF exist [317]:

Modeling responses with contingency tables, regression models: This class of DIF methods consists of conditional effects that study the effect of grouping variables and interaction terms while conditioning on the total score of a test. After conditioning on differences in item responses due to differences in knowledge being measured, DIF exists if item responses for different groups still differ. This difference can be a main effect of group differences (*uniform DIF*) or an interaction between group and knowledge (*nonuniform DIF*).

Item response theory (IRT): For IRT methods [78, 307], DIF exists if item *trace lines* are different between groups. IRT methods measure DIF as the area between logistic trace lines (or equivalently, comparing parameters such as difficulty and discrimination). IRT approaches match items not on total score but on latent variable modeling, so the scale for knowledge level of students and item difficulty (θ) is arbitrary and must be calibrated. Examples include signed area tests (for uniform DIF), unsigned area tests (which allow for nonuniform DIF), and nested model testing via

a likelihood ratio tests.

Multidimensional models: These types of methods relax the common undimensionality assumption that a test measures a single latent factor. Instead, these types of methods assumes that tests are, to some extent, multidimensional (e.g. a test to measure programming skills also measures another dimension such as reading comprehension). Simultaneous item bias tests (SIBTEST) DIF detection methods are an example of multidimensional methods. Because these methods involve a type of factor analysis, they require analysis of sets of items, rather than individual items for DIF. Multidimensional models have also incorporated contextual and sociological variables [318].

4.2.2 Interpretations and uses of DIF

A question that exhibits DIF disadvantages a certain group (e.g. women students) and may warrant follow-up analysis to determine whether the question should be revised or removed [138]. Within the context of computing education, Davidson et al. 2021 demonstrated the use of DIF to identify potential unfairness in an introductory CS exam, arguing for more broad use of DIF in the validation process of CS assessments [76].

Organizations instituting high-stakes testing (e.g. Educational Testing Service) have used DIF analysis to categorize questions according to fairness, identify topics and contexts to avoid in question design, and adjust test scores if they discovered that some questions exhibited DIF after test administration [314]. Because of test security requirements, they typically rely on review of items by expert psychometricians.

But DIF and bias are *not* synonymous. DIF does not prove bias, and the lack of DIF does not prove lack of bias [313]. Judgement is required to act on results of DIF analysis and address potential bias issues, but prior work focused on contributions of psychometric experts revising high-stakes tests. This paper is the first to consider DIF interpretation by stakeholders who are not psychometric experts, which is critical to test validity because the fairness of a question depends on how instructors, students, and other stakeholders interpret and use scores [149]. So this study explored how curriculum designers used DIF statistics to better understand what knowledge and skills their tests were trying to measure and understand common sources of DIF that confounded

that measurement. By doing so, we can understand the feasibility of a new use of DIF, where domain experts such as curriculum designers may be able to contextualize DIF results to make informed judgements that equitably improve their assessments and curriculum.

4.3 Context: CSD curriculum & assessment design

To understand how curriculum designers interpreted and used data on potential assessment bias, we analyzed responses from Code.org's Computer Science Discoveries (CSD) 2019-2020 course [52]. CSD is for 6-10th grade students, with the median age of students in our sample being 13 years old and 86% of students being 11-16 years old. Mapped to the Computer Science Teachers Association standards, CSD took a wide lens on CS, covering topics including problem solving, programming, user-centered design, and data. CSD was typically used for in-person, synchronous instruction led by a teacher. Designers wrote CSD for "*new-to-CS teachers*"[52].

The CSD curriculum guide recommended that a typical 10-12 week term cover Units 1-3 (of 6), which covered problem solving, web development, and interactive animations and games. Unit 1 focused on the problem solving processes where students learned to use a structured problem solving process to address problems and design solutions that used computing technology. For the unit's final project, students proposed an app to solve a problem of their choosing. Unit 2 focused on web development, where students learned HTML and CSS to create and style content, how different languages allowed them to solve different problems, and how solutions could generalize. Students used Code.org's Web Lab programming environment to create personal portfolio websites for their final project. Unit 3 taught students to create interactive animations by using basic programming constructs (control structures, variables, user input, randomness). Students used Code.org's hybrid blocks and JavaScript programming environment to design games with animated sprites. Taken together, these units taught programming as a fun and creative form of expression.

Each unit ended with a post-project assessment. A post-project assessment included four to seven multiple choice and matching questions, as well as three open ended reflections on the final project of the unit (which we did not analyze). These tests aligned to the learning framework of each unit and were designed to assess parts of the framework that may not have been covered by the

project rubrics. Teachers must decide to enable post-project assessments for students to even see the assessment. The curriculum guide left it up to teachers to decide how to use assessments (e.g. for formative feedback or summative grading), but curriculum designers we interviewed stated the assessment was for formative purposes. Students could only submit each post-project assessment once.

For our analysis, we focused on the multiple choice and matching questions because they had dichotomous correctness (graded as entirely correct or incorrect) that enabled modeling with traditional psychometric techniques. Multiple choice items required students to select one or two options from five options (scored as correct only if all correct options selected but not incorrect ones). Match questions required students to correctly place four or five options in the correct location (e.g. placing comments in appropriate locations in the code). Table 4.2 describes the items.

4.4 Quantitative Analysis with DIF Analysis

We conducted a psychometric analysis to understand how effectively dichotomous items in the post-project assessments for CSD Units 1-3 measured students' understanding. In this section, we describe the response and demographic data we analyzed, provide basic item statistics, examine IRT assumptions, and then report race-based and gender-based DIF.

4.4.1 Data: 6 - 10th graders' demographics & test responses

For our analysis, we focused on Units 1-3 because the curriculum guide recommended them and they had the most responses (>10% of students in sample responded to each item). We analyzed 139,097 responses from 19,617 students who used CSD for the 2019-2020 academic year and reported both gender and race. Table 4.1 shows reported demographics for students.

Because this was an optional formative assessment, responses were sparse. Of the 333,489 potential responses (86,584 students to 17 items), students only provided 139,097 responses (41.7%). Of the 139,037 provided responses, 64,481 were scored as correct (46%) an the remaining 74,616 (54%) were scored as wrong. We reported proportions of students not responding to each item in

Table 4.1: Reported gender and race. Students could report one gender and one or many races.

	female	male	non-binary	total
African American/ Black	2,549	3,253	49	5,851
Hispanic/Latinx	1,736	2,640	52	4,428
Native American/ Alaskan Native	365	542	18	925
Pacific Islander/ Native Hawaiian	150	244	9	403
white	3,455	6,211	96	9,762
Asian	470	997	27	1,494
total	7,469	11,953	195	19,617

the *NR* column of Table 4.2.

4.4.2 Item statistics & reliability are acceptable

DIF methods analyze item-level responses, so we report classical test theory (CTT) item statistics including difficulty, discrimination, and reliability. CTT statistics are common, simple, and provide limited but useful information about the quality of a measurement instrument [4], shown in Table 4.2. *Difficulty* is the proportion of respondents getting an item correct, with a lower number indicating a more challenging item. Difficulty ranged from 0.27 (*U3, Q5*) to 0.75 (U3, Q4), with 10 of 17 items having a difficulty of < 0.50. Furthermore, three multiple choice items had an incorrect option (known as a *distractor*) selected more frequently than the correct response ($\Diamond$ in Table 4.2), which may be problematic. This assessment was fairly challenging. We used point-biserial correlation (r_{pbis} to measure of *discrimination*, or how effectively an item differentiates a test-taker of higher knowledge from one with lower knowledge. It is an association between a response to a single item and the overall score [4, 78]. r_{pbis} can range from -1.0 to 1.0 but should always be > 0, with $r_{pbis} > 0.3$ being considered acceptable. Only one item, fell below this threshold (*U3, Q5*, $r_{pbis} = 0.27$), suggesting items had acceptable discrimination. We used change in Cronbach's α to

judge change in internal-consistency *reliability*. The test as a whole had a Cronbach's $\alpha = 0.732$, which is acceptable for low-stakes formative use [169, 214]. Removing any of the 17 items resulted in a decrease in α ($\Delta\alpha < 0$), so we analyzed all items.

4.4.3 Three IRT assumptions mostly hold

To use Item Response Theory (IRT), we must first confirm its three assumptions of *conditional independence*, *unidimensionality*, and *functional form* [78]. The conditional independence (or *local independence*) assumption states that responses to an item are independent of responses to any other item, conditional on a person's knowledge. That is to say that there is no interdependency between items. Justifying the conditional independence assumption requires looking at the design and implementation of the test. The test did not have a time limit, so speededness likely did not affect test-takers responses. And with the exception of two items (Unit 2, Question 5 & Unit 2, Question 6), no items referenced shared information. U2, Q5 and U2, Q6 both referenced the same image of code. While this is a violation of unidimensionality, we justified keeping these items in the data because they were the only interdependent items and simulation studies have shown that, when only a small number of items violate this assumption, removing those items leads to more biased estimates [71]. Our choice was also justified by the results of factor analysis.

To verify unidimensionality, we conducted exploratory and confirmatory factor analysis. Exploratory factor analysis suggested a single factor according to the eigenvalues > 1 criterion [264]. Confirmatory analysis with one factor showed a strong model fit (RMSEA = 0.018, CFI = 0.916, TLI = 0.903) [34].[3]

Verifying the functional form assumption involves comparison of multiple models to see which one best fits the data. We fitted IRT models with one (1PL), two (2PL), and three (3PL) parameters. The 1PL model has a difficulty parameter and assumes all items share the same discrimination value. The 2PL model has a difficulty and discrimination parameter. The 3PL model is a 2PL

[3]EFA was conducted using R with `psych::fa()` [240] using a maximum likelihood factor analysis with a varimax oblique transformation. CFA was conducted with `lavaan::cfa()` [249] with fixed residual variances (`std.lv=T`) and full information maximum likelihood (FIML) approach for handling missing data.

with an additional parameter to account for guessing. We compared model fit using the Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) [78]. While the 3PL had the lowest AIC (706353) and BIC (706831), and 1PL had the highest (AIC=710885, BIC=711053), we ended up selecting the 2PL model (AIC=708878, BIC=709196) because of model fitting issues relating to data sparsity when grouping by reported gender and race for DIF analysis.

4.4.4 Results: Checking for DIF by gender & race

For this study, we used an IRT method for detecting DIF by reported gender and race. We used a Likelihood Ratio Test (LRT) DIF analysis [42] with a 2PL model because total score was arbitrary and there is significant non-response (so contingency tables and/or regression models would be less appropriate). To adjust for multiple comparisons, we used a Benjamini-Hochberg p-value correction [21], an adjustment that maximizes power while controlling the false discovery rate to the nominal value (in this case, 5%) [190]. While more advanced DIF methods enable comparison of a reference group to multiple focal groups (e.g. [303, 302]), we could only compare two groups (single reference group, single focal group) at a time because of limitations of data related to sparsity of responses and few students in the focal groups. Specifically, we used LRT DIF to check for DIF between students who reported as non-binary, female, and male through three pairwise comparisons. We also checked for DIF between AHNP[4] (African/Black, Hispanic/Latinx, Native American/Alaskan Native, Pacific Islander) and WA (white, Asian) students. We choose these groupings because AHNP racial groups tended to be minoritized in computing education [143, 86, 319, 278, 94], and WA racial groups tended to be dominant [299, 247, 186, 88].

Because we used a 2PL model, DIF would manifest as groups having statistically significant differences in difficulty and/or discrimination parameters. For LRT DIF, we used a χ^2 statistic and p-value to determine whether groups had significantly different parameters. To measure effect size, we used the signed in-sample differences (SIDS) and unsigned in-sample differences (UIDS) [196]. Because these are dichotomous items scored as 0 or 1, we can interpret SIDS to be the average

[4]We interpreted AHNP to be equivalent to *BIPOC* (Black, Indigenous, people of color). We referred directly to ethnic groups instead of using new labels to avoid ambiguity and potential harm [301].

difference in probability of selecting the correct answer between groups. We considered SIDS and UIDS values of $0 - 0.05$ to have a negligible effect, $0.05 - 0.10$ to have medium/ intermediate effect, and > 0.10 to have a large effect [85].

For LRT DIF with a 2PL model, uniform DIF indicates a difference in difficulty parameters between groups, while non-uniform DIF indicates a difference in discrimination parameters. An item with uniform DIF disadvantages the group with a significantly *greater* difficulty parameter. If a model exhibited *non*-uniform DIF (item disadvantages groups differently based on different knowledge levels), then we would expect the discrimination parameters to be statistically significant between groups (but not necessarily the difficulty parameter); the SIDS and UIDS would likely be different. For non-uniform DIF, the item traces for the two groups would be two logistic curves of different slopes that intersected at some point.

DIF for gender: Uniform DIF favors reported male, non-binary

Table 4.3 shows the results of three pairwise comparisons for each item to understand DIF between students who reported as non-binary, female, and male. We found that six test items disadvantaged students who reported as female (compared to reported male and/or non-binary students), one item disadvantaged reported male students (compared to non-binary), and no items disadvantaged reported non-binary students.

Table 4.3 shows difficulty and discrimination parameters for students of different reported genders, as well as effect size (abbreviated as *e.s.*). When comparing students who reported as **female and male** (blue rows in Table 4.3), we found that two items (*U1, Q3*; *U3, Q6*) exhibited uniform DIF with a non-negligible e.s. Both items had significantly greater difficulty parameters ($p < 0.001$) for students who reported as female compared to as male, no significant difference in the discrimination parameter, and equivalent SIDS and UIDS. This uniform DIF for these items suggested that students who reported as female were less likely to answer these items correctly even after controlling for knowledge levels, as shown in Figure 4.1. *U1, Q3* has a medium e.s. that says that on average, students who reported as female got this item wrong 5.2% more than those who reported as male. *U3, Q6* had a large e.s. that we interpreted to say that on average, students

who reported as female got this item wrong 10.3% more often.

Although the effect sizes were negligible (SIDS < *0.05*), two items exhibited DIF slightly disadvantaging students who reported as *male*. *U2, Q3* and *U2, Q4* had significantly lower ($p < 0.05$) difficulty parameters for students who reported as female compared to as male. So on average, students who reported as female were more likely to get these items correct.

Taken together, we can say that matching items related to app development and commenting code most disadvantaged students who reported as female, with multiple choice items on web development and good coding practices providing a statistically significant but negligible advantage for them.

When comparing reported **female and non-binary students** (white rows in Table 4.3), we found that five items exhibit uniform DIF that disadvantaged students who reported as female. Three items in Unit 1 had uniform DIF disadvantaging students who reported as female: *U1, Q1* (medium e.s.), *U1, Q3* (large e.s.), and *U1, Q4* (large e.s.). *U1, Q3* actually disadvantaged students who reported as female when compared to both male and non-binary. Items *U2, Q6* (medium e.s.) and *U3, Q4* (large e.s.) also disadvantaged students who reported as female compared to non-binary students.

When comparing reported **non-binary and male students** (gray rows in Table 4.3), we found that one item disadvantaged students who reported as male (*U1, Q4*, medium e.s.).

Table 4.2: Item information (type, description) and statistics. Difficulty, discrimination (r_{pbis}), reliability (change in α from 0.732), and proportion of students not responding (NR) are reported. ♦: distractor selected more frequently than correct answer. ■: interdependency between items.

	type	description	difficulty	r_{pbis}	$\Delta\alpha$	NR
U1, Q1	select 2	select 2 best ways to define computer	0.70	0.32	-0.004	0.30
U1, Q2	match	match steps to painting mural to problem-solving process	0.35	0.35	-0.01	0.30
U1, Q3	match	match weather/outfit app actions w/ computer system parts	0.52	0.42	-0.02	0.31
U1, Q4	select 1	identify which of two problems with school is better defined	0.41	0.36	-0.01	0.31
U2, Q1	select 2	select 2 tasks HTML is "most important language for"	0.45	0.37	-0.003	0.49
U2, Q2	select 1	identify problems with using single language for web dev.	0.46	0.34	-0.0001	0.50
U2, Q3	select 1	when to use classes for website	0.29	0.40	-0.02	0.51
U2, Q4♦	select 1	identify causes for styling to not appear on a specific webpage	0.32	0.38	-0.01	0.51
U2, Q5■♦	select 1	given HTML code and web page view, select CSS to produce	0.31	0.43	-0.02	0.51
U2, Q6■	select 2	given same HTML & view, select 2 ways to make text larger	0.60	0.45	-0.02	0.51
U2, Q7	select 1	select true statement about copyright	0.59	0.44	-0.02	0.51
U3, Q1	select 2	select 2 options that improve code readibility	0.66	0.40	-0.01	0.86
U3, Q2	select 2	select 2 uses for functions	0.53	0.41	-0.02	0.86
U3, Q3	select 1	given code (in blocks and text), determine stored value in var.	0.38	0.35	-0.01	0.86
U3, Q4	select 1	determine which is not best to decide before beginning to code	0.75	0.37	-0.01	0.86
U3, Q5♦	select 1	identify potential causes of problem w/ "platform jumper game"	0.27	0.27	-0.01	0.86
U3, Q6	match	given 22 lines blocks code, match comments to location in code	0.36	0.42	-0.02	0.86

Table 4.3: Likelihood Ratio Test (LRT) DIF results for pairwise comparisons between reported gender (non-binary, female, male). Significant difference in difficulty parameter denoted with * for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$. Effect sizes for uniform DIF denoted with • for medium (signed in-sample differences/SIDS ≥ 0.05), •• for large (SIDS ≥ 0.10). ε denotes p-value that is < 0.001. P-values adjusted with Benjamini-Hochberg procedure.

	uniform DIF					non-uniform DIF					effect sizes	
	difficulty			sig. test		discrimination			sig. test			
	non-b.	female	male	χ^2	p	non-b.	female	male	χ^2	p	SIDS	UIDS
U1, Q1*		-1.146	-1.141	28.727	ε		0.654	0.779	8.737	0.075	0.025	0.026
U1, Q1**•	-1.473	-1.146		9.650	0.007	0.796	0.654		0.502	0.935	0.078	0.078
U1, Q2***		0.913	0.852	34.370	ε		0.955	0.836	6.888	0.110	0.029	0.029
U1, Q3***•		0.164	-0.048	86.984	ε		1.133	1.237	3.335	0.534	0.052	0.052
U1, Q3*** ••	-0.382	0.164		26.566	ε	1.544	1.133		1.601	0.749	0.152	0.152
U1, Q3** ••	-0.382		-0.048	12.533	0.002	1.540		1.237	0.866	0.935	0.100	0.100
U1, Q4***		0.927	0.529	72.547	ε		0.655	0.758	6.786	0.110	0.047	0.047
U1, Q4*** ••	0.143	0.928		17.549	ε	0.676	0.655		0.094	0.957	0.117	0.117
U1, Q4*•	0.143		0.529	6.552	0.033	0.678		0.758	0.188	0.957	0.069	0.069
U2, Q1**		0.526	0.478	12.641	0.004		0.807	0.669	10.953	0.032	0.022	0.026
U2, Q3***		1.076	1.101	26.257	ε		1.342	1.110	16.019	0.003	0.025	0.029
U2, Q4*		0.979	1.074	9.725	0.013		1.090	0.887	16.738	0.003	0.014	0.024
U2, Q6**		-0.213	-0.254	14.499	0.002		1.221	1.548	24.686	0.001	0.020	0.036
U2, Q6*•	-0.405	-0.213		5.985	0.041	1.836	1.221		3.247	0.365	0.076	0.086
U3, Q4* ••	-1.462	-0.855		6.442	0.033	1.617	1.422		0.017	0.971	0.141	0.141
U3, Q4** ••	-1.463		-0.746	9.768	0.007	1.605		1.382	0.002	1.000	0.169	0.169
U3, Q6*** ••		1.116	0.634	46.081	ε		1.455	1.561	0.603	0.935	0.103	0.103

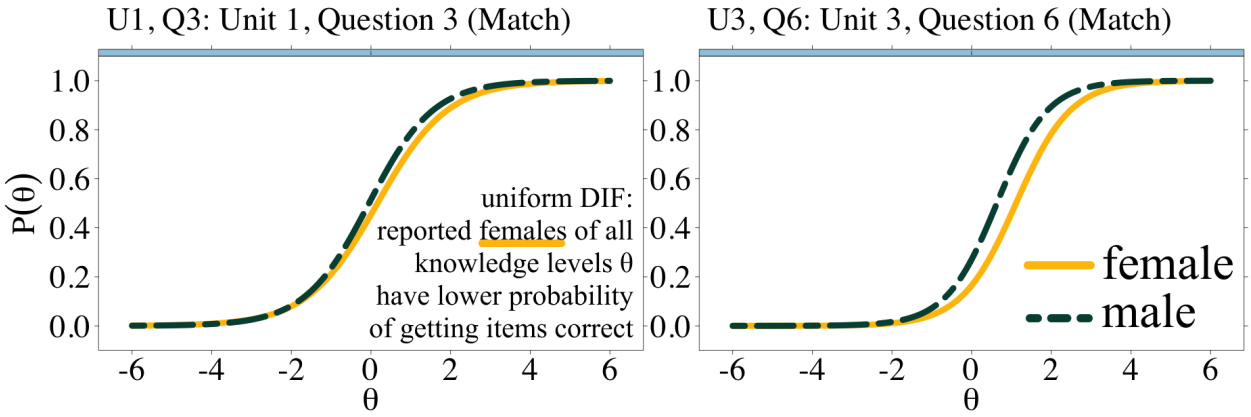


Figure 4.1: Traces for items that exhibited (uniform) gender-based DIF of medium or large effect. (Items w/ • in blue rows of Table 4.3)

DIF for race: uniform DIF disadvantages AHNP

When comparing **AHNP (African/Black, Hispanic/Latinx, Native American/Alaskan Native, Pacific Islander) students to WA (white, Asian) students**, we found that 13 of 17 items exhibited uniform DIF with medium or large effects of disadvantaging AHNP students, as shown in Table 4.4. All four items in unit 1 (*U1, Q1-4*), the later three items in unit 2 (*U2, Q5-7*), and all six items in unit 3 (*U3, Q1-6*) had significantly greater difficulty parameters for AHNP students ($p < 0.001$), suggesting these items disadvantaged AHNP students. While some items had significantly different discrimination parameters (*U1, Q4; U2, Q4; U2, Q6* for $p < 0.01$ and *U1, Q2; U1, Q3; U3, Q3* for $p < 0.05$), there was no difference in SIDS and UIDS (or negligible difference for *U2, Q4*). So, we interpreted all 13 race-based DIF items to exhibit *uniform* DIF. So on average, AHNP students had a 5.9% (for *U1, Q2*) to 18.6% (for *U1, Q3*) lesser chance of getting items correct compared to WA students. Figure 4.2 shows the trace plots for items that exhibited uniform DIF with large e.s..

Table 4.4: Likelihood Ratio Test (LRT) to detect DIF with AHNP students (African/Black, Hispanic/Latinx, Native American/Alaskan Native, and Pacific Islander) as the focal group and WA students (white, Asian) as the reference group. Significant difference in difficulty parameter denoted with * for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$. Effect sizes for uniform DIF denoted with • for medium (signed in-sample differences/SIDS ≥ 0.05), •• for large (SIDS ≥ 0.10). ε denotes p-value that is < 0.001. P-values adjusted with Benjamini-Hochberg procedure.

	uniform DIF				non-uniform DIF				effect sizes	
	difficulty		sig. test		discrimination		sig. test			
	AHNP	WA	χ^2	p	AHNP	WA	χ^2	p	SIDS	UIDS
U1, Q1***•	−1.071	−1.718	134.215	ε	0.696	0.684	0.045	0.869	0.084	0.084
U1, Q2***•	1.021	0.580	52.298	ε	0.819	0.980	6.931	0.024	0.059	0.059
U1, Q3***••	0.276	−0.457	583.094	ε	1.077	1.307	8.746	0.013	0.186	0.186
U1, Q4***••	1.087	0.176	222.595	ε	0.574	0.770	13.729	0.001	0.114	0.114
U2, Q1***	0.473	0.291	21.186	ε	0.797	0.685	4.100	0.091	0.039	0.040
U2, Q2***	0.414	0.227	17.339	ε	0.745	0.656	2.677	0.192	0.036	0.036
U2, Q3	1.122	0.902	2.763	0.096	1.124	1.295	5.061	0.059	0.023	0.025
U2, Q4***	0.988	1.052	15.989	ε	1.123	0.871	14.471	0.001	0.026	0.034
U2, Q5***•	1.054	0.761	37.284	ε	1.274	1.316	0.280	0.725	0.059	0.059
U2, Q6***••	−0.080	−0.589	289.125	ε	1.212	1.550	14.881	0.001	0.151	0.151
U2, Q7***••	−0.021	−0.603	259.283	ε	1.163	1.234	0.847	0.480	0.147	0.147
U3, Q1***••	−0.435	−0.976	61.144	ε	0.952	1.064	0.813	0.480	0.122	0.122
U3, Q2***••	0.348	−0.303	93.439	ε	1.137	1.195	0.196	0.746	0.161	0.161
U3, Q3***••	1.254	0.411	39.539	ε	0.761	1.079	7.471	0.021	0.113	0.113
U3, Q4***••	−0.788	−1.269	64.150	ε	1.152	1.357	1.791	0.279	0.123	0.123
U3, Q5***•	2.332	1.328	17.693	ε	0.601	0.776	2.503	0.193	0.070	0.070
U3, Q6***••	1.011	0.445	55.141	ε	1.300	1.325	0.027	0.869	0.130	0.130

Items that exhibited uniform race-based DIF spanned the first three units in the CSD curriculum. Unit 1 items focused on basics of a computer and problem solving, asking students to do things such as select the two best ways to define a computer (*U1, Q1*) and match steps to painting a mural to a pre-defined problem-solving process (*U1, Q2*). The first four items in unit 2 that exhibited negligible amounts of DIF were all multiple-choice items that asked conceptual questions about creating a website. U2, Q4 actually exhibited an uniform DIF (of negligible e.s.) in the opposite direction, where on average AHNP students scored 2.6% *better* than WA students. This question asked students to identify potential causes for styling to not appear on a specific webpage. The remaining items in unit 2 asked questions about a code snippet (*U2, Q5-6*, which are interdependent on the same code) and about copyright. Items in unit 3 assessed students on constructs and patterns to create interactive games, asking students about things including the benefits of using functions (*U3, Q2*) and what is NOT best to decide before beginning to write code (*U3, Q4*).

A majority of items exhibiting DIF could suggest that the LRT DIF method was failing to match students of equivalent knowledge level using latent variable modeling. To see whether the LRT DIF results were reasonable, we used logistic regression (LR) DIF, which matches students by total score. A DIF item detected by multiple methods is more likely to truly be a DIF item[76], so similar results from the LR DIF analysis would suggest that the LRT results were accurate. We used LR DIF with purification and a Benjamini-Hochberg correction [21] to check for uniform DIF. LRT DIF found 13 DIF items with a medium or large effects; 12 of those were also detected with LR DIF ($p < 0.001$, except U1, Q1 which was $p < 0.01$); U2, Q5, was only trending towards significance ($p = 0.08$). Because 12 of 13 items that LRT DIF found to exhibit DIF with non-negligible effect also exhibited DIF for LR DIF, we have stronger evidence to suggest that most items exhibited uniform DIF that disadvantaged AHNP students.

Taken together, most of the assessment exhibited uniform DIF that disadvantaged AHNP students, but items relating to website design (Unit 2) exhibited the least disadvantage (and in one instance, a negligible advantage). Figure 4.3 shows the substantial effects of DIF on students' scores across all 17 items, comparing between gender and racial groups and the average number correct for three knowledge levels.

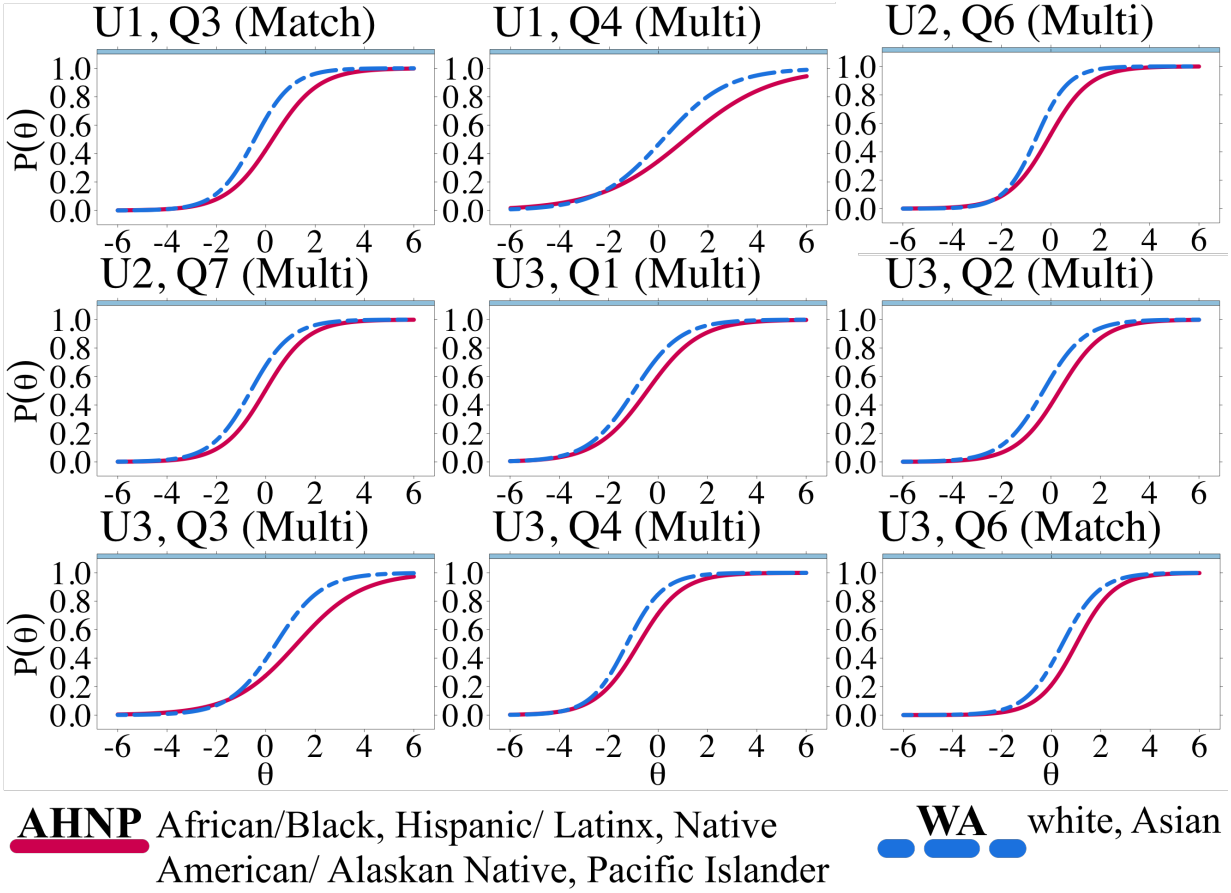


Figure 4.2: Traces for items that exhibited (uniform) race-based DIF with large effect size. (Items with ●● in Table 4.4)

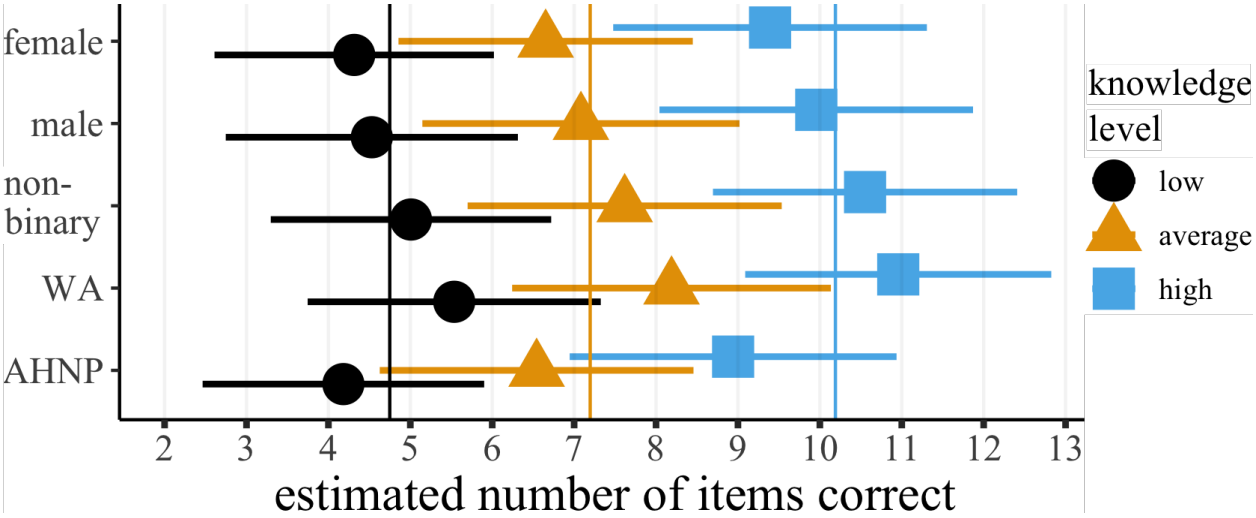


Figure 4.3: Expected number of items a student would get correct (out of 17) by gender and racial groups for three different knowledge levels. Knowledge levels were calculated with an IRT model assuming no DIF, where *average* is the median knowledge level in our sample ($\theta = -0.07$), *low* is a standard deviation (1σ) below ($\theta = -0.81$), and *high* is 1σ above ($\theta = 0.65$). Vertical bars indicate simulated mean number correct with no DIF. Shapes indicate mean number of items correct for each group from 1000 simulations, with horizontal error bars showing 1σ.

4.5 Qualitative Results: Designers' Interpretation

To understand how domain experts interpreted and used DIF results, we conducted a workshop with seven curriculum designers at Code.org as well as individual follow-up interviews with three who expressed interest. In this section, we describe the workshop, follow-up surveys, and conversations; provide background on them (demographics, perspectives on assessments and equity); and elaborate on themes they identified when interpreting DIF data. All this was in an effort to understand a new use for DIF: improving equity in learning by informing domain experts of potential issues.

4.5.1 Workshop with curriculum designers to interpret DIF

We conducted a remote workshop with curriculum designers to understand how they interpreted and considered using results from our DIF analyses. Seven curriculum designers participated in

a remote, recorded workshop in place of a regularly scheduled team meeting. The workshop was organized with a stated goal of thinking about how assessment design relates to Code.org curricula more broadly. It began with anonymous visible responses (via Poll Everywhere) to the following questions: *How can instructors and/or students benefit from using assessments in Code.org?* and *For Code.org,what are challenges to designs an equitable learning experience?* The goal of having participants respond to these questions and discuss them was to prompt them to think more about assessment and equity. After that, we gave a 5-10 minute presentation introducing the study, item response theory, and DIF.

To understand curriculum designers' interpretations of DIF, we randomly split them into two separate groups (3-4 people per group) where each group was given a collaborative document with information about items that exhibited DIF (gender-based DIF for one group, race-based DIF for the other), as well as links to the curriculum and assessment items with solutions. To describe each item that exhibited DIF, we included item trace plots (like Fig. 4.1 and 4.2) as well as brief description following the following format: *For this question, {X}% of boy students (dotted green line) would get it correct and {Y}% of girl students (solid yellow line). This is an {intermediate/large} effect size*. We intended for this information to be consistent with something that could be automatically generated by an analysis package. Groups were then given 20 minutes to discuss and take notes on the following questions: 1) How do you interpret this data? 2) What actions might you consider taking? 3) What additional information are you missing? How could that new information help you? After this, everybody reconvened and members from each group took turns sharing their findings. After the workshop, we sent each participating curriculum designer a post-survey asking them about the benefits of reviewing DIF, difficulties or challenges of reviewing DIF, potential uses of data that identifies unfairness, as well as demographic information. Four participants filled out that survey. Three others also initiated follow-up discussions.

The data we analyzed include video and audio recordings of the workshop, responses to questions (Poll Everywhere, post-survey), the collaborative documents that each group shared when reviewing DIF data, and message transcripts from follow-up discussions. All quotes in this section came from Code.org curriculum designers who participated in the study. To preserve anonymity

(especially amongst curriculum designers), we do not provide further attribution to any quotes.

4.5.2 Curriculum designers: tests are formative, equity is hard

The curriculum designers we worked with had domain expertise in developing and managing computer science curriculum, though only some worked on CSD specifically. The follow-up survey found that the designers reported genders including men, women, and non-binary, and ethnicities including Asian, Black/African, Pacific Islander, and white. Multiple had Master's degrees in education, with one having previous experience in psychometrics. So we can say that this diverse group had domain expertise relating to curriculum design for computer science courses for elementary, middle, and high school students. All seven designers saw assessments for formative purposes (to "*pin point areas where students need extra help*," "*inform later instruction for a class or individuals*.").

When asked about challenges to designing an equitable learning experience, curriculum designers noted challenges relating to scaling online curricula to a diverse global audience. Three designers noted the challenges of using an online platform to provide curriculum such as "*embedded limitations*" and "*varying fidelity of implementation*." Three also noted the challenges related to "*designing activities that can benefit students even with such a wide range of school implementations or teacher mindsets*." Two noted the need to design curriculum that supported teachers: "*designing curriculum that works well with our [professional learning] program but also serves those who are using it without [professional learning].*" One curriculum designer noted the role of teachers "*to create equitable spaces for their students based on the community they serve*." And finally, one also called for "*more diversity in people behind the curriculum and [professional development/professional learning]*." Curriculum designers tended to frame designing equitable learning experiences as a holistic endeavor that involved multiple stakeholders (e.g. teachers, students) and multiple efforts (e.g. curriculum design, professional learning).

4.5.3 Curriculum designers' interpretations of DIF

At Code.org, seven people made up the curriculum team that designed and maintained online instructional materials for the largest in-person implementations of CS curricula in the world. They often worked with professional development specialists, product managers, software developers, and others to develop and improve three curricula targeting different age groups. But how equitably a curriculum serves members of a diverse community of teachers and students is a constant uncertainty for organizations like Code.org that produce online instructional materials used by over a million students annually whom they will never meet. To understand how curriculum designers interpreted DIF results for gender and race, we reported themes that curriculum designers identified after reviewing data on items that exhibited DIF, as well as statements they wrote or said during the workshop or in a follow-up conversation.

Considering DIF relative to item features

When looking at gender-based DIF, curriculum designers considered item design and knowledge the items assessed. For *U1, Q3* and *U3, Q6*, designers noted how "*Female students are performing lower on matching [type] questions that are both computer science concepts and code tracing*." But they also noted a difference in the magnitude of DIF: "*Comparing both of these graphs, female students are performing lower on code tracing than vocab matching*." So curriculum designers noted similarities in item type and differences in the knowledge that items assessed as well as magnitude of DIF.

From there, curriculum designers considered how performance on other items could help them. Curriculum designers only saw DIF items, but wanted to see data from all items. They considered questions related to other items of the same type (matching): "*How did other matching questions throughout the course do?*" They also sought to compare DIF results to items of another type: "*What about in comparison to single-answer multiple choice questions. Are students doing better or worse on those? By gender?*" So curriculum designers sought to compare DIF results of items of similar and different forms.

Alignment between assessment and curriculum

Designers considered how the CSD curriculum prepared students for knowledge that items assessed. For example, when reviewing gender DIF, they noted how the item assessed commenting code but CSD did not emphasize this: "*comments are not very well emphasized in CS Discoveries at all. So this may be the very first that students are seeing this idea of putting a comment to a block of code*." This item raised the broader question of "*how are these assessment questions showing up in the curriculum leading up to this point?*"

Given this insight, they discussed conducting an audit to check alignment between item format, curriculum, and learning objectives. One curriculum designer stated that "*an action you might consider is auditing how frequently these types of assessment prompts appear earlier in the course. Are [students] actually prepared for this?*" So it is not just preparing a student for with the knowledge necessary to answer an item (e.g. how to read code to identify higher level goals), but also ensuring students are familiar with the format of the item itself (e.g. placing comments in code). Another curriculum designer did note that "*it might be kind of hard to map some of the questions we were looking at to lessons or objectives covered in the curriculum*." Nevertheless, curriculum designers considered "*yearly audits of assessment questions as part of our summer updates*." This ultimately led curriculum designers to frame DIF results as informative to an equity-focused curriculum improvement process: "*I could see us using [DIF] as one of the data points we use to evaluate our curriculum as a whole in terms of how we are serving the populations of students traditionally underrepresented in computer science... using this data as a starting place for a conversation around where to focus our efforts first and foremost, on improvements to the lessons in the curriculum or to the assessments themselves*."

Finally, a designer considered how social context influenced student responses and how there may have been a lack of alignment between items and curricula: "*Some of the questions I could imagine if they're given independently of the unit, that some students could answer based on experiences they've had before coming into the classroom. Because of the fact that they're not that tightly aligned with things in the curriculum, probably, there would be cases where favoring would*

be just as present after teaching this course as before."

Differing goals for assessments

While designers generally saw assessments as useful for formative evaluation, they also considered how different goals impacted interpretations of DIF data. Assessments were optional and their uses were left ambiguous in the curriculum guide, leaving one designer to question the use of the data: "*CSD assessments are optional.. so I wonder about the quality of the data being collected in the first place?*"

Curriculum designers also questioned the authenticity of an item because they felt its challenge was not consistent with a "*authentic real world type questions.*" *U3, Q3* asked students to trace variable values as they updated. While this knowledge of changing program state aligned with learning objectives, curriculum designers thought the code snippet "*was puzzly and tricky, but nothing you would actually write as a program... if you're more of like 'I'm taking this course because I want to make meaningful things', then this [question] does not fall into that category.*" Curriculum designers identified a tension between designing an item with a goal of measuring specific knowledge precisely compared to reflecting more authentic tasks: "*the ability to answer these questions [doesn't] tell us a lot about how well students could accomplish these tasks/demonstrate these skills on a real project.*"

Challenges to reviewing DIF data

Curriculum designers identified challenges to using DIF related to interpreting data on uncertainty, as well as limitations to understanding causes of bias with DIF.

Item trace plots deviated from one designer's expectations, so they relied on their colleagues to understand the data: " *I'm used to seeing % look something like 16% of male students are proficient rather than 'only 16% of students who reported as male would have a >50% chance of getting that question right'... I leaned on my colleagues to help fill in some blanks.*"

Curriculum designers also noted how evidence of DIF did not provide them information on the

causes of bias. One curriculum designer felt that investigating DIF did not provide the most relevant information for addressing bias: "*I'd like to see more info on how the curriculum is actually being used in a holistic sense. Who is teaching, where do they teach, what environment do the kids go home to, etc.*" They went on to suggest that analyzing DIF may be detracting from a more challenging conversation on disparities in STEM education by race and gender: "*I didn't feel we really discussed *why* there was a disparity between bipoc and white/asian students... There is already a ton of literature on STEM assessments, race, and gender, so I'd start by reviewing that stuff before making any assumptions [about biases from item design].*"

4.6 Discussion: How DIF Informs Domain Experts

In this study, we analyzed gender and race bias in one of the largest online curricula for CS education, and then conducted a feasibility study to explore how domain experts could use DIF results. Our analysis found that five items disadvantaged students who reported as female compared to male and non-binary students with non-negligible effect sizes, and 13 items disadvantaged AHNP students compared to WA students with non-negligible effect sizes. These items (denoted with • in Tables 4.3 and 4.4) should be reviewed for revision or potential removal. Our workshop with curriculum designers found that they interpreted DIF relative to student identities, curriculum, and assessment goals, identifying critical nuances for making valid interpretations and uses of assessment scores.

There are multiple ways to interpret our findings on how curriculum designers interpreted DIF results. One interpretation is that designers lacked the psychometric expertise to interpret DIF results. Indeed, designers noted some trouble interpreting DIF results (e.g. trace plots) and having a psychometric expert available could have been beneficial. But practically, organizations creating curricula for online or in-person use often do not have access to psychometric expertise. And even without a psychometric expert, curriculum designers were able to consider DIF results relative to item design, student identity, curriculum, and assessment goals. And these interpretations are consistent with those of an external psychometric expert.

And furthermore, curriculum designers are ideal interpreters of DIF results because they can

consider them with respect to the intended uses of the test scores, a crucial consideration to test validity. In contrast to the high-stakes summative assessments that psychometric experts typically analyze, designers framed these assessment items as formative and low-stakes, intended to provide feedback to support students' learning. So they considered the role of formative assessment in equitable learning experiences when assessment items exhibit potential bias. And while some items perpetuate bias that exists in cultural contexts, some items may introduce or further exacerbate bias, causing potential harm to test-takers of minoritized groups. Future work can explore how biased formative assessment items affect test-takers from different genders, ethnicities, and other identities, perhaps considering stereotype threat [272, 293, 256] and test-taker self-efficacy [266, 139, 141].

Another interpretation is that domain experts needed quantitative analysis with more contextual variables to understand what causes biases and inequities. DIF indicates the potential existence of bias, but it does not provide insight into the cause of the bias (e.g. test design, pedagogical practice). The latest research on DIF emphasizes the role of the *testing situation* as well as characteristics of an item [317]. And while providing more features to a model may enable more nuanced insights, this also comes with a greater demand for data which may further minoritize minority groups. We had to group AHNP students and WA students together to ensure IRT model convergence, while also recognizing that these groupings were reductionist (students of different ethnicities often have different lived experiences) and potentially harmful [137]. Quantitative complexity comes at the cost of reduction/ aggregation.

Furthermore, reductionist quantitative labels do not reflect the experiences of groups, so designer's domain expertise may help prevent misrepresentation. Our analysis identified that five items favored students who reported as non-binary. But coming out as non-binary in school is a constant and challenging process [207], and a majority of non-binary students may have been assigned female at birth [50]. So while quantitative analysis suggests that items biased in favor of non-binary students, more contextual understanding nuances this interpretation. That nuance also highlights that quantitative analysis can be a *starting point* which domain experts can use to augment their existing expertise. Future work can explore how to present quantitative data on bias

to situate domain experts' interpretations relative to their existing beliefs and expertise.

Yet another interpretation is that DIF analysis is not beneficial to domain experts' understanding of equity because it does not get at the causes of inequities. Curriculum designers viewed DIF as a confounded indicator of potential bias because it is unclear if the bias came from the item, the curriculum, the classroom context, or broader sociocultural context. But the results of our feasibility study showed that that providing domain experts even limited information can help them focus their investigation. For example, curriculum designers in our study focused on the items that did *not* exhibit significant race-based DIF, wondering if perhaps these items on web development suggested a more equitable entryway into the curriculum for AHNP students. So DIF can provide precise and measurable metrics reflecting potential bias that can help domain experts develop their existing knowledge or bring about new ideas related to equitable curriculum design.

A final interpretation is that domain experts must contextualize DIF findings to ensure results do not invite harmful misinterpretations. While data on test validity and fairness is often used by psychometrics experts for the purpose on improving test design, it may also benefit domain experts like curriculum designers as they review and revise their curriculum materials. In our study, we identified ways that curriculum designers considered DIF results relative to the domain expertise they had on curriculum design as well as stakeholder (e.g. teacher, student) needs and goals, findings that data alone could not show. This suggested that domain expertise can enable more nuanced understanding than DIF alone, which typically relies on reductionist labels and dichotomies to compare a dominant group to a minoritized group. And quantitative measures of bias such as DIF can augment domain experts' understanding of how to improve equity in learning experiences by identifying who is affected and where to focus improvement efforts.

Iterating towards more equitable learning experiences requires measuring factors we cannot easily intuit, and using domain expertise to contextualize these findings with understanding we cannot easily measure. So interactions with quantitative data such as DIF can enable domain experts to recognize what is happening to better inform them as they use contextual knowledge to identify how they can address inequities.

4.7 Conclusion: Prior knowledge and cultural competence informed, power relationships scoped

Connecting the findings of this design exploration back to the framework I defined in Fig. 2.2, this study suggested that curriculum designers were able to engage their *relevant prior knowledge* and *cultural competence* to contextualze data on potential test bias (DIF). The collaborative workshop provided designers with an opportunity to consider data in relation to prior knowledge they had with the design and use of the test and curriculum. And some curriculum designers with more cultural competence from their prior education and lived experiences were able to draw potential connections between potential test bias and systemic challenges that extend beyond curriculum design. But *perceptions of power relationships* scoped the conversation to what actions were feasible given their responsibilities and limited time and resources.

An alternative design that better engaged participants' cultural competence and perceptions of power relationships could deliberately shift the focus of the workshop. Curriculum designers attributed causes of DIF to varying scopes, from instructional design to differences in instructional use to systemic biases in educational systems. But the conversations largely focused on instructional design, because that is what they had the most capacity to impact. Scaffolding the conversation to deliberately consider causes of bias at different scopes by relaxing immediate constraints of what was deemed actionable could have yielded richer interpretations of the data on potential test bias.

Chapter 5

STUDENTAMP: CONTEXTUALIZING STUDENT FEEDBACK TO HELP TEACHING TEAMS IDENTIFY INEQUITIES

In this chapter, I describe a study that investigated how to support the equity-oriented goal of identifying inequities in large, remote courses. I built *StudentAmp*, a student feedback tool that collected student-reported data on demographics, challenges they faced, and perceptions of their peers' challenges. It then showed teaching teams challenges that students reported contextualized with demographic information and a score representing how disruptive students found given challenges.

I conducted an evaluation with five large, remote courses during the COVID-19 pandemic to understand how teaching teams used data on contextualized student feedback. From their discussions and collaborative engagement with the data, I discovered how cultural competence from prior training (e.g. coursework in public health, anti-racist seminars) enabled some to engage more readily with topics of identity and minoritization. I also discovered a tension between relying on prior knowledge of personal experiences and recognizing that lived experiences of students varied from their own. And finally, I uncovered how perceptions of power relationships left even professors feeling somewhat powerless to address systemic inequities that extend beyond their courses. This work informs design of contextualized student feedback that equitably considers the needs of students at scale while also ensuring the privacy and well-being of the minoritized groups.

The work in this chapter was conducted in collaboration with Alannah Oleson, Jayne Everson, and Dr. Amy J. Ko and was accepted for publication at 2022 PACMHCI and will appear at CSCW

2022 [311]. [1] [2]

5.1 Introduction: Contextualizing Feedback to Understand Inequities

Teaching equitably is important in computer and information sciences (CIS), where there are many inequities in formal CIS courses in university and higher education contexts. Formal higher education in CIS is a primary pathway for participation in the computing community, yet CIS courses face persistent diversity, equity, and inclusion issues [296, 143]. In part due to the growing demand for computing skills in the workforce, CIS enrollment numbers have surged recently, straining the capacity of instructors to scale teaching [208]. Despite their popularity, CIS courses continue to face challenges retaining and supporting diverse students in both high school [94] and college [319, 278, 296]. This results in the loss of diverse potential contributors to computing fields [208] and raises social justice issues around who can access and engage with computing communities [163, 257, 258].

One way to to try to teach more equitably is by sourcing feedback from students and responding to it [59]. Feedback is especially critical to an equitable learning environment because students from minoritized groups face unique challenges that, if left unaddressed, can pose serious impediments to science and technology learning [69]. Instructors in higher education have used student feedback as a way to monitor and improve teaching equality, especially in distance and remote learning environments [146]. However, it is not enough to simply be made aware that these challenges exist: To help turn student feedback into action, instructors need context regarding their students' lived experiences to understand how challenges affect different students [60, 188].

Student feedback tools in large (100-500+ student) courses must be scalable. Prior CSCW works have examined massive open online courses (MOOCs), looking at student motivations and

---

[1] This study was pre-approved with exempt status by the UW Institutional Review Board (IRB) as STUDY00008871. This material is based upon work supported by the National Science Foundation under Grant No. 12566082, 1762114, 1539179, 1703304, 1836813, 2031265, 2100296, 2122950, 2137834, 2137312, a Google Cloud Research Grant, and unrestricted gifts from Microsoft, Adobe, and Google. Supplemental material can be found at `https://github.com/codeandcognition/archive-2022cscw-xie`.

[2] I will use "we" instead of "I" in this chapter to acknowledge the shared contributions of all authors.

retention [312], the impact of a reputation system on the student experience using forums [53], and how matching students across locations helped students to earn higher grades [167]. But in contrast to MOOCs, remote courses typically have more synchronous interactions and feedback mechanisms between instructors and students. Scalable feedback in synchronous courses involves ensuring convenience for students to share feedback [157, 110], and convenience for instructors and teaching assistants (TAs) to collect, analyze, and discuss the feedback [146, 241]. In addition to this requirement, we argue that student feedback for equity-oriented goals must also provide the context to help instructors and TAs consider feedback within the context of students' lived experiences while also ensuring students' privacy.

Context is important to support equity-oriented goals, but existing student feedback mechanisms lack the context to connect feedback to lived experiences. At the scale of hundreds of students to a single *teaching team* consisting typically of one instructor and a few TAs, a teaching team typically cannot respond to all feedback. As a result, commonly used electronic response systems, such as anonymous online surveys sent to students during the term, often lack context about lived experience [48]. This loss of context results in less actionable feedback as it obscures perspectives of minoritized groups as they become lost amongst the majority of perspectives which are typically from students of dominant groups [188].

Contextualizing feedback can come in tension with protecting student privacy, another critical aspect for equity-oriented feedback. Students from minoritized groups are often most at risk when their information is exposed without their informed consent. Feedback methods that are interpersonal and conversational, such as conversations between a student and an instructor after a lecture or with a teaching assistant (TA) during office hours, are common within large computing courses. While these methods provide context by revealing students' identity, they also privilege students who are more willing and able to speak up, such as white and Asian men with prior programming experience [110]. Interpersonal techniques can be especially problematic for students of minoritized groups due to a lack of anonymity potentially introducing stereotype threat [272, 300] and social-desirability biases [123, 108]. Student feedback for equity-oriented goals must also ensure students' privacy when they share feedback.

To explore the design of a student feedback tool that supports equity-oriented goals by 1) being scalable, 2) providing context, and 3) ensuring student privacy, we designed *StudentAmp*. StudentAmp contextualizes student feedback on challenges in their life with demographics and peer perspectives from other students. Using StudentAmp, students self-report challenges that interfered with their learning as well as demographic information (e.g. gender, ethnicity, prior programming knowledge). Students then consider random pairs of challenges their classmates/peers shared and determine which challenge they would consider more disruptive. StudentAmp aggregates students' *meta-feedback* responses to produce a ranking of perceived challenge disruptiveness. StudentAmp then uses this data to produce a report for an instructor detailing challenges students reported, contextualizing challenges with demographic information as well as ranking them according to student perceptions of disruptiveness.

To evaluate the effects of using StudentAmp to collect and report student feedback contextualized by demographic information and perceptions from classmates, we conducted a formative study with teaching teams and students of five large remote computing courses (163 - 628 students/course). We considered the following research questions:

1. What do students share about challenges interfering with their learning?

2. How do students perceive the values and risks of sharing information on challenges they face, contextualized with demographics and peer-perceptions?

3. How do teaching teams of large computing courses use different types of information to contextualize students challenges for equity-oriented interpretations?

We found that students considered the privacy of themselves and others when sharing feedback on challenges that were often about their lives beyond computing courses. Seeing anonymous peers' challenges also helped students empathize and develop a sense of belonging with peers. Instructors used demographic data to connect challenges to student experiences by situating challenges in lived experiences that may differ from dominant norms, while finding data on peer perspectives questionable and unreliable. We interpreted these findings as design trade-offs between

contextualizing feedback with demographic data to inform stakeholders of inequities at scale and ensuring the privacy and well-being of students.

This paper makes the following three contributions:

1. A large scale thematic analysis of 810 challenges that 604 students faced while learning computing in large remote courses during the COVID-19 pandemic;

2. An artifact (StudentAmp) that is a design exploration into how contextualizing student feedback may support more equitable learning experiences in large, remote courses; and

3. A rich qualitative investigation of

   (a) Students' experiences sharing feedback through StudentAmp and viewing the feedback of their peers, and

   (b) Teaching teams' experiences using StudentAmp to better understand inequities in their courses by interpreting contextualized student feedback.

5.2 Background: Equity, Perspective Taking, and Theory of Action

In this section, we provide our framing of equity within the context of higher education computing courses. Then, we describe how student demographics and perspectives can contextualize student feedback data by providing opportunities for perspective taking and empathizing. Finally, we describe a Theory of Action that scaffolds interpretations and uses of student feedback data.

5.2.1 Equity Involves Understanding Experiences of Minoritized Groups

We framed equitable learning as ensuring students from diverse backgrounds can successfully access and engage with a learning experience to realize their dignity and potential. Within the context of computing education, understanding students' diverse backgrounds involves considering intersectionality [70], or how different aspects of students' identities intersect and interact. These aspects of identities include students' ethnicities, genders, disabilities (physical, mental, social),

preparatory privilege, current situation outside of the course (e.g. familial or financial responsibilities), and past educational experiences (whether they are a transfer student, first generation). Learners are complex individuals beyond a single demographic label.

When considering equity, we must consider not only students, but also the structures of society that students exist in. We establish this framing of equity upon upon Structuration Theory, which defines a recursive process where society and its structures shape the activity of individuals and individuals shape and condition the structures of society [164, 54]. Understanding equity involves not just individuals involved, but also the context of the economic, social, cultural, and political conditions of the time and place [103, 206]. Equity has a social justice goal where corrective measures must adjust for aggregate harm from social inequalities [231]. As a result, understanding equity involves considering how learners are situated within complex environments that they also shape.

Within computing courses, improving equity would require not just improving access to computing education, but also supporting successful participation and achievement by diverse students learning computing [172]. Structural and systemic inequities embedded in and around computing courses can manifest as barriers to participation (e.g. unconscious bias of instructors excluding students of color from successful participation [244]), affect students' sense of belonging and identity (e.g. instructional materials promoting gender bias [197]), and exacerbate existing disparities in privilege (e.g. students cannot synchronously engage with instructors and other classmates because of timezone differences, work commitments, or familial responsibilities) [172]. Inequities arise when structures and norms fail to include or serve students of minoritized groups. Addressing inequities often involves interventions that support the needs of specific groups, such as a one hour social-belonging intervention to support the long term career, mental health, and community building for Black students [31, 294].

For this paper, we referred to *minoritized* as a descriptor of identity groups that are typically not dominant within computing communities in the United States (US). Dominant groups are positively privileged [299], unstigmatized [247], and generally favored by the institutions of society [186], particularly within social, economic, political, and educational systems [88]. For the con-

text of college computer and information science programs in the US, we characterized dominant groups as including white and Asian men who started college shortly after high school (not transfer students), do not have disabilities, have little or no financial or familial responsibilities, have English fluency, and have at least one parent who completed a four year college degree. Minoritized groups, then, are groups that are not positively privileged or favored and often stigmatized. In our context of study, minoritized groups include students who are women, non-binary, African-American/Black, Hispanic/Latinx, Native American/Indigenous, Pacific Islander, transfer students, not fluent in English, and/or first-generation, as well as students who have disabilities and/or have financial or familial responsibilities. While some may consider minoritized groups a small proportion of the population, these groups can actually make up a large proportion of society while still being minoritized by systemic injustices. Systemic cultures and norms tend to favor dominant groups and disadvantage minoritized groups.

5.2.2 Perspective Taking to Better Understand Students' Situations

To understand others' situations, humans rely (at least partially) on empathy. Empathy is a multidimensional construct, a set of interrelated yet distinct social behaviors and abilities that enable understanding of other peoples' unique contexts [77, 260, 245]. According to some models, there are two major kinds of empathy: affective empathy, which involves responding with one's own emotion to another person's mental or emotional state; and cognitive empathy, which involves the ability to understand another person's mental state [77]. For this investigation, we focused specifically on promoting *perspective-taking* behavior, which is a facet of cognitive empathy involving adopting others' points of view [77].

Perspective taking as a means of empathizing can be useful tool in understanding others' situations and needs. For instance, in design contexts, designers often use some form of perspective taking to try and better understand the needs of different groups of stakeholders. Prior work suggests that encouraging perspective taking through the use of personas or cognitive walkthroughs can help promote better understandings of minoritized groups [121, 199], including people with different genders [39, 275], cultures [6], socioeconomic statuses [198], and abilities [27, 210].

Supporting proper perspective-taking behavior can be challenging, especially in educational contexts. While there remains comparatively little work on promoting empathy in traditional programming courses, prior work in the area of HCI and software design education suggests that perspective taking can be difficult to teach and to learn [233, 121, 218, 219]. This is especially true in higher education computing contexts, which tend to be dominated by young, cognitively and physically high-performing students who may lack exposure to perspectives and viewpoints that are very different than their own [173]. Poorly executed perspective-taking activities may also lead to stereotyping, or making erroneous assumptions about a particular individual based solely on some limited information about them, such as their demographics. Stereotyping is a particular danger when asking people from dominant groups to perspective-take with people in minoritized or less contextually dominant groups [22, 36].

Stereotyping is an innate human behavior and cannot be done away with entirely [281]. For instance, if no particular traits about users are specified, software designers practicing perspective taking may fall back on implicit assumptions that a user is of a contextually dominant race, gender, age, culture, and class, who is heterosexual, affluent, comfortable with technology, and not disabled [68]. However, prior work suggests that providing enough rich contextual information about the target person's identities and behaviors can preclude some of the harmful effects of stereotyping [135]. Providing more information about a person's experiences and identities also can reduce tendencies toward single-axis analysis [68] which can erase the lived experiences of those with intersectional identities. To have the best chances of perspective taking being effective, comprehensive, and beneficial instead of harmful, providing more information about a person can support more holistic understandings of their unique situation.

5.2.3 Theory of Action to Guide Data Interpretation

Becoming aware of challenges is only a first (but critical) step towards addressing them. That is to say that showing somebody information will not necessarily translate to action. To scaffold this connection between data and action, we used Theory of Action, a framework that helps educators develop evidence-based stories that explains the specific changes they intend to make to improve

teaching and learning [41].

We drew upon Theory of Action (ToA) to connect instructors' interpretations of data from StudentAmp to action [41, 140, 56]. ToA relates individual actions to systemic functioning by articulating the underlying logic of work and starting assumptions about how and why actions will lead to desired outcomes [140, 56]. In ToA, actions involve information that stakeholders find valuable within their societal structures and can use to affect power dynamics [56]. While originally derived from studies of individual and organizational learning [7, 8], educational policymakers and administrators have used ToA to make changes to improve teaching and learning [140, 41].

Ongoing development and communication of a ToA can help instructors improve teaching and learning. Most of the work that uses ToA as a guiding framework to improve teaching and learning has primarily focused on schools and school districts teaching primary/elementary and secondary education (e.g. [161, 140]). It is typically an iterative process where leaders look at data to understand students' learning experiences, as well as reflect on how teachers' instruction affects student learning and how school principals' practices affect teachers' instruction [41, 140]. For ToA to make changes that would improve teaching and learning, leaders must articulate ToA and reform plans in terms that are compelling and understandable to multiple stakeholders and lay framework for ongoing "reform conversation" [140].

While leadership in primary and secondary educational institutions in the United Stated tends to be more centralized to school and district levels, post-secondary institutions (e.g. colleges and universities) tend to afford individual teachers (professors) more autonomy over their own classes [38, 130]. Leveraging this, we re-framed ToA to remove the principals and instead have teachers/instructors (e.g. professors, lecturers) as the leaders. This new framework positions instructors within the leadership role of 1) looking at data to understand students' learning experiences, 2) reflecting on how instruction affects student learning, and 3) identifying how the context surrounding the course affects instruction. This context affecting learning is broad and can include departmental policies (e.g. grade being used for acceptance into competitive major) and current events (e.g. global health emergency, political unrest).

To successfully use ToA to improve a course involves having course instructors serve as stew-

ards who continuously develop, communicate, and advocate for actions. Honig et al. 2010 saw stewardship as critical to the ongoing process of reform with ToA [140]. They identified tasks that stewards must take which we adapted from a school/district level to a course level: Ongoing development of a theory of action for the transformation of course; communication with others to help them understand the theory of action, including strategies used and underlying rationale for these strategies; and strategic brokering of external resources and relationships to support the overall course transformation process.

We framed StudentAmp as a tool to provide data that course instructors could use to inform the creation, iteration, and application of Theories of Action to improve their course. We scaffold the implementation of StudentAmp in courses within the context of creating Theories of Action (defined in [41]).

5.3 Design of StudentAmp

To understand the design of StudentAmp involves first understanding the positionality of the researchers who designed the tool, as well the design considerations we considered. We describe these first, then describe StudentAmp and how we intended for students and instructors to interact with it.

5.3.1 Critical Self-Reflexivity: Acknowledging Researchers' Positionality

This research required a reduction of people to the responses they were willing to share, so we acknowledge our assumptions and values in this section. By doing so, we follow critical approaches to quantitative methods which require researchers "to engage in critical self-reflexivity as a necessary first step for the long journey of deracializing statistics" [115]. As part of this process, we define assumptions and commitments that were the foundation of this research.

Firstly, we recognize the power structures and heterogeneity of people within different roles. Direct stakeholders in this research included teaching teams (including faculty members leading the teaching of a course and teaching assistants (TAs) supporting the teaching) and students in-

volved in the course. Even within these groups, there were differences. Faculty members leading instruction ranged from tenured research-track faculty who had worked for years at the institution to teaching-track lecturers with comparatively less teaching experience. The TAs were all undergraduate students, but their experiences with their respective courses ranged from never having taught or taken it to having years of previous experience taking and teaching the course. Full or part-time students attending these courses may or may not have been accepted to their major (admissions to CIS majors is very competitive and not guaranteed as part of admission to the university). Most students were enrolled to take the course, but some may have had listener status where they were not taking it for official credit. Some transferred from other higher education institutions (e.g. two-year institutions) with different norms, while others came directly from high school. A common theme across all stakeholders: The data we collect is a partial and biased lens into their experiences in a select few courses as part of a much larger educational experience.

Secondly, we acknowledge the tensions between labeling people in data, the *intersectionality* of people's identities (students in particular), and ensuring privacy. Intersectionality denotes the various ways in which ethnicity and gender (and other demographic labels) interact to shape the peoples' lived experiences [70]. Prior work has found that simplistic labeling of people can harm minoritized groups in particular. Labels of demographics (e.g. ethnicity, gender) academic experience (e.g. year in school, major, transfer or not), and lived experience (e.g. disabilities, familial language) are overly-simplistic. Furthermore, we needed to balance the nuance of the labels we select between how representative they were to diverse individuals and how anonymous they were such that instructors could not map responses back to individuals or small groups of students. Despite these risks, we believed that instructors could still use these labels in such a way to help stakeholders contextualize relationships between challenges and intersectional groups of people. Our perspectives align with the notion that "race is a measure of a relationship – not an inalterable trait" [316].

Finally, we acknowledge that models are always wrong in that they never fully reflect the complex phenomena we want them to represent, but they can be designed such that they are useful in informing stakeholders of hidden challenges. We framed the work we did as producing simplified

models of the complex phenomena of inequity in classes. We do not believe that in itself models will help, but they can support conversations, interactions, and interventions that address the systemic issues we sought to bring to light [59]. The objective of this work was to help instructors identify equity issues in their class, and that is a first of many steps in making learning experiences more equitable and just.

5.3.2 Design Considerations to Support Scalability, Context, and Privacy

We used the following design considerations to guide the design of StudentAmp:

(DC1) **Privacy/anonymity/safety**: Providing feedback should not harm a student. That is, instructors or other students should not be able to map responses back to specific students and information collection should not distress students. Because the anonymity of electronic responses systems can increase students' propensity to engage in providing feedback [110] and low response rates are a common issue with student feedback systems [271, 97], we believe that anonymity will support more inclusive participation.

(DC2) **Potential lack of awareness**: Students are not necessarily aware of all possible challenges they're facing and instructors are not aware of all possible challenges in their classes.

(DC3) **Person-in(fluencing)-environment**: Challenges are artifacts of inadequate support of students from their environment, not inadequacies of individuals.

(DC4) **Time constraints**: Students are limited in their availability and motivation to provide information and teachers are limited in their availability to analyze it. Furthermore, instructors need time to enact changes to their courses.

(DC5) **Relative disruptiveness of challenges**: Some challenges affect a student more or less than other ones.

(DC6) **Proximal, perceived value of participation**: Both students and teachers should perceive tangible and timely value for their participation.

(DC7) **Intersectional perspective of students**: *Intersectionality* denotes the various ways in which race and gender (and other demographic labels) interact to shape the multiple dimensions of underrepresented peoples' experiences race [70]. Whenever possible, StudentAmp should provide insight into the intersectionality and complexity of identity.

These design considerations are not without tensions. In this study, we focused on the tension where designing for equity-oriented goals involved a balance of protecting the privacy of minoritized groups (DC1) while also conveying their intersectional identity (DC7) such that others can better understand their experiences.

5.3.3 StudentAmp Enables Sharing of Contextualized Student Feedback

We designed StudentAmp as a responsive website to enable broad use by students and teaching staff. In initial interactions with the tool, instructors created their own *sections* to be an instructor of a new course. To grant other users instructor access (e.g. teaching assistants), instructors in the study provided a list of emails to researchers, who then manually gave those accounts access. Students then created accounts by signing up by email or Google account, using a six digit character code provided by their instructor to join a section as a student. Students could join multiple sections, and within the context of this study, we did find that two students were enrolled in multiple courses that used StudentAmp. Users could switch between being students in courses they were enrolled in (via section code) and instructors in courses for which they had instructor permissions, if they had access to any.

Student View: Sharing challenges, demographics, perspectives on other challenges

For this study, we designed StudentAmp's student view with the intention of enabling a student to be able to share feedback within a few minutes. Students could access StudentAmp from any modern web browser (e.g. Firefox, Chrome, Safari).

Figure 5.1 shows an example of the StudentAmp interface as it appeared to students. Students first shared "the biggest challenge in [their] life getting in the way of this class," with helper text

which prompted students to think beyond the scope of the class. Text also appeared which encouraged students to share more ("Keep writing so others understand your challenge!") if their response was < 100 characters and more if their response was ≥ 160 characters ("You wrote quite a bit! Consider condensing your writing so others can read it quickly."). From our pilot testing, we found that a message of 100-160 characters (approximately the maximum length of a tweet on the social media platform *Twitter*) represented sufficient description for another student or instructor to understand a challenge response without being too burdensome to read.

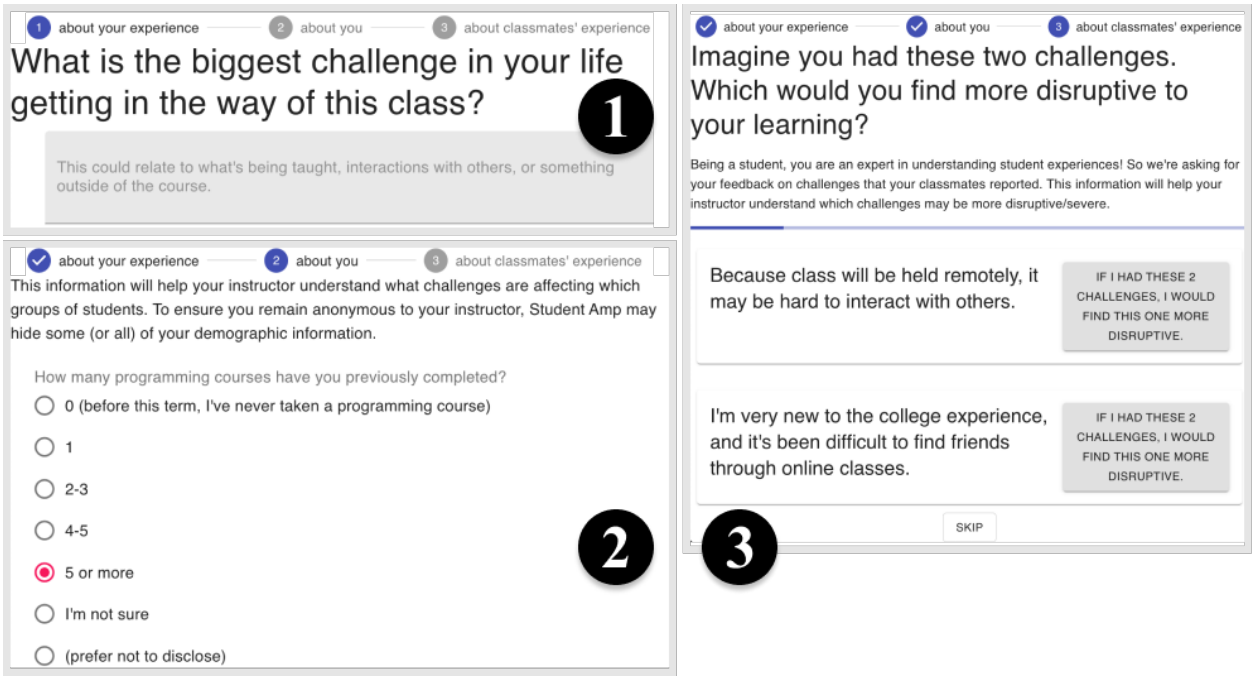


Figure 5.1: StudentAmp student view: Students shared 1) a challenge they faced, 2) demographics (pre-populated if they've previously filled in), and finally 3) meta-feedback by selecting which of two random challenges their peers shared was more disruptive, repeating this step two to eight times depending on class size.

After sharing their challenges, students had the opportunity to self-report demographics, as shown in Figure 5.1, step 2. We based StudentAmp's demographic questions on factors which prior work found to be impactful to students' learning experiences, including prior programming experience, whether they were a transfer student [168], whether they were first-generation, whether their

familial language is the same as the language the course is taught in (English), gender, ethnicity, and physical, mental, or social disability status. While we required answers for all multiple choice demographics questions, each question included an option for "(prefer not to answer)." If students had previously filled out demographic questions (e.g. in a previous feedback session), StudentAmp populated these questions with the student's prior responses. The demographics questions and options were as follows:

1. *How many programming courses have you previously completed?* 0 (before this term, I've never taken a programming course); 1; 2-3; 4-5; 5 or more; I'm not sure; (prefer not to disclose). [select one]

2. *Did you previously attend another college/university? (e.g. 2-yr community college, another 4 yr university).* YES, I previously attended another college/university; NO, my current college/university is the first one I have attended; I'm not sure; (prefer not to disclose). [select one]

3. *Are you a first-generation college student? (first-gen if parent(s) did not complete a 4 yr college/university degree).* YES, I am a first-generation college student; NO, my parent(s) completed a 4 yr college degree; I'm not sure; (prefer not to disclose). [select one]

4. *Is the language your family primarily speaks at home the same as the one used to teach this class?* YES, the language my family primarily speaks at home is the same as the one used to teach this class; NO, my family speaks a different language than the one used to teach this class; I'm not sure; (prefer not to disclose). [select one]

5. *Are you currently working or searching for a job? (select all that apply).* I am actively looking for a job; I work part-time (20 hrs a week or less); I work full-time (more than 20 hrs a week); I am neither working nor looking for a job; I'm not sure; (prefer not to disclose). [select one or more]

6. *What is your gender? (select all that apply).* woman; man; non-binary; prefer to self-describe**; (prefer not to disclose). [select one or more]

7. *What is your ethnicity? (select all that apply).* Asian; Black/African; Hispanic/Latinx; Native American; Pacific Islander; white; prefer to self-describe[3]; (prefer not to disclose). [select one or more]

8. *Rate to what extent a physical/bodily disorder hinders your learning experience.* 0: not at all; 1: to a small extent; 2: to some extent; 3: to a moderate extent; 4: to a great extent; 5: to a very great extent; I'm not sure; (prefer not to disclose). [select one]

9. *Rate to what extent a mental or social disorder hinders your learning experience.* 0: not at all; 1: to a small extent; 2: to some extent; 3: to a moderate extent; 4: to a great extent; 5: to a very great extent; I'm not sure; (prefer not to disclose). [select one]

After sharing their challenges and demographics, students finally shared *meta-feedback* on their classmates' responses, as shown in Figure 5.1.3. In this phase, StudentAmp randomly selected two challenges classmates had reported and asked students to imagine they had these two challenges, then to select the one that they imagined would be more disruptive to learning. To support data integrity of the meta-feedback, students could *skip* any responses. The meta-feedback pairwise comparison process was repeated 2-8 times depending on the class size ($2 \leq 2 * log(class\ size) \leq 8$).

Instructor View: Designing for instructor-led data exploration

Once students shared feedback, StudentAmp presented teaching teams with a report on student feedback, as shown in Figure 5.2. We designed StudentAmp to augment instructors' domain

[3]To support more inclusive demographics reporting [267], a free response follow-up question appeared with the prompt "Please self-describe your {gender, ethnicity}" after a student selected "prefer to self-describe" for gender or ethnicity. This information was not shared with teaching teams due to privacy concerns.

knowledge related to the course and their students by enabling exploration of contextualized feedback data. StudentAmp enabled this data exploration by 1) informing instructors of challenges, which student groups they affected, and how severe students perceived them to be and 2) supporting situated annotations through labels and notes so instructors could review previous findings.

As mentioned previously, instructors within our study each created a *section* for their course which students joined via a unique 6 character code. Each time an instructor wished to use StudentAmp to gather feedback, they created new feedback session. Once students shared feedback via StudentAmp instructors viewed the results, as shown in Figure 5.2. Teaching teams could review this data to identify how certain types of challenges disproportionately affected certain groups. Instructors and teaching teams could browse reported challenges, sorted by disrupt score (Fig. 5.2e), identifying trends and patterns. They could then create labels and assign them to challenges (Fig. 5.2f). We designed StudentAmp's labels to help teaching teams organize and prioritize feedback to better identify trends within and across feedback sessions. Similar to GitHub labels [120], teaching teams could define any labels they wanted, then assign one or many labels to any responses in any feedback session, similar to a tagging system. While the use of labels did require teaching manually labeling individual feedback (e.g. we did not provide automated labeling), it also enabled filtering of responses to use demographics charts (Fig. 5.2c) to explore how challenges affected demographic groups. In our study, we saw that this helped teaching teams understand which types of challenges disproportionately affected different groups of students across the nine demographic features we collected (enumerated in section 5.3.3).

In StudentAmp, we aimed to support beneficial and effective perspective taking for instructors. Teaching teams could look at challenges (unfiltered or filtered by label) and use demographic information associated with individual challenges (Fig. 5.2d) to perspective take. To help instructors better understand the nuanced ways that different challenges affected different students, we provided demographic information alongside each challenge to promote more informed perspective taking through the addition of richer contextual information. By doing so, we hoped to avoid stereotyping by encouraging instructors to see their students as unique individuals with many different kinds of identities and contexts, rather than defaulting to the assumption of an "average"

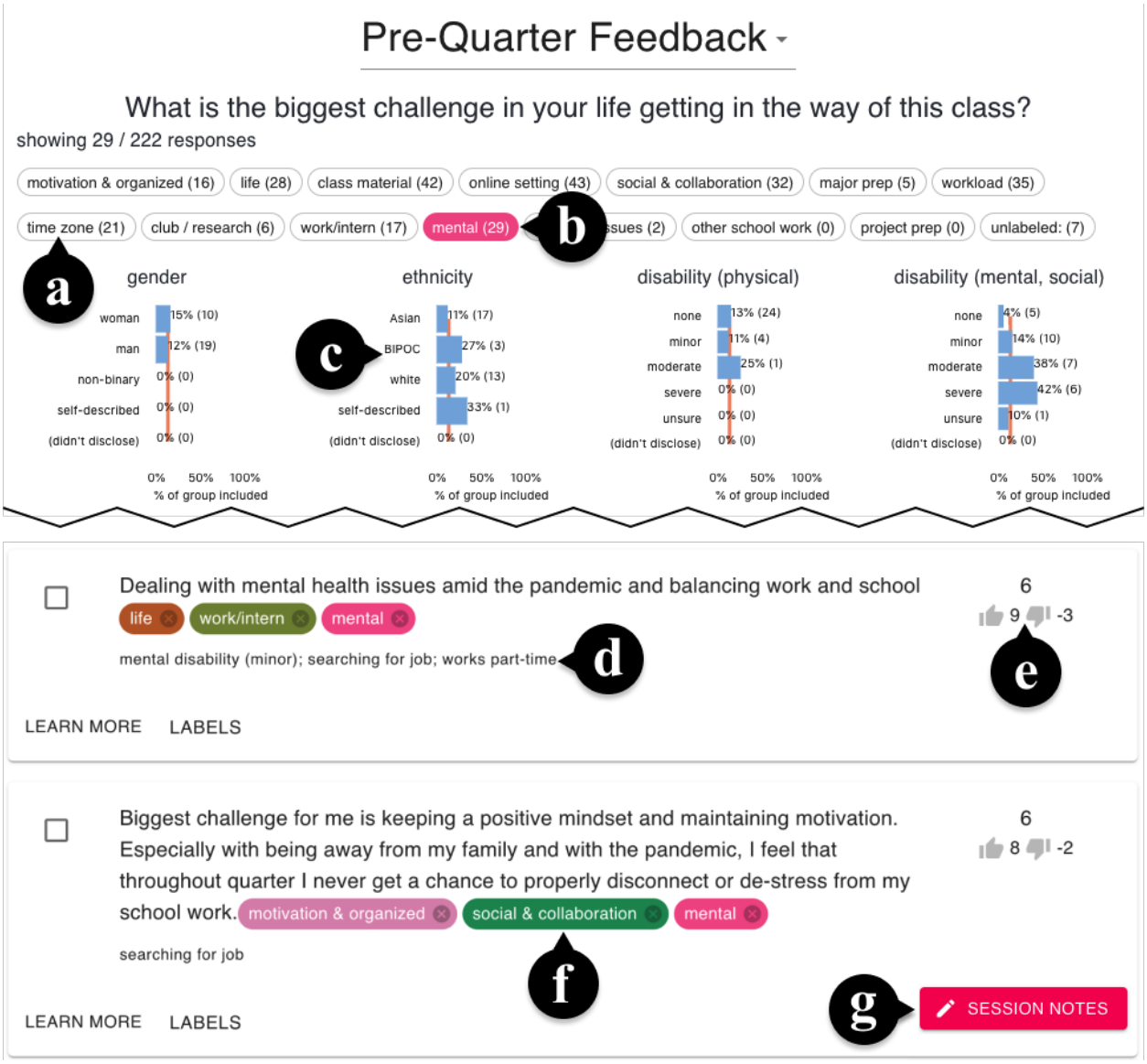


Figure 5.2: StudentAmp instructor view: Teaching teams could organize challenges by creating custom labels (a), which they could select to filter responses (b). The filters enabled teaching teams to use charts of demographic information (c) to see how challenges disproportionately affected certain groups (e.g. how the 29 challenges labels "mental" disproportionately affected BIPOC students and students with moderate or severe disabilities. The instructor view also included each challenge that included the selected label(s). Each challenge was contextualized with demographics for minoritized groups that students identified with (d), disrupt score (e), and labels that the teaching team assigned to that challenge (f). Teaching teams could also share collaborative notes (g), which have prompting based on our Theory of Action.

(likely dominant) student if no information was given. However, we also had to balance this against the need to protect student privacy and preserve anonymity, which we discuss further in our design considerations.

Throughout this entire process, teaching teams could use StudentAmp's collaborative *notes* (Fig. 5.2g) to review Theory of Action principles and also share notes with other members of the teaching team. Notes were open-ended text fields shared by all teaching teams. Each feedback session had a separate notes section from which instructors could write notes about their findings within the report. To help guide teaching teams according to our Theory of Action, we included prompting above the notes. This prompting asked teaching teams to consider 1) What's going on in students' learning experiences?, 2) How does what instructors are doing (or not doing) affect learning experiences? and 3) What factors external to the course help or hinder students' learning experiences? Which students? It then asks students to fill in the blanks as many times as possible: "If I/we __, then the course will change by __, so that students who are __ will be able to __."

Design of Disrupt Score

We designed StudentAmp's *disrupt score* with an intention to draw attention to challenges that were more disruptive to students' learning, rather than those that were simply more frequent or relatable to the majority of students. Prior work [69] and our pilot testing identified that students of minoritized groups reported challenges that were often unique from what their peers shared. As a result, we designed StudentAmp to support equitable feedback processes by organizing and presenting feedback in a way that went beyond showing the most common challenges. The disrupt score provided a quantification metric intended to represent how disruptive to students' learning some challenges were compared to other challenges.

In its most literal sense, the disrupt score was the net number of times a random student within the course decided a challenge was more disruptive to learning than another randomly selected challenge. It was based on a pairwise comparison of random challenges, as shown in Figure 5.1.3 (see also the description of *meta-feedback* above). We based this strategy on the Copeland method of pairwise scoring to determine ranked voting [252, 61]. Each challenge began with a score of

0. Each time a student shared meta-feedback (described in 5.3.3), the challenge they selected had its disrupt score incremented by one ($+1$). The challenge they did not select correspondingly had its disrupt score decremented by one(-1). No score changes occurred if a student chose to skip a meta-feedback comparison. These scores were then aggregated and used to track challenges in the instructor view in addition to being shown alongside each challenge (Fig. 5.2e).

A key assumption to the disrupt score was that students were able to perspective take and consider the disruptiveness of challenges they may not even have. Our initial pilot testing with students thinking aloud as they considered challenges found that students were more likely to select the challenge they most associated with having. As a result, for the study results we report in this paper, we adjusted StudentAmp's design by adding more information in prompts and button text to explicitly encourage consideration of disruptiveness rather than readability. As we report and discuss in the following sections, think-alouds with interviewed students suggested that students conceptualized the meta-feedback process in different ways (section 5.5.2), which resulted in several teaching teams having trouble interpreting disrupt scores (section 5.5.3).

5.4 Study Design: Deployed StudentAmp and conducted interviews

5.4.1 Context: Five large, remote, computing courses during a pandemic

To answer our research questions, we conducted a formative evaluation where five large computing courses used StudentAmp over the course of an 11 week term and we surveyed and interviewed professors, TAs, and some students twice throughout the term. An institutional review board (IRB) approved this study prior to any data collection.

We recruited culturally competent [297] instructors teaching large computing courses (100+ students). We chose to directly contact instructors who had previously demonstrated cultural competence [297] for participation in the study, as evidenced by their prior research efforts or participation in seminars on anti-racism. We focused on culturally-competent professors to avoid interpretations of student feedback and demographics focused on innate ability, such as certain people or groups having the "geek gene" and being more suited to computing [226, 192]. We selected large

courses to better ensure anonymity and to have a scale at which analyzing individual feedback of an entire course would become time-consuming and challenging. To avoid potentially unproductive and harmful uses of StudentAmp, we chose to exclude instructors who had previously stated beliefs about demographic groups' interest and abilities in computing [226].

Of the eight professors we reached out to, five professors teaching large computing courses participated in this study and used StudentAmp. Table 5.1 provides an overview of those five courses and how they used StudentAmp. Courses all took place within the same term, during which a global COVID-19 pandemic and a global reckoning with racial injustice were both ongoing issues. Because the university was shut down to in-person learning, all courses in this study were taught remotely, with students located in time zones all over the world. Given courses B, D, and E were introductory courses with no prerequisite requirements, some students in the study were first year college students who had yet to have an in-person college experience.

The five courses in our study were from the two departments of the same research university. This public research university in the United States was located in a major city with significant presence of large technology companies. All participants from this study (professors, TAs, and students) taught at or attended this university. We interpreted this study context to be Western, Educated, Industrialized, Rich, and Demographic (a.k.a. WEIRD, [133]). While most people around the world are not from WEIRD societies [132], many computing education contexts in the United States tend to be WEIRD societies [319, 278].

Courses A, B, and C were different courses offered within the computer science (CS) department. Based on data collected earlier in the same academic year that this study occurred, the department has 1,668 enrolled undergraduates. The CS department reported 31% of their students as female and 69% of undergraduates as male (only binary gender was collected on this survey); 8% of students as under-represented minority/URM (African American, American Indian/Alaska Native, Hawaiian/Pacific Islander and Latinx/Hispanic) and 75% as non-URM; 20% as first-generation (none of their parents completed four year college degrees); and 17% as international.

Courses D and E were the same course taught by different teaching teams and from the information science department, a separate department from the CS department. Based on enrollment

data collected the term after this study, the information science department had 526 undergraduates. The information science department reported 43% of undergraduates as female and 57% as male (only binary gender collected). The department reported 45% as Asian, 12% as URM, 16% as white, 28% unknown. Whether a student was Hispanic/Latinx was reported separately, with 4% (23) reporting as Hispanic/Latinx.

In addition to using StudentAmp, teaching teams relied on other tools and methods to collect student feedback for various purposes. Other feedback tools included feedback after assignments (e.g. to find out how long an assignment took, P-B), during lecture (e.g. "to ascertain skill acquisition," P-D), during the middle of the quarter (e.g. mid-term feedback conducted with an instructional consultant, P-A, P-C), and at the end of the term. They had previously used feedback to "respond to small conveniences that students requested" (P-C). They also invited students to reach out to them directly through email or similar mediums, but P-E noted how students from dominant groups tended to speak up more through these channels:

> P-E: *"students that have taken a bunch of programming classes and already done this stuff...they're the ones who speak up, who talk, engage [...] and this is all tied to race and gender."*

Teaching team demographics: Professors from Dominant Groups, TAs were gender-diverse

The teaching teams were led by five professors who were white or Asian men with prior teaching experience. Professors of all five courses reported as white or Asian men with no physical, mental, or social disabilities. Two professors (for courses A and B) were teaching their courses for the first time but had experience teaching related courses; the other three had taught that same course multiple times before. All five professors had been teaching courses remotely for at least the two terms prior to the study.

Of the 26 teaching assistants (TAs) who had access to or reviewed StudentAmp responses, 17 responded to a survey to report their their demographics. For ethnicity, 1 TA reported as Hispanic/Latinx, 12 as Asian, and 4 as white (non-Hispanic). For gender, 1 reported as non-binary, 9

as women, and 8 as men. Three TAs reported mental or social disabilities, such as anxiety. All but one TA who responded had previously either taken the course or an equivalent one to the course they were serving as a TA for. That one TA who had no prior experience as a student or TA for the course they were teaching had previously served as TA for the course B professor (P-B) for other courses in the past. Taken together, we can say that TAs were predominantly white or Asian, identified as a diversity of genders, and generally had prior experience with the course material as a student and/or a TA.

5.4.2 Data: StudentAmp responses, interviews w/ students & teaching teams

To answer our three research questions, we collected data from Student Amp and conducted two rounds of individual interviews with students and group interviews with teaching teams.

Data for RQ1: 810 challenges shared with StudentAmp

To understand what challenged students shared, we analyzed challenges shared by students with StudentAmp. In total, 604 unique students shared a total of 810 challenges across the five courses through StudentAmp over the duration of the 11 week term. The *Responses* column in Table 5.1 shows the number of responses in the feedback sessions across the five courses that used StudentAmp. We included incomplete responses because those responses were only incomplete because those students did not provide meta-feedback.

Data for RQ2: Rounds of interviews with 5 students of minoritized groups to understand perceptions

To understand factors that may have impacted what students shared with their instructors through StudentAmp, we conducted two rounds of semi-structured interviews with students. We recruited students who used StudentAmp and indicated interest in conducting follow-up interviews for compensation ($50 for two 1-1.5 hr interviews, which was slightly above minimum wage in the area at the time). We conducted interviews remotely, recording video (including screen share) and audio

with the consent of students.

We interviewed five students from minoritized groups. Of the 234 students who indicated potential interest in a follow-up interview, we identified 39 who were from minoritized groups (as evidenced by their ethnicity, gender) and/or reported a unique challenge. We contacted those 39 students by email and ended up interviewing all five students who replied. These five students came from three courses (three from course A, one from C, one from D). Three students were Asian women, one was a Hispanic/Latinx woman, and one was a white non-binary person[4]. Two were third year undergraduates studying majors related to computer and information sciences (CIS), two were first year undergraduates interested (but not yet enrolled) in CIS majors, and one was second year Master's student studying information science. While taking this course, interviewed students reported other commitments including submitting more than 100 job applications, moving physical locations, having familial responsibilities, and taking almost double the recommended course load.

We collected data via two rounds of retrospective think-aloud style interviews [95] to understand how students interpreted the prompts and how they decided to share what they did.

The first round of interviews occurred within the first five weeks of the term, after students had shared feedback using StudentAmp at least once. Students answered questions about their experiences learning computing and expectations about using StudentAmp, and then shared their screen as they walked through their prior usage of StudentAmp (three steps in Fig. 5.1), reviewing and reflecting on their previous responses and the context surrounding them. We asked them to interpret what each page was asking them to do and also about their perceived risks and benefits for sharing information at each page. We closed the first interview by asking students to reflect.

The second round of interviews occurred during the final two weeks of the term and included a sorting activity to have students consider additional challenges they did not share as well feedback on an example instructor view (Fig. 5.2). These interviews asked students to reflect on their course experience. We then had students look at all the challenges they shared over the course of term (2-3 challenges per students), asking them to identify potential trends, explain how they decided

[4]Regardless of their reported gender identity, we choose to refer to all interviewed students, professors, and TAs using *they/them* pronouns in this paper to discourage re-identification.

to share these challenges, and consider how peers potentially seeing their reported challenges affected their decisions to share.We then showed students the StudentAmp instructor view to get their perspectives on the utility and risks of contextualized feedback.

Data for RQ3: Rounds of interviews with 5 teaching teams to understand use of StudentAmp

To understand how teaching teams used contextualized student feedback for equity-oriented interpretation, we conducted semi-structured interviews with teaching teams who used StudentAmp.

Prior to or during the first two weeks of the course, we helped members of the courses' teaching teams (professors and TAs) set up their StudentAmp accounts online so they could collect and analyze feedback from their students. We also asked professors to fill out a survey to share their prior experiences with student feedback, the course they were teaching, and demographic information. TAs filled out a similar survey at the end of the term. To understand how teaching teams of large computing courses used different types of information to contextualize students challenges for equity-oriented interpretations, we conducted three interviews with members of the teaching teams of the five courses in our study: One interview with professors individually, followed by two group interviews with the entire teaching team of a course. The interviews with individual professors took place either before the term began or within the first four weeks of the course. In these interviews, we asked professors to share about the course they were teaching, how they intended to collect feedback from students about various aspects of the course (whether that method was StudentAmp or not), and about their personal definitions of equity in educational spaces (see Table 5.1).

The two group interviews were spaced out across the duration of the course, with one occurring towards the beginning of the term (after at least one feedback session with StudentAmp had been completed) and one at the end. We began the first group interviews by explaining the features of the StudentAmp interface (Fig. 5.2c,d,e,g), allowing time for the teaching team to ask questions as needed. We then engaged in a process following our Theory of Action (Section 2.3) in which the teaching team tried to understand students' learning experiences using StudentAmp, considered how their teaching practices or external factors might be affecting student learning, and

then proposing changes to the course and articulating how those changes might affect students of different groups, identities, or experiences. We asked one member of the teaching team to share their screen during this process, so participants could have a single shared view to discuss. After ten minutes of unstructured exploration of students' responses, if teaching teams were not already doing so, we prompted them to begin identifying broader patterns and trends they saw among student responses, create labels (Fig. 5.2a), and add labels to challenges (Fig 5.2f). After a few more minutes of labeling (the exact duration of which depended on each teaching team's level of discussion), we prompted teams to consider the demographic information provided at the top of the StudentAmp interface (Fig 5.2c) and explore any potential interactions of demographics with the labels they had created to explore how groups of similar challenges might disproportionately affect certain groups of students. The second group interviews followed a similar process, with the main difference being that the student responses under review were from subsequent StudentAmp feedback sessions.

Interviews were conducted remotely with the Zoom video conferencing tool, which enabled synchronous video and audio conversation, messaging, and recording. We choose to use this tool because all five teaching teams were familiar with it as they used it in their remote teaching. We recorded video and audio for all interviews. All members of the teaching team consented to the inclusion of their audio and visual recordings, as well as their survey data in this study. Participants were also compensated at a rate of $15/hr (approximately minimum wage for the area).

5.5 Analysis & Results: Analyzing StudentAmp Responses, Interviews

In the following sections, when we present quotes and other data tied directly to an individual, we gave them a unique anonymous ID associated with their role and course. These three part IDs take the form of <*role*> - <*course*> - <*optional number*>, where role is a character that denotes the individual's role within the class (S-Student, P-Professor, T-Teaching assistant), course is the character corresponding to the individual's associated course from Table 5.1 (A, B, C, D, or E), and number denotes an individual student within the course (from 1 to course enrollment) or an individual TA (from 1 to the maximum number of course TAs). Professors do not have numbers

attached to their IDs (e.g. P-A, P-B).

5.5.1 RQ1: Students shared challenges beyond the scope of the course

In total, 604 unique students shared a total of 810 challenges across the five courses through StudentAmp over the duration of the 11 week term. To better understand what students shared, we conducted an inductive thematic analysis and a subsequent round of qualitative coding using themes from the initial analysis. Three researchers participated in the qualitative analysis:

- The first author, a critical data studies and computing education researcher with seven years of research experience in data equity in computing education. The first author had expertise in designing interactions with data in educational contexts and mixed methods, having previously taught high school and college courses on introductory computer science and data science. He led the design of StudentAmp and the evaluation.

- The second author, a computing education researcher with seven years of research experience in HCI and design methods, including four years researching computing education within that space and a year of teaching experience in higher education computing contexts. The second author had expertise in qualitative methods and led the analysis of student-reported challenges.

- The third author, a computing education researcher with eleven years of middle and high school teaching experience, and two years of educational research in computing-related contexts.

First, all three researchers participated in collaborative affinity diagramming of 100 randomly selected challenges to inductively generate initial themes with a sensitizing concept ([29, 227]) of *types of challenges*. We used these themes as the basis of our code set. All three researchers collaboratively coded 40 (5%) randomly sampled challenges with the initial code set, discussing discrepancies and iteratively refining the code set and code definitions as needed. As can be seen in

many of the below quotes, even though we asked students to report the single biggest challenge they were facing in the class, students often reported multiple, often interwoven, challenges. As a result, we allowed for multiple codes per challenge during the qualitative coding effort. Our goal with this coding effort was to achieve consensus, so all codes applied to a challenge had to be agreed upon by all coders. After that, two researchers (the second and third authors) continued collaboratively coding another 160 (15%) challenges to ensure both of them had similar interpretations of the code set definitions and to address any confusions that arose. Finally, the two researchers divided the rest of the data and each qualitatively coded half of the remaining student-reported challenges. Once they finished their respective analyses, the two researchers asynchronously verified each others' code applications, marking any instances of disagreement. Finally, the two researchers met synchronously to discuss and come to consensus on the codes applied to the final 610 student-reported challenges, discussing interpretations and eventually achieving full agreement on all codes.

Table 5.2 shows our code set, comprised of the major themes that arose from our analysis and one "Other" code that was applied when a challenge was too unclear to code or when it otherwise did not fall into a coded category. These 16 types of challenges in Table 5.2 represent different learning difficulties that students conveyed to instructors through StudentAmp. For this analysis, we adhere to Hammer and Berland's perspective on qualitative coding [125], treating the results of our coding effort as organizations of claims about data rather than quantitative data in and of itself. As a result, we do not report specific code frequencies, instead focusing on representative descriptions of the themes observed within our data. In the following section, IDs preceding quotes indicate the speaker's role within the class (S for student), the course in which the student was enrolled (see Table 5.1), and a randomly generated number unique to each student within the course.

Even though StudentAmp's instructor view shows student-reported challenges alongside some demographic information about the student who wrote it, for the purposes of this paper, we choose not to report demographic information of the speaker for each individual quote. Instead, we report the demographics or our participants in aggregate, to illustrate the diversity of perspectives and experiences represented by the data while still preserving our participants' anonymity and reducing

the risk of community or peer re-identification. The following subsection contains quotes from 21 different students to illustrate the kinds of challenges students reported through StudentAmp. 20 of the 21 students who provided these quotes identified as belonging to at least one minoritized group, and often several. Of these 21 students[5], at the time of the study:

- 10 identified as women, 10 as men, and 1 declined to provide gender information.
- 12 identified as Asian, 7 as white, 2 as Hispanic/Latinx, 1 as Black/African, and 1 as Pacific Islander.
- 3 reported taking their first programming course.
- 4 reported attending another institution prior to their current one (transfer students).
- 5 reported as being first-generation college students.
- 6 reported that their family spoke a language at home than was different from the one used in the course (English).
- 6 students reported that they were currently working part-time, and 2 students full-time. 10 students were actively job-searching (e.g. applying to jobs, attending interviews). 5 students were neither working nor job-searching.
- 12 students reported that they did not have a physical/bodily disorder that hindered their learning experience (0 on a scale of 0-5). 8 students reported that they had physical/bodily disorders which hindered their learning to a minor extent (1-2 on scale), and 1 to a severe extent (4-5 on scale).

[5]Numbers reported for demographic facets may total more than 21, since students could belong to multiple categories simultaneously (e.g. holding more than one ethnic identity, or both working full-time *and* actively job-searching).

- 12 students reported that they did not have a mental or social disorder that hindered their learning experience (0 on a scale of 0-5). 2 students reported that they had mental or social disorders which hindered their learning to a minor extent (1-2 on scale), 5 to a moderate extent (3 on scale), and 3 to a severe extent (4-5 on scale).

The quotes presented below have been edited as little as possible to preserve authenticity. When clarifications or minor edits for anonymity were necessary, or when some less relevant parts of the quote were removed for length reasons, we designate any edits with square brackets.

We found the types of challenges students reported to fall into six broad categories (see the second-from-the-left column of Table 5.2 for an overview).

Course-related feedback focused on course & remote learning

The first category was that of feedback related to the course itself, represented by the *Course structure*, *Course content*, and *Remote learning* codes. Students who reported challenges with *Course structure* codes often wrote about their difficulties keeping up with the pacing of the course:

> S-A-102: *"The structure of this class because we simultaneously learn stuff for the assessment while learning the stuff for the following week forcing me to sacrifice one for the other."*

Other students who reported *Course structure* codes faced challenges with managing the course's required virtual learning tools or adapting to the instructor's pedagogical style.

When students reported challenges that contained *Course content* codes, they often mentioned the stress that came from trying to learn computing topics:

> S-E-28: *"I have never learned coding/data analysis ever in my life, things are just intimidating. IDK this class is STRESSING ME OUT."*

Many students who reported this challenge also reported not having much prior experience with computing, or who hadn't programmed in a long time.

Remote learning challenges were often reported by students who disliked the virtual format of classes, which was mandated by the university in response to the ongoing COVID-19 pandemic:

> S-E-6: *"The biggest challenge is just the general lack of structure that is inherent in online classes, regardless of how well the instructor organizes the course."*

Some students felt that virtual classes were not as conducive to learning as in-person classes, or that they did not feel like they got as much out of remote classes:

> S-C-109: *"I think the biggest challenge will truly just be the online nature of life right now. Screen fatigue is a big issue for me, & especially knowing that a programming class will require large amounts of screen time after class is a bit daunting. [...] I am worried about feeling intellectually gratified just because of Zoom fatigue."*

Other students who reported challenges containing *Remote learning* mentioned issues with poor Internet connections that made it challenging for them to attend virtual classes and difficulties connecting with peers and teaching staff.

Overall, challenges coded as *Course structure*, *Course content*, and *Remote learning* codes were likely similar to the kinds of feedback instructors might get with traditional feedback mechanisms (surveys, teaching evaluations, etc.).

Broader academic life focused on academic commitments beyond the course

A second higher-level category students reported through StudentAmp was that of challenges in their academic life outside of that particular course, represented by the *Other classes*, *Extracurriculars*, and *Academic context* codes.

Students who wrote about *Other classes* challenges often described the heavy course load they were taking alongside the course in which StudentAmp was used, forcing them to have to prioritize what work they did. Several students wrote about feeling overwhelmed by their academic workload:

> S-D-57: *"my course workload for my other classes is very heavy and my life is being consumed with all of my classes"*

Similarly, some students reported that being involved in various *Extracurriculars* impacted their available time:

> S-D-38: *"This quarter I am doing a few too many club activities and thus it's making it difficult to focus on my classes. It's my own fault."*

The university at which the study took place has a strong culture of student extracurricular involvement, in part due to the fiercely competitive climate within the university's computing-related departments.

Challenges having to do with the particular departmental or university-wide climate were reported by students in challenges involving *Academic Context* codes, such as those that described self-comparison to peers within the computing major:

> S-A-21: *"comparing myself to others; imposter syndrome; competitive environment in computing majors at [university]"*

The competitive, closed major system of the university, in which students were not guaranteed to get into their first choice of major, also contributed to students' stress and was listed as a common challenge due to the timing of the study, which occurred during major application cycles.

These three types of challenges represented by difficulties students were facing that still had to do with their academic lives, but that were explicitly outside the scope of the course itself.

Non-academic roles include familial and job commitments

A third higher-level theme of challenges which surfaced during our analysis was that of non-academic roles and responsibilities, represented by the *Home & family* and *Job* codes.

Students who reported *Home & family* challenges often described difficulties focusing on coursework in their current environments, which often co-occurred with *Remote learning* codes.

Sometimes, students simply mentioned that it was difficult for them to focus in their home environments, leading to them having to re-watch lectures or take extra time reading course materials. Other students wrote about their roles as caretakers of other family members, which took time away from their own responsibilities:

> S-B-18: *"I have a sibling that is disabled and another sibling that just started kindergarten. Because of this, I have to help my parents with making sure they attend their classes and do their homework, which is time consuming and also takes time out of other responsibilities around the house, plus work."*

Other household stressors, such as sick pets or siblings, also affected students' abilities to focus, whether due to stress about their well-being or having to provide transportation to medical appointments.

> S-E-27: *"My dog is getting eye surgery today, because of suspected cancer that caused Gluacoma. We don't know if the cancer is malignant yet but it's hard knowing that each day could be his last. :(so far our luck has been pretty bad but I really hope that the other tumors are benign like the one on his stomach. [...]"*

Many students worked jobs or internships during the quarter, or were actively job searching, as represented by the *Job* code. Often, this challenge was discussed in terms of time constraints, which sometimes made it difficult for students to engage with instruction or find time to complete their assigned work. Several students mentioned they were working full-time or part-time jobs alongside their full-time course loads. Other students described their roles as primary providers for their families:

> S-A-71: *"I am the only one supporting my family economically, so they depend on me working and getting money for our dependancies."*

Sometimes, the job environment or tasks themselves contributed to overall student stress:

> S-B-29: *"My job. I work at a homeless shelter. I only work two shifts a week, but I deal with a lot of very high stress situations (fights, 911 calls, suicidal ideation, sexual assaults, mental health crises, etc). Balancing school stress and job stress can often be difficult. It has been harder recently as I am regularly exposed to Covid positive individuals, so the likelihood of me catching Covid is very high."*

Both *Home & family* and *Job* codes represent roles and responsibilities in students' broader lives that placed demands upon their time and available physical, mental, and emotional resources.

Environment and context focus on broader contexts during a pandemic

Sometimes students reported challenges that had to do with their broader contexts, such as those that were classified as *Location*, *Political*, or *COVID-19* codes.

By far the most commonly mentioned challenge within the *Location* code was that of being in a different time zone than the university (likely due to remote learning mandates), making it difficult to attend synchronous classes, work with group members on class projects, and attend office hours.

> S-C-189: *"I think is time. I currently living in [other country] so that I need to get up at 5 o'clock to have this class."*

Students who described challenges relating to *Location* codes sometimes mentioned the weather in the place they were located impacting their mood, and thus ability to learn, as well.

Political codes were somewhat rare, but seemed to strongly impact students when they arose. This study took place at a U.S.-based university at a time when nationally relevant events were regularly occurring, which was stressful and distracting for students.

> S-C-117: *"It is hard to focus on the course during the global pandemic and political instability. I'm very distracted."*

Similarly, many students reported the ongoing COVID-19 pandemic as the greatest challenge detracting form their learning, as represented by the *COVID-19* code. Several students simply

wrote some variation on "COVID" or "the pandemic" as their response. Others described the impact the long-term stress and lockdown conditions had on their ability to learn and complete coursework, especially in the context of remote learning. Some students had family members or close friends that had contracted or were recovering from COVID-19 as well, which caused worry and extra stress:

> S-C-124: *"I guess the biggest challenge right now is the health of my family (relatives) since a lot of my aunts, uncles, and grandparents are old (thus highly susceptible to COVID). It was pretty stressful during Fall Quarter because one of my aunts got COVID and had to go the hospital for awhile"*

The *Location*, *Political*, and *COVID-19* codes represent categories of challenges that were persistent undercurrents in students' environments, causing worry and stress. These broader contexts in which students learned and lived certainly seemed to impact students' ability to engage with their classes and complete their coursework, though these were challenges not directly related to the course itself.

Well-being included physical and mental well-being

The fifth higher-level category of challenges students reported was that of personal well-being getting in the way of their learning, as is the case with the *Physical health*, *Mental health*, and *Isolation* codes. Students who mentioned *Physical health* sometimes described bouts of illness that caused them to fall behind in their courses or not feel well enough to do coursework.

> S-A-92: *"I had the worst case of mono for about the first three weeks of class, and now I am trying to catch up and relearn the basics of those first three weeks. I am having a difficult quarter"*

Other *Physical health* codes involved complications related to remote learning, such as sleep schedules being disrupted by having to attend synchronous classes in the middle of the night, or having physical symptoms from virtual learning through screens.

Students who reported *Mental health* challenges often wrote about stress, anxiety, and depression.

> S-B-41: *"Depression and anxiety, the pressure everyday"*

These kinds of challenges often co-occurred with reports of heavy course loads, difficult course content, or low engagement with others due to remote learning settings.

Somewhat similarly, *Isolation* codes had to do with students' socio-emotional well-being, and especially a lack of meaningful interactions with others.

> S-A-35: *"Being alone and lonely doing CS"*

Challenges that contained *Isolation* codes were often reported in conjunction with statements about the COVID-19 pandemic (due to quarantines, lockdowns, etc.) and remote learning.

> S-D-37: *"Being online is very isolating and does not allow for as much connection between students and instructors."*

All three these types of challenges – *Physical health*, *Mental health*, and *Isolation* – have to do with students' inner well-being, and health is an important prerequisite for effective learning.

Self-regulation involved motivation and time management

Finally, a sixth category of challenges students reported had to do with their own current self-regulation capacities and abilities. These were represented by the *Motivation* and *Time management* codes.

When students wrote about challenges that we coded as *Motivation*, they mentioned struggles with procrastination, distractions, focus, and feeling capable of completing coursework to their own standards.

> S-C-205: *"I have trouble finding the motivation to do school work nowadays."*

Oftentimes, *Motivation* codes co-occurred with *Remote learning* or *Isolation* codes, when students pointed out that their current environments were making it more difficult for them to focus or feel engaged within the course, or

students also wrote about the challenge of *Time management*, having to balance many demands and time constraints from their various roles as students, workers, family members, and humans in general. As a result, *Time management* codes co-occurred with many other codes, since students would often identify time as the challenge, then go on to describe the different facets of their lives that made time management difficult.

> S-C-16: *"I work about 30 hours a week as well as taking 17 credits this quarter, so time will be a challenge."*

Balancing all these demands could be difficult, especially when students struggled to set their own routines or to maintain a semblance of work/life balance.

> S-A-60: *"I only have so much time in the day/week for this class, other classes, and personal projects. It's honestly really rough to balance productivity and sanity :/"*

Both *Motivation* and *Time management* challenges had to do with students' current self-regulation capacities. Taken in context with many students reporting burnout from sources like the ongoing pandemic, competitive academic environments, and other stressors outside the class, it is perhaps not entirely surprising that students reported these kinds of challenges getting in the way of their learning.

Overall insights: Course-related feedback, external responsibilities, internal well-being

Overall, the six higher level categories of challenges students reported cover a wide range of potential learning difficulties. Some had to do with the course directly, as seen in the first section on course-related feedback codes. This set of difficulties is fairly similar to the feedback instructors might received from traditional methods (such as those described in the introduction), and perhaps the most "traditionally" actionable kinds of barriers to learning for instructors.

However, the other types of challenges reported through StudentAmp—challenges having to do with students' academic lives outside the course, their non-academic roles and broader contexts, their inner well-being and self-regulation skills—are particularly interesting to see. These latter categories of challenges might not show up through traditional student feedback methods. Instructors might never become aware of them if they exclusively used those kinds of methods, meaning that they likely would not be able to address them directly.

5.5.2 RQ2: Students' perceptions of sharing contextualized feedback

To better understand students' perceptions of sharing contextualized feedback, we conducted a thematic analysis on the transcripts of the interviews we conducted with students of minoritized groups (interviews previously described in Section 5.4.2). In the remainder of this section, we report four major themes students shared related to their perceptions of StudentAmp's purpose, as well as the ways in which different aspects of StudentAmp's design may have influenced what they shared or how they interacted with the tool.

Feedback was deemed important, even though purpose of StudentAmp was unclear

While interviewed students found feedback to be important to share, they expressed uncertainty about the purpose of StudentAmp. When we asked students what they thought the tool's intended purpose was, all five interviewed students expressed some uncertainty with two framing StudentAmp as a tool to improve the course. Students also compared StudentAmp to other feedback tools, such as direct emails, mid-term feedback, and end-of-term feedback. One student felt StudentAmp focused on and enabled conversation about a different context than other feedback they may have given:

> S-D-57: *"The feedback I said on StudentAmp was more 'What's going on with you? What are your challenges?' Whereas the feedback I gave for my instructor and my TA was 'Is the way they're teaching us helping? Are they getting back to us in time with questions?' I think they were just two different types of context in terms of what was*

being asked from us as students."

Two participants noted how StudentAmp was unidirectional: Professors could get feedback from students, but not the other way around. One participant didn't expect much benefit from StudentAmp because of this lack of bidirectional feedback:

> S-C-88: *"[StudentAmp] is a place where professors could hear the voices of students, to some extent. But students could not hear what professor thought [...] in this case, it's not [a] participatory process, not exactly like that."*

Challenges beyond the scope of the class were worth sharing, but privacy mattered

When discussing value and risks of sharing challenges with StudentAmp, interviewed students talked about how differing goals and relationships with instructors affected what challenges they shared.

StudentAmp asked students to share the biggest challenge in their life, where that challenge could go beyond the scope of the class itself. Three interviewed students noted how their biggest challenge was often beyond the scope of the course, including S-D-57, who provided a metaphor of school as one of many "bubbles" in life:

> S-D-57: *"I've always thought that it's important that teachers or professors or people you interact with know a little bit about who you are and a little bit about what's in your surrounding bubbles. School is only one bubble of a student's life, so knowing all knowing a little bit about those other aspects about student life, you know family, emotional, work, relationships, friendships. Just knowing a little bit about those things can give you general knowledge of how it could be impacting the school bubble."*

Another interviewed student also recognized that life outside of school affects class experiences, but they had concerns with that overlap occurring:

> S-A-148: *"Stuff in your personal life definitely affects class, but because this is a school thing it makes me not want to 'cross those wires' almost. I'm worried about*

> *inappropriate timing, or something. I don't want to be 'that person.' Which is weird or doesn't really make sense, because that's what the survey is asking about. But I don't know. That can kind of be the 'little fear' in the back of your head."*

Students perceived multiple risks to privacy and safety that sometimes limited what they shared through StudentAmp. One student noted that they didn't want a person they lived with potentially seeing what they wrote while they interacted with the tool, and therefore opted not to share particular challenges. Another student noted how professors may not be the best person to respond to certain types of challenges, such as mental health:

> S-A-128: *"Maybe I could share these mental health [challenges] but, I don't know. I think I wouldn't just share those with a teacher, because you could share them with a professional who would be able to better help."*

Sharing about themselves made students feel vulnerable even when supposedly anonymous. One student justified this risk by feeling it was necessary as part of validating their challenges to the teaching team:

> S-A-148: *"I just hope that a professor would never think that I'm trying to take advantage of their kindness. [...] I wanted to let [my professor] know that I was serious, while staying anonymous, while asking for an extension specific for me. So I don't really know how you can satisfy all that."*

Demographic information was seen as an asset, although risk of re-identification existed

Interviewed students identified ways that sharing demographic information could help instructors understand the positionality of students facing different challenges. However, they also identified potential risks related to acceptance of minoritized identities and potential re-identification.

The students we interviewed generally felt that demographics could help instructors interpret challenges that students shared, with a few caveats. For instance, one student felt that instructors

should have the cultural competence ([297]) to understand how societal structures affect learning experiences of students from minoritized groups:

> S-D-57: *" So if an as an instructor sees that the students who are BIPOC [...] and they're not doing as well as white students. I feel like a good informed instructor would know the racial understandings and the gender understandings as why certain groups with demographics will not be doing as great as other [groups]. Simply because of the world we live in, and the kind of.. structure our society is built upon. So I think a good instructor would know how to interpret that information and how to better help those students because they're all just trying to be at the same end goal."*

Another student found it relevant that StudentAmp asked demographic questions relating to other life commitments. Multiple interviewed students were searching for jobs while also taking this course, with one saying how StudentAmp helped connect job searching to course experience in a way that instructors previously had not:

> S-C-88: *" I appreciate that [StudentAmp] cares about whether we have jobs. Because I previously chatted with some other instructors and they said 'for job searches, that's kind of something different. I first care more about whether you learn well in this class.' Which makes me feel like I need to separate the job searching and course work. But they are not separate things. They're definitely things happening at the same time in my life."*

Interviewed students also shared concerns related to their identities being seen as valid by teaching team members. One student shared uncertainties about how accepting the teaching team would be with regards to their minoritized gender identity:

> S-A-148: *"A risk might be [professors and TAs] don't take me seriously if they disagree with my identity or don't think my identity is valid [...] If this is a class I'm trying to do well [in] and take seriously, especially a class that's relevant to my major*

> *where I might see this professor again, or it matters a lot that I do well in this class, I'd be worried about not being taken seriously."*

Another student identified the potential risk of re-identification. Even though courses were remote, teaching teams had additional information about students through their interactions with them as well as learning management systems. This information included a list of full names of all enrolled students. One student saw a potential risk of re-identification by connecting multiple pieces of demographic information with popularity of names in different cultures:

> S-A-128: *"With more [demographic information], like first-gen BIPOC, I feel like it would really narrow it down to a select few people [...] and there's tons of people who have similar names from certain regions. Like sometimes I can figure out where someone's from based off of their name."*

Seeing others' challenges fostered community, but students questioned disrupt scores

Several interviewed students noted how seeing other students' feedback helped them feel less alone. Recall that students saw random pairs of their classmates challenges through StudentAmp (as described in Section 3.3.1). One student found that seeing challenges similar to their own made them feel less alone. Another student found the variety of challenges their classmates reported reassuring to see, especially during remote learning:

> S-A-148: *" It was nice to see that there's a variety [of challenges], that people are going through different things, or getting different things out of the class. But then when it's the same challenges as me, that's also reassuring, because then I am like 'okay I'm not the only one that's facing this right now, or having difficulty with this part of the class.' "*

Challenges that our interviewed students reported tended to have negative disrupt scores in the instructor view, suggesting that classmates found their challenges less disruptive compared to other challenges. StudentAmp aggregated meta-feedback responses into disrupt scores. Disrupt scores

were the net number of times a classmate selected a given challenge over another challenge. The five interviewed students shared 13 challenges which had a median and mode disrupt score of -2, with the minimum being -6 and maximum being +2.

Students tended to question the aggregate disrupt score associated with challenges they reported, especially given how low they were. This did not bother some students, as they still personally felt their challenges were valid. However, other students recognized the impact that negative disrupt scores might have on instructors' awareness of needs of minoritized groups:

> S-C-88: *"As a user, when I see minus score, I would feel negative feelings definitely there was some judging behind it. And I understand probably people want to use this way to sort the results to help people browse information efficiently. But minority, disadvantaged, underrepresented people, they don't have many members or great numbers in the whole community. But still, they need to have their voice. It's not necessary because they are minority people and they have emergent needs, so other people would [...] probably be experiencing different things so that's my concern."*

One possible explanation for the observed variety in disrupt scores is that students interpreted the meta-feedback prompt differently than intended. To better understand how students perceived the request being made of them on the meta-feedback page, we asked interviewees to recall and reflect on their perceptions of it. During our first interviews, one student recalled interpreting the prompt as we intended (i.e., that they should choose which challenge would be more disruptive if they personally had it), two could not recall what they thought of the prompt, and the remaining two noted being surprised or confused by the prompt. Whereas the goal was to have students select the challenge they found more disruptive, one student interpreted the prompt as asking which challenge they also had and another as which challenge best represented the challenge they wrote.

Another potential explanation for low disrupt scores was that how students articulated their challenges affected how their classmates perceived them. One student thought that some students did not select their challenge because they used informal language (e.g., "whack" to describe a level of difficulty) and was not as verbose as some other students had been.

5.5.3 *RQ3: Teaching teams used demographics to support perspective taking about challenges beyond the scope of the course*

To analyze the interviews we conducted with teaching teams (previously described in Section 5.4.2), we conducted a collaborative thematic analysis on the transcripts with a sensitizing concepts of *teaching team interactions with StudentAmp* and *teaching team perceptions of student feedback*. Our approach was guided by the frequency of the topics raised by the teaching teams of the five courses as well as their relevance to answering our research question. In the remainder of this section, we report on the four major themes that arose from the interviews.

How teaching teams organized challenges reported in StudentAmp

To understand how teaching teams interpreted challenges that students reported in StudentAmp, we analyzed teaching teams' processes for creating and assigning labels to challenges. Professors and TAs could create custom labels to represent categories or groupings of challenges, and then assign them to challenges that they felt fit into those categories.

Some teaching teams focused on challenges most proximal to the course, such as those that dealt with course structure and course content, because they felt those challenges were most actionable. A TA in course A (T-A-6) read through all 139 responses in course A's week 4 feedback and labeled seven as "feature request,"[6] which were challenges they felt were actionable.

> T-A-6: *"'feature request' is a [label] name. It's just kind of actionable feedback that might help us make the course slightly easier for everybody and so it's more focused on [Course A] directly and things we do that may negatively impact how people learn."*

Rather than focusing on challenges that related to the course, other teaching teams looked at challenges more holistically. P-D worked with someone with qualitative research experience

[6]For context, of 7 challenges that TA labeled as "feature request" we coded four as *Course structure*, with the others being coded as *Remote learning*, *Motivation*, and *Other* in our analysis of student-reported challenges (RQ1, see Table 5.2)

(outside of the research team) to analyze the first feedback session and developed the following labels and descriptions:

1. *Mental Health*: Distinct from Stress below, a condition (such as Depression) that is experienced by the student

2. *Stress + Time*: A (temporary) feeling resulting from excessive demands on time and energy, lack of time to complete work, challenges of work life balance

3. *Motivation*: difficulties with being proactive, staying motivated, struggling to keep up, resisting burnout

4. *Learning Environment*: issues with the physical space (e.g., noise, distractions), issues with internet connectivity, also isolation: feeling alone, lacking a sense of community

5. *Group work*: difficulty collaborating with other students

6. *New to Coding*: expression of intimidation, frustration, difficulty getting started, feeling "lost"

While P-D took a more holistic, top-down approach, a TA in course C (T-C-2) took a more bottom-up approach by looking at the first three pages of responses (75) in the first feedback session and creating labels that reflected causes of challenges. After TAs in course C worked together to label all 222 responses in the first feedback session, the three most commonly used labels were *online setting* (43), *workload* (35), and *social & collaboration* (32).

While some teaching teams focused on challenges that were more directly related to the course and more actionable (e.g. course structure), other teaching teams considered challenges even if they were beyond the control of the teaching team (e.g. lack of social interaction in remote learning).

How teaching teams considered demographics

In general, teachings teams tended not to consider demographics unless we prompted them during the interviews or unless they had prior training related to cultural competence. For instance, a TA in Course D (T-D-5) was familiar with speaking about equity and privilege from coursework in public health, and T-C-2 and T-C-3 were both currently enrolled in a course on educational equity and diversity. However, when members of the teaching team did consider demographics, they connected challenges to rich personas of students that deviated from expectations of dominant groups.

When Course B was reviewing challenges, they focused for several minutes on a specific challenge from their data: *"I'm unsure of my ability to train my brain to think this way"* (the student reporting this identified as part of several minoritized groups, including having mental and physical disabilities, taking their first programming course, and being a transfer student; disrupt score -2 = 3-5). When considering the challenge, a TA (T-B-1) who had the same gender identity as the student who wrote the challenge connected the challenge to their own experiences as a student and thought to remind students in lab section that "it's normal to struggle a little bit; it is challenging material and you're learning really fast." When prompted about the students' demographics, P-B focused on the transfer student label to identify implicit assumptions in the course design:

> P-B: *"a transfer student that makes me think of someone who's more likely than not probably coming in from a community college so may have the academic background, but doesn't necessarily know the way to navigate a four year institution effectively [...] physical disability minor [...] that could be someone who may be wearing a cast [...] And then severe mental disability could be any number of things as well, but definitely that would be something that would interfere with student's schedule or their ability to focus or their self esteem and their confidence and actually passing the course and and completing the assignments."*

P-B then went on to propose improvements such as clarifying how to use university email and access the course's learning management system, reassuring students that they could succeed in

the course, and granting individual extensions on assignments.

Showing demographics did not necessarily translate to understanding on what to do with the information. When considering demographics for a challenge we labeled as job and course structure (*"I work 40 hours a week 8[A.M.] - 6[P.M.] so it can make it challenging to connect with TAs who only offer office hours during the middle of the day during the week..."*), P-E was uncertain how to consider the demographic information this student who identified as part of several minoritized groups, including being a transfer and first-generation student, working full-time, and job searching.

> P-E: *"In my head mentally, I still often see students as the standard undergrad 18 to 20 year old [...] living on campus or an apartment somewhere. The 'non-traditional' students as they're often frame are a different kind of aspect. I'm not sure what to do here now...in my head, this a challenge for everybody."*

P-E framed problems of students from minoritized groups as similar to students from dominant groups, but more severe. And while demographic information challenged the archetype of the "standard undergrad," P-E was unclear how to use this new information.

P-D and P-E reviewed feedback together. While P-E was unsure how to consider demographic information, P-D connected this challenge from a student from minoritized groups to systemic challenges at the university:

> P-D: *" what it feels like to read something like this is it is somewhere between heartbreaking and frustrating and angering. That is instructors were put in this really awful position where the university pressures people to take more courses than they can handle because [tuition] is so expensive."*

How teaching staff considered disrupt score

Disrupt scores were not taken literally, as the affordances of the interface design resulted in confusion amongst teaching teams and comparisons of such diverse challenges potentially confounding the aggregate disrupt score.

Because StudentAmp ordered challenges by disrupt score, affecting how teaching teams viewed and interacted with the data. StudentAmp ranked in each feedback session by disrupt score, resulting in teaching teams seeing challenges ordered from highest disrupt score to lowest. Disrupt score was shown as a number that is the difference between a positive number next to a thumbs up icon and negative number next to a thumbs down icon, as shown in Fig. 5.2e. It represented the net number of times a students selected that challenge over a random other one when asked to determine which challenge they found more disruptive.

One TA looked at the first three of nine pages (75 responses) and created and added labels to them, using the disrupt score as a stopping criteria:

> T-C-2: *"I just [went] over all the responses that are first three pages of responses and try to categorize and that basically settled all the tags [...] I think there is like a thumbs up, thumbs down. So I guess students get to like and dislike, or agree or disagree with certain statements. So up to page three, getting to a point [the disrupt score] is up one. So I think that's probably enough for telling what the students find most challenging."*

But later on, as they submitted a response as a student to explore the student view, the instructor questioned their interpretation of disrupt score as a measure of how many students related to a challenge:

> T-C-2: *"the net disruptive score [...] I don't like the minus sign. The first time I read this, I thought it means 'people do not agree with this.' [...] And then later on, when I actually [submitted] a student response, I pretended [I] was the student. I review the process, and then I realized it's asking which one is more disruptive instead of which one resonates more with your circumstance. I felt that there's a difference there and it's not really clear when I first viewed it."*

The thumbs up and down icons were similar with iconography used in many software interfaces to indicate rankings, but indicated something slightly different because selecting a response over

another is not exactly the same as "upvoting," and choosing not to select a response is not exactly the same as "downvoting."

The professor for course A felt that the disrupt score was difficult to interpret in part because the diverse content of challenges made some comparisons uninformative. They gave an example of the disruptiveness of the global pandemic as being far greater than anything related to this class:

> P-A: *"The pandemic's huge, and to say something is less disruptive than a global pandemic that's not a very high bar. EVERYTHING should be less disruptive than a global pandemic. But it is nice that nobody is saying 'yeah there's something so wrong with this class, that is the biggest problem, even though there is a pandemic,' I suppose that's a win."*

Teaching teams really began to question low disrupt scores which corresponded to challenges they thought were disruptive. Of the responses which reported a challenge, responses with the lowest disrupt scores in each feedback session varied in content, but tended to be either vague or involved challenges that were not relatable to most students. The challenge with the lowest disrupt score in our dataset (-19=1-20) related to a phone being a distraction:

> S-D-69: *"My phone is the biggest challenge I am very addicted and it takes all of my focus during class times."*

TAs from course D labeled this challenge as *Motivation*. When reviewing this challenge, P-D wondered if other students were either downplaying the severity of the challenge or questioning its authenticity:

> P-D: *"Maybe people sort of downplaying the severity of that [challenge] or maybe that addiction is maybe something some people don't think is real."*

After hearing P-D's comment, P-E proposed an explanation wondering how the "non-clinical" language and focus on the phone may have caused peers to not take this challenge seriously:

> P-E: *"I wonder how many people are like 'oh haha yeah my phone is a joke too oh yeah no totally animal crossing is like definitely like my biggest distraction at the moment.' Whereas it's supposed to being like 'I have severe problems focusing on anything and I'm constantly spending time doom scrolling,' like there's actual things are going on there, but because of how [challenges] are framed and presented, they get read in very different ways."*

The disrupt score deviating from expectations caused P-D and P-E to propose alternative explanations for disrupt scores that deviated from their expectations. These explanations included students misunderstanding the meta-feedback prompts or not taking challenges seriously because of the way students reported challenges or because peers were unfamiliar with the challenges

P-A also noted a similar case where a challenge may have gotten a low disrupt score because it only affected a subset of students. For week 7 feedback in course A, we coded the challenge with the lowest disrupt score (-15 = 1-16) as location:

> S-A-117: *"Time difference"*

After seeing multiple challenges mentioning time zone differences with negative disrupt score, P-A acknowledged the impact of this challenge on a select few students:

> P-A: *"There's a few [responses] on timezone differences, but they are pretty consistently downvoted. Which I've heard enough now to believe that the timezone difference is a big deal for a small population of students."*

This theme of a low disrupt score for mental health related challenges also appeared in Course D. In week 1, a student who reported being part of several minoritized groups, including having a mental disorder which severely impacted their learning experience, stated a challenge related to severe depression and suicidal thoughts as part of week 1 feedback for Course D. This challenge received a disrupt score of -2 (6 - 8). In response, P-D talked about mental health in their next lecture and provided links to university resources to support students' mental health.

> P-D: *"So the students ranked severe depression and suicidal thoughts they rank that lower than the other thing you know eight out of 14 times which either means that students misunderstood the prompt or they misunderstand severe depression and suicidal thoughts."*

For week 7 feedback in Course A, S-A-202, who reported belonging to multiple minoritized groups including having a minor mental or social disorder which hindered their learning experience, stated a challenge related to depression and having to go to regular doctors appointments to manage their depression. This challenge had a disrupt score of 1, with 9 students selecting it over another random challenge and 8 students selecting a random challenge over this one. Upon seeing the lower disrupt score, P-A questioned the disrupt score and gave an explanation related to a lack of familiarity with depression:

> P-A: *"'My depression,' that's unfortunate. That should have a much disruption rating...I would bet you that the eight people who downvoted that don't have any history of mental illness or depression in themselves or their families, because if you know what that's like– that should be much higher."*

Teaching teams used StudentAmp to adjust course, training TAs, discuss systemic issues

While this evaluation focused on how students and teachers interpreted contextualized student feedback, we also identified three ways that teaching teams used this information.

First, teaching teams considered changes to course structure to be more accommodating to diverse students and their needs. Examples of this include supporting more community building amongst students who felt isolated by remote learning, making deadlines and office hours more flexible to accommodate students from different time zones and those who worked jobs, and supporting students' mental health by raising awareness of free university resources and finding new ways to express empathy.

Another way teaching teams used StudentAmp was for development opportunities for the teaching teams themselves. While we framed StudentAmp as a way to improve teaching practices

by responding to student challenges, P-C saw StudentAmp more of as an opportunity to discuss diverse student experiences with course TAs to build empathy.

> P-C: *"The point of a StudentAmp survey is not to collect data on students or even to improve instruction (as in formative course evaluations), but rather to amplify student experiences that might otherwise fall between the cracks. In response, the instructor's 'fireside chat' offers a natural mechanism for the instructor to recognize and validate student experiences revealed through StudentAmp."*

Finally, StudentAmp did foster some discussion about systemic issues which extended beyond the course and even beyond the university. Teaching teams felt limited in what actions they could take in the middle of their large remote courses, but they still used StudentAmp to discuss broader systemic challenges that their students faced:

> P-D: *"There are certain dials that [professors] can turn and they're still contextualized within the university system where all the other courses they're taking have firm deliverables and it's contextualized within a broader social and economic system in which, if they don't get a good job they can't go to the doctor later or pay off their huge loans"*

5.6 Discussion: Tension between providing context and protecting well-being

In this paper, we designed, developed, and evaluated StudentAmp, a student feedback tool that supported equity-oriented goals by asking students to report challenges that may be beyond the immediate scope of the course and contextualizing those challenges with self-reported demographic information as well as an aggregate score reflecting peer perceptions of challenges. We evaluated StudentAmp with five large computing courses (150 - 750 students) that were taught remotely during dual pandemics of COVID-19 and racial injustice. We found that students used StudentAmp to share challenges beyond the scope of the course, including challenges in non-academic roles and challenges related to the well-being of themselves, their families, and their peers. Interviewed

students identified a tension between wanting to share more information about their lives to justify their needs while also wanting to preserve their anonymity and safety. Teaching teams used this contextualized feedback to consider not just challenges, but also the positionality of the student reporting the challenges. Taken together, this paper contributes a design exploration into how contextualizing student feedback can support equity-oriented goals in large, remote computing courses.

In this section, we describe multiple ways to interpret our findings. We focus in particular on the primary tension that this study: How to support equity-oriented goals by contextualizing student feedback while also ensuring the privacy, well-being, and trust of students, especially students of minoritized groups.

5.6.1 Limitations: Self-selection bias

One interpretation of our findings is that they may lack validity because of self-selection bias throughout the study. Indeed, participants self-selected into the study at multiple phases, with instructors choosing to participate in this study and use StudentAmp, a subset of students choosing to share feedback with StudentAmp, unequal usage of Student across the five courses in the study, and five students choosing to interview with us. This type of bias was likely exacerbated by the context during which we conducted this research, in which students were still adjusting to remote learning and trying to do so during dual pandemics of COVID-19 and racial injustice. It was partially due to these challenging times that we decided to conduct this research, because students were facing new or worsening challenges to learning and because teaching teams needed to understand how to support them, but lacked the capacity to do so.

We tried to mitigate self-selection bias by compensating instructors, TAs, and students for their time and allowing for flexibility in regards to what participants were comfortable disclosing and with scheduling. We also conducted repeated (at least two) interviews with each teaching team and student interview participant. A follow-up survey with all students could have provided corroboration to themes that we identified in our interviews, and future work on this topic would do well to explore more deeply the saturation and relative frequency of the themes we surfaced.

As a result, while self-selection bias is indeed a limitation of this study, we can still nevertheless interpret our findings as a contribution to a larger body of knowledge that seeks to understand how to design contextualized student feedback for equity-oriented goals.

5.6.2 StudentAmp focuses on students' experiences in a broader context

Another interpretation is that StudentAmp is similar to other feedback tools that already exist. All five courses that used StudentAmp also used other common student feedback techniques, such as online surveys, direct conversations, mid-term feedback, and end-of-term feedback. However, interviewed students felt that StudentAmp was different than these techniques because it afforded them an opportunity to share feedback about their experiences beyond the immediate scope of the course, which most other feedback tools focused on. The anonymity of StudentAmp enabled them to be more open and vulnerable in ways that identifiable techniques such as direct conversations and emails do not afford. We found that seeing these broader challenges as well as demographic information helped teaching teams discuss student challenges that went beyond course structure and affected certain groups of students disproportionately (e.g. mental well-being, timezone differences, and job or familial commitments). In these cases, StudentAmp enabled broader and more contextualized feedback, and future work can explore how to better integrate aspects of StudentAmp into different pedagogical practices with student feedback.

5.6.3 StudentAmp avoids reducing students to labels, supports perspective Taking

Another interpretation is that StudentAmp is harmful through its reductions of people's diverse lived experiences. Data technologies can promote material, symbolic, and other violences by reducing people to broad demographic groups and promoting incremental changes that do not address larger systemic issues [137, 136]. To avoid data violence related to stereotyping, we designed StudentAmp to show multiple dimensions of demographic information, enabling consideration of intersectional identities for perspective taking. StudentAmp enabled aggregation and reduction of challenges (through user-defined labels), but not by demographic groups. These design decisions

required teaching teams to consider challenges as contextualized by multiple dimensions of student demographics. We found that discussing demographic information was challenging for teaching teams, with members with more cultural competence (e.g. training in culturally responsive teaching) more able to lead these discussions.

Showing high-dimensional demographic information likely enabled consideration of intersectional identities in perspective taking. Rather than see one or two demographic labels (e.g. gender and ethnicity), teaching teams could see up to nine (enumerated in section 5.3.3). Making visible multiple demographic features at once provided a mechanism to discourage simple and harmful stereotyping and instead supported more nuanced perspective taking. Furthermore, this feature provided teaching teams the option to talk about aspects of demographics they were most comfortable with. For example, we noticed in multiple interviews how teaching teams discussed that a student was a transfer or first-generation and how this was more common amongst BIPOC students. All five professors and many TAs identified as coming from dominant groups, so speaking about some aspects of demographics may have been less comfortable with (e.g. gender, ethnicity) and other aspects of demographics may have been more comfortable with (e.g. transfer student, first-generation student). From our data, it is unclear what the impact of using more comfortable demographic labels in discussions about equity might be, since it could be an inroad to starting "more difficult" discussions, or it could obscure more direct systemic issues about gender and ethnicity. This could be an exploration for future work.

5.6.4 StudentAmp affords some anonymity, but privacy risks still exist

Another interpretation is that StudentAmp poses a privacy risk to students. In particular, multiple demographic labels make common privacy guarantees such as k-anonymity [276] impossible. Furthermore, background knowledge attacks [184], where an adversary (e.g. member of teaching team) uses information from other data sources to re-identify a respondent may be of particular risk within a learning context. Given how instructors, TAs, and students know and interact with each other frequently and instructors and TAs have additional information about students (e.g. names and pictures). We mitigated this risk by recruiting culturally competent instructors and not enabling

aggregation or filtering by demographic group in the StudentAmp interface. However, one interviewed student noted how re-identification could still be possible with background knowledge (e.g. name to infer ethnicity or gender) and how this could be especially dangerous if an instructor or TA was less culturally competent. Future work can explore how to reduce the risk of re-identification by disclosing demographic information in a non-uniform way that provides relevant context for specific challenges (e.g. mention mental disabilities when a challenge relates to mental health, but not familial language) or adjusting their buckets (e.g. use BIPOC label if there are only a few Black/African students but more Hispanic/Latinx students). McDonald's framing of privacy from a vulnerability perspective could guide improvements to privacy and safety of minoritized groups in particular [193].

5.6.5 Organizing information in a scalable, equitable, and privacy-preserving way is an open design space

Yet another interpretation is that StudentAmp was not equitable because it did not draw attention to the needs of minoritized groups. We tried to show the information about challenges in an equitable way by doing two things: First, StudentAmp attempted to call attention to minoritized groups' needs by showing demographic information with challenges only if that person was from a minoritized group. This enabled perspective-taking behavior, as discussed in Section 5.6.3. Second, StudentAmp also ranked challenges by disrupt score in an attempt to organize challenges by disruptiveness, but teaching teams found disrupt scores to be a confounded measure.

The goal of the disrupt score was to introduce a mechanism that organized challenges by disruptiveness and not frequency or commonness, as minoritized groups could make up a small proportion of students and/or have unique challenges. In our interviews with teaching teams, we found multiple instances of them being uncertain of how to interpret the apparent contradiction of low disrupt scores for seemingly severe challenges, such as mental health concerns. Potential explanations include StudentAmp providing an unclear prompt or explanation for the meta-feedback task, as evidenced by multiple interviewed students interpreting meta-feedback prompts in different ways and being uncertain about what they were doing. Asking a student to consider the

disruptiveness of challenges that they may never have encountered for the purposes of calculating the disrupt score is also a difficult task. Future work can explore potential improvements including the effect of different challenge selection procedures (e.g. asking students to consider challenges that are similar to their own, or from people from similar demographic groups) or contextualizing meta-feedback with demographic information or other information to enable more informed consideration.

An improved disrupt score is only one of many potential mechanisms to support the organization of information in an equitable yet efficient way while also ensuring privacy and well-being of minoritized groups especially. A large remote course with hundreds of students is a dynamic environment where students' needs must be met in a timely manner. Efficiency in data collection and analysis is key to taking action. Furthermore, students of minoritized groups may face challenges that are unique and unlike challenges their peers have, so there must be a way to organize information so perspectives of minoritized groups are not lost. Lastly, addressing students' needs often involves vulnerability and asking students to share information about their experiences both in and out of the class. Because students of minoritized groups are perhaps most vulnerable to privacy violations, we must ensure the privacy and well-being of minoritized groups in particular. These tensions present a rich design space for future work that values equity and human well-being while also wrangling with the pragmatics involved with the need to scale.

5.7 Conclusion: Cultural competence and demographic data supports perspective taking

Connecting the findings of this design exploration back to the framework I defined in Fig. 2.2, this study suggested *cultural competence* supported teaching teams' interpretations of demographic data to consider perspectives of minoritized groups. Instructors and TA had developed their cultural competence through additional training (e.g. seminars, coursework, research) demonstrated more capability to engage with the demographic data to consider perspectives of minoritized groups. This helped teaching teams shift away from an implicit assumption that all students had experiences similar to their own or that students aligned with dominant groups. And similar to the workshop with curriculum designers in the previous chapter, *prior knowledge* related to teaching and taking

the course helped teaching teams contextualize challenges students reported, and *perceptions of power relationships* focused the conversation largely on factors that they could control.

An alternative design could have framed the cultural competence of instructors and TAs as a shared asset to share with the entire teaching team. Some instructors and TAs demonstrated a hesitancy in engaging with demographic information (gender and ethnicity in particular) implied a potential uncertainty in how to engage with this information. To better engage all instructors and TAs, we could have designed the collaborative sessions as opportunities to develop cultural competence amongst the teaching team as well. By creating a space for instructors and TAs to ask (perhaps anonymous) questions, we could leverage strong cultural competence of some members as a shared asset to benefit the entire teaching team. By framing cultural competence as a shared asset, teaching teams may have been able to have richer interpretations of student feedback data to identify how challenges disproportionately affected certain groups.

Table 5.1: Context about five courses in study and their Student Amp usage: Course content and structure, professors' definitions of equity, number of students who completed course, number of students, number of responses in each StAmp feedback session (number of incomplete responses in parentheses), and who amongst the teaching team had Student Amp access.

ID	Course Content	Course Structure	Prof's dfn. of equity	Students	Responses	Access
A	Intro. to CS II. Data structures, complexity, sorting in Java. One prerequisite course.	optional, recorded lectures w/ professor (3 / wk) and lab section w/ TA (1 / wk)	"everybody should be able to succeed... my focus has been to remove as many structural barriers within the course"	500-750	wk 2: 148 (+3) wk 4: 139 (+7) wk 7: 86 (+8)	professor & lead TAs (6)
B	Intro to CS for non-majors. Control & data abstraction, file processing, visualization in Python.	optional, recorded lectures w/ professor (3 / wk) and lab section w/ TA (1 / wk)	"there {[}are{]} a lot of cultural problems in the CS space... elitism and racism, to some degree, and sexism. It was an important to me that I can try to address those impressions"	150-200	wk 0,1: 30(+8) wk 2: (8) wk 3: (1) wk 4: (2) wk 7: (4)	professor only (all 7 TAs saw responses)
C	Design, analysis, and critique of data structures and algorithms in Java. Course A is prerequisite.	Students meet in small groups w/ TA 4 days / wk. Assignments: three group projects, each two weeks long.	"How are we engaging with students' identity in the course"	250-300	wk 0,1: 218(+4) wk 4: 19 (+10)	professor & head TAs (3)
D	Introduction to collection, storage, analysis, and visualization of data in R.	optional, recorded lectures w/ professor (2 / wk) and lab section w/ TA (1 / wk)	"students should have equal probabilities of success regardless of their background... putting forth the support and resources necessary to balance out the playing field "	150-200	wk 1: 58 (+4) wk 4: 35 (+5)	professor & all TAs (10)
E	(same as D)	(same as D)	"[students are] all able to get to the same ending objective... put the most resources that I have (time and energy) towards supporting [students with the farthest to go]"	150-200	wk 2: (8) wk 5: 13 (+7)	professor only

Table 5.2: Types of challenges that students reported through Student Amp, used as the codeset for our RQ1 analysis. The rightmost columns indicate courses in which at least one student reported an instance of that challenge. The leftmost columns represent categories of themes which arose during our analysis.

				Courses with 1+ reported (Total num. challenges)				
		Code	**Definition: Challenges related to...**	**A (386)**	**B (46)**	**C (253)**	**D (104)**	**E (31)**
	Course-related feedback	*Course structure*	The "how" of the course: assignments, materials, speed, tools and platforms, teaching methods, office hours, etc.	✓	✓	✓	✓	✓
		Course content	The "what" of the course: Topics of instruction, such as computing, math, or programming	✓	✓	✓	✓	✓
		Remote learning	Online instruction methods and tools, not taking the course in-person	✓	✓	✓	✓	✓
External responsibilities, roles, and contexts	Broader academic life	*Other classes*	Workload or time constraints from taking other courses concurrently	✓	✓	✓	✓	✓
		Extracurriculars	Student life-related activities outside the scope of courses, such as clubs, sports, etc.	✓	✓	✓	✓	✓
		Academic context	Departmental or university-wide academic landscape, such as the highly competitive student climate, changing majors, etc.	✓		✓	✓	✓
	Non-academic roles	*Home & family*	Household or familial responsibilities, including roommates, partners, and other relationships	✓	✓	✓	✓	✓
		Job	Work and internship-related activities, including job searching	✓	✓	✓	✓	✓
	Environment and context	*Location*	Geographic location, especially that which differs from the university	✓	✓	✓	✓	✓
		Political	Politically or nationally relevant events, contexts, and/or climates	✓		✓		
		COVID-19	Explicit mentions of the COVID-19 global pandemic, quarantine, lockdown, etc.	✓	✓	✓	✓	✓
Well-being, health, and individual challenges	Well-being	*Physical health*	Physical injuries, bodily wellness, exercise, nutrition, sleep, etc.	✓		✓	✓	
		Mental health	Anxiety, depression, pressure and/or stress, etc.	✓	✓	✓	✓	✓
		Isolation	Being alone or lonely, including difficulties making friends in a course or connecting with others	✓	✓	✓	✓	✓
	Self-regulation	*Motivation*	Ability to focus on and finish a task, including references to procrastination and perceived lack of productivity	✓	✓	✓	✓	✓
		Time management	Ability to balance many competing responsibilities from classwork, jobs, family, personal lives, etc. within time constraints	✓	✓	✓	✓	✓
		Other	Challenges that were listed as "N/A" or "nothing", or that did not contain sufficient data to interpret	✓	✓	✓	✓	

Chapter 6

DISCUSSION & FUTURE WORK: ENABLING STAKEHOLDERS TO CONNECT DATA INTERPRETATIONS WITH DOMAIN EXPERTISE

In this dissertation, I conducted three design explorations to understand how stakeholders interpreting data may be able to support equity-oriented goals. To support more equitable self-directed online learning for learners of varying self-efficacy, I designed Codeitz to provide adaptive recommendations while also affording them agency over their own learning experiences (Chapter 3). To support more equitable assessment of learning, I explored how showing emperical evidence of potential test question bias to curriculum designers may inform changes to address bias by gender and race (Chapter 4). To support more equitable remote learning experiences, I explored how contextualized student feedback could inform teaching teams of which challenges disproportionately affected which sub-groups of students (Chapter 5). In Table 6.1, I summarized my findings from these studies and frame them relative to the three factors that affect interpretations of data for equity that I outlined in Section 2.4.2.

6.1 Discussion: Interpretations of study findings and future work

There are multiple ways to interpret the findings from these three studies as they relate to designing interactions to enable stakeholders to use data to support equity-oriented goals. A key theme across these interpretations is equity is a goal that is situated in a dynamic social context. Thus, I argue that we must embrace a plurality of approaches that are unified around a common commitment of problematizing the dominant social norms by questioning what we accept as normal in computing education and who we minoritize in the process. All this is in an effort to change dominant structures, systems, and discourses to imagine and work towards more equitable and just futures in computing education.

6.1.1 Limitation: WEIRD bias in studies is not representative of world (but is representative of computing norms)

One interpretation of these studies is that there are limitations related to the bias of the study populations. Indeed, these studies largely engage with Western, educated, industrialized, rich, and democratic (WEIRD, [132]) societies that are over-represented in much of academic research despite being some of the most psychologically unusual people on Earth. Recruitment for the Codeitz was from and around a top research university in the United States, resulting in most participants being undergraduates from dominant groups (white and Asian men). And while we did not consider geographic location in our analysis of DIF due to privacy concerns of minors, Code.org reports indicate that usage of its platforms disproportionately occurs in rich urban/industrialized areas in the Western world [131]. The StudentAmp study also occurred at a top research university in the United States. And all five instructors in the course were from dominant groups (white and Asian men).

Future work must explore less WEIRD contexts that are more representative of the global population (e.g. at Historically Black Colleges and Universities [81, 221], in rural areas [229, 215, 126]). But a dominant norm in computing is that much of it happens in WEIRD contexts, such as in Silicon Valley (CA), Seattle (WA), and other rich urban centers in the United States. The studies in this dissertation explore more equitable futures in computing education by considering incremental change from existing WEIRD norms.

6.1.2 Data is reductionist, but it is one of many tools for equity-oriented goals

Another interpretation of these findings is that the reductions and simplifications that come with data are harmful to equity-oriented goals. This is a common and valid critique amongst critical data scholars (e.g. [30, 159, 79]), situated on a history of using quantitative methods to justify and perpetuate minoritization and harm [92, 55]. While data and associated quantitative methods are dangerous, they may not be inherently oppressive. Prior work has adopted this more nuanced framing of data to use it for social justice through cultural competence in research teams and

activism-oriented action research [115, 55].

In this dissertation, I argue data can support equity-oriented goals when interactions with data promote open interpretations and connections with other forms of knowledge. The intention of the design of Codeitz was to provide recommendations to provide learners with the opportunity to consider it relative to other information they had. In the workshop with DIF, I scaffolded the experience to have curriculum designers consider what actions they may take or what information they were missing when they interpreted data on potential test bias. Similarly for StudentAmp, I asked teaching teams to explore the data on student feedback and demographics and consider it in relation with relevant prior knowledge related to teaching. Combined, these studies explore the use of data as one of many tools by putting it in the hands of stakeholders who have the domain expertise to interpret it amongst other tools. Future work may explore in how to scaffold connections of data interactions with other knowledge to promote equitable action and dissuade unproductive deficit framings of minoritized groups [285].

6.1.3 Having data non-experts interpret data is dangerous

Yet another interpretation of these findings is that asking stakeholders who are not necessarily experts in data analysis invites misinterpretation. Indeed, researchers and data scientists may have the expertise to interpret data that stakeholders may not. But they are often far from the context from which the data is intended to reflect, leaving them to often analyze the data in isolation. Alternatively, positioning stakeholders as the data interpreters enables them to leverage their domain expertise in socially situated manners that equity-oriented goals often require.

So it is the role of designers of interactions with data to scaffold experiences to support interpretations of data that are promote equity. By design, I refer to the creation of sociotechnical experiences of interpreting data in a situated context, extending beyond the design of a particular technical system that presents the data. For example, the design of the Codeitz system may have enabled agency to be possible, but the study design likely failed to consider how agency was unfamiliar to students. An improved design may have incorporated activities to teach students to consider the benefits of guiding their own learning experiences. In the DIF and StudentAmp

studies, I relied on researchers being scaffolds to guide conversations towards interpretations and conversations that were productive to equity-oriented goals. Equity is socially situated, so I argue the design of interactions with data must also consider the context in which interpretations occur as well as how stakeholders form their prior beliefs. Future work can explore the how differing sociotechical contexts in which situated stakeholders (e.g. teachers, students, policymakers, parents) affects their interpretions of data (e.g. [229]).

6.1.4 Designing interactions that position stakeholders as interpreters of data can support equitable change

A final interpretation of these findings is that we must design interactions with data that consider how stakeholders form prior beliefs that affect their interpretations. For these studies, I considered how stakeholders' cultural competence, relevant prior knowledge, and perceptions of power relationships affected their interpretations of data for equity. Table 6.1 summarizes how I interpreted study findings relative to these factors. Across these studies, we identified how stakeholders engaged with their prior knowledge when interpreting data. We also identified how perceptions of power relationships focused much of the interpretations on factors that they could control (e.g. curriculum and test design for curriculum designers, course structure for teaching teams). This suggests promising future work related to multiple stakeholders interpreting the "same" data, perhaps in collaboration. And while we did not foster the development of cultural competence, we saw how stakeholders with prior experiences relevant to cultural competence (e.g. coursework in public health, lived experience as part of minoritized groups) made interpretations richer by framing intepretations within a broader systemic context of minoritization.

These findings point to future work that explores how to engage with these factors to design for stakeholder interpretations of data that promote more breadth and depth in imagining equitable changes. Just as students are not empty vessels and bring with them prior knowledge when learning [32, 209], stakeholders interpreting data bring with them relevant prior knowledge. Sociotechnical designs that engage with this prior knowledge can ensure that prior knowledge provides context to the data interpretations. Stakeholders themselves are situated in a network of power relationships,

so we must consider their relationships within social systems and structures. By designing with power relationships in mind, we can better understand how different stakeholders come to different interpretations. And finally, measuring or even promoting cultural competence can enable stakeholders to consider data relative to broader structures of inequity.

6.2 Future work: Sharpening framework, shifting fundamental beliefs around data

Future work can explore how this framework can inform the design of interactions with data for equity-oriented goals in contexts beyond computing education as well as how to engage with data in theoretical framings that go beyond positivism.

6.2.1 Applying framework in new contexts and elaborating on it

Future work can explore how to apply and expand on the framework I proposed to inform future designs of interactions with data for equity-oriented goals. The framework I proposed (Fig. 2.2) was a result of design explorations that I conducted. Future can explore upon how to apply this framework in new contexts as well as how to expand upon it.

Future work can explore how this framework could apply to support equity-oriented goals in contexts beyond computing education. Computing education's challenges with inequities and minoritization of many groups also applies to other contexts, such as neighboring educational disciplines (e.g. science, math, engineering education) and perhaps other communities that engage with computing (e.g. technology companies, government agencies considering the use of technology).

Further application of this framework will come in tandem with further elaboration of this framework. Elaboration of this framework begins with further investigation of how to engage with each of the three factors. Relevant prior knowledge, perceptions of power relationships, and cultural competence differ by by context, time, and by positionality of different peoples. So to understand how to design for them, future work an explore participatory design and design justice approaches which not only engage stakeholders in the design process, but also partner with them to share discoveries and benefits with them throughout the process [68]. By doing so, we can focus

on equity-oriented goals, which emphasize the well-being of people above all.

6.2.2 Epistemological shifts: Data as one of many tools

If we are to use data to support equity and challenge existing systems of oppression, future work must also consider data under a theoretical framing that extends beyond positivism. Traditional statistical methods when used with a theoretical framing of positivism perpetuate norms of dominance and oppression [79]. Statistical methods have the most "power" when we accept reductions of the world to dichotomies and accept a regression to a common average. But these false dichotomies typically do not reflect the diversity of people, and the assumption that dominant norm reflects minoritized groups are often what equity-oriented actions must overcome.

My analysis of gender-based DIF data in Chapter 4 is an example of how considering data beyond a postivism framing provided more nuanced insights. Analysis of the student response data and incomplete self-reported demographics data identified test questions that positively favored non-binary students (see 4.4.4). The data-driven conclusion would be that non-binary students were positively biased when compared to reported male students. But when considered in the context of additional information on experiences of non-binary K-12 students (see 4.6), we come to a more nuanced understanding of how the data may not reflect the experiences of non-binary students.

Future work that shifts away from postivism could consider data from many funds of knowledge to provide more nuanced perspectives on lived experiences intersecting with systems of power. This can include diversifying data sources (e.g. photos, life history, campus maps [82, 274]), developing new methodologies to embody data with lived experiences [82, 10], and developing new theories that are more situated (e.g. how liquid modernity considers how the social world is always in flux [16]). The emerging field of QuantCrit [115, 55, 269, 270] is a promising shift in this direction, where data can be one of many tools used in equity-oriented movements.

6.3 Conclusion: Equity by designing for socially situated interpretations of data

Considering data for equity-oriented goals can be uncomfortable or uncertain to stakeholders with their own prior knowledge, power relationships, and cultural competence. A lack of consideration of factors such as these can result in interpretations of data that are not situated in a social context. These acontextual interpretations will often fall short of support equity-oriented goals that are socially situated. Therefore, I conclude by reiterating my thesis statement:

Interactions with data that consider prior knowledge, perceptions of power relationships, and cultural competency can enable computing education stakeholders to connect their interpretations of data with their domain expertise in service of equity-oriented goals.

Table 6.1: Summary of findings from across the three studies and their relationship to the three factors that affect interpretations of data for equity-oriented goals.

			relationship to factors that affect interpretations		
study	context	key finding	relevant prior knowledge	perceptions of power relations	cultural competence
Codeitz	self-directed online learning	agency had no effect on learning, perhaps because considering recommendations to exert agency was unfamiliar for university students	agency was unfamiliar compared to expectations of how to engage w/ learning environment		(did not consider)
DIF	assessment bias	curriculum designers were able to identify potential changes to curriculum and test by interpreting data and engaging curriculum design expertise	drew upon domain expertise w/ curriculum and test design when considering how to address bias	focused on what they could control (curriculum & test design)	varied by prior training and positionality; enabled broader consideration of bias
Student-Amp	student feedback	teaching teams were able to use student-reported challenges and demographic data to perspective take and consider how challenges could affect sub-groups differently	drew upon prior experience taking and teaching course, at institution	focused on changes within course, which felt limited	prior training enabled more consideration of student identity (e.g. ethnicity, gender)

BIBLIOGRAPHY

[1] Anne Adams, Peter Lunt, and Paul Cairns. A qualititative approach to HCI research. In Paul Cairns and Anna Cox, editors, *Research Methods for Human-Computer Interaction*, page A qualititative approach to HCI research. Cambridge University Press, 2008.

[2] June Ahn, Fabio Campos, Ha Nguyen, Maria Hays, and Jan Morrison. Co-Designing for privacy, transparency, and trust in K-12 learning analytics. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, LAK21, pages 55–65, New York, NY, USA, April 2021. Association for Computing Machinery.

[3] Elizabeth J Allan and Aaron R Tolbert. Advancing social justice with policy discourse analysis. In Kamden K Strunk and Leslie Ann Locke, editors, *Research Methods for Social Justice and Equity in Education*, pages 137–149. Springer International Publishing, Cham, 2019.

[4] Mary J Allen and Wendy M Yen. *Introduction to Measurement Theory*. Waveland Press, December 2001.

[5] Ammara. "f*ck the algorithm"; a rallying cry for the future. `https://medium.com/digital-diplomacy/fuck-the-algorithm-the-rallying-cry-of-our-youth-dd2677e190c`, August 2020. Accessed: 2021-10-3.

[6] Farshid Anvari, Deborah Richards, Michael Hitchens, and Hien Minh Thi Tran. Teaching user centered conceptual design using Cross-Cultural personas and peer reviews for a large cohort of students. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)*, pages 62–73, May 2019.

[7] Chris Argyris and Donald A Schön. *Organizational Learning: A Theory of Action Perspective*. Addison-Wesley Publishing Company, 1978.

[8] Chris Argyris and Donald A Schön. *Organizational Learning II: Theory, Method, and Practice*. Addison-Wesley Publishing Company, 1996.

[9] Richard C Atkinson. Optimizing the learning of a second-language vocabulary. *Journal of experimental psychology*, 96(1):124–129, November 1972.

[10] Lucy E Bailey. Thinking critically about "social justice methods": Methods as "contingent foundations". In Kamden K Strunk and Leslie Ann Locke, editors, *Research Methods for Social Justice and Equity in Education*, pages 91–107. Springer International Publishing, Cham, 2019.

[11] Ryan Baker and George Siemens. Educational data mining and learning analytics. In R Keith Sawyer, editor, *The Cambridge Handbook of the Learning Sciences*, pages 253–272. Cambridge University Press, Cambridge, 2 edition, 2014.

[12] Ryan S Baker. Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education*, 26(2):600–614, June 2016.

[13] Albert Bandura. *Self-Efficacy: The Exercise of Control*. Macmillan, February 1997.

[14] Albert Bandura. Social cognitive theory: An agentic perspective. *Annual review of psychology*, 52(1):1–26, 2001.

[15] Albert Bandura. Toward a psychology of human agency. *Perspectives on psychological science: a journal of the Association for Psychological Science*, 1(2):164–180, June 2006.

[16] Zygmunt Bauman. *Culture in a Liquid Modern World*. John Wiley & Sons, May 2013.

[17] Brian Beaton. How to respond to data science: Early data criticism by lionel trilling. *Information & Culture*, 51(3):352–372, August 2016.

[18] Aida Behmard. Feynman, harassment, and the culture of science. `https://caltechletters.org/viewpoints/feynman-harassment-science`, October 2019. Accessed: 2021-11-2.

[19] Michael Benitez, Jr. Resituating culture centers within a social justice framework. In Lori D Patton, editor, *Culture Centers in Higher Education: Perspectives on Identity, Theory, and Practice*, pages 119–134. Stylus Publishing, LLC., March 2010.

[20] Ruha Benjamin. Race after technology: Abolitionist tools for the new jim code. *Social forces; a scientific medium of social study and interpretation*, 98(4):1–3, June 2020.

[21] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, January 1995.

[22] Cynthia L Bennett and Daniela K Rosner. The promise of empathy. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, May 2019. ACM.

[23] Melanie Bertrand and Julie A Marsh. Teachers' sensemaking of data and implications for equity. *American Educational Research Journal*, 52(5):861–893, 2015.

[24] James R Bettman, Mary Frances Luce, and John W Payne. Constructive consumer choice processes. *The Journal of consumer research*, 25(3):187–217, December 1998.

[25] Sylvia Beyer. Why are women underrepresented in computer science? gender differences in stereotypes, self-efficacy, values, and interests and predictors of future CS course-taking and grades. *Computer Science Education*, 24(2-3):153–192, July 2014.

[26] Marie Bienkowski, Mingyu Feng, and Barbara Means. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. Technical report, U.S. Department of Education, 2012.

[27] Aikaterini Bourazeri and Simone Stumpf. Co-designing smart home technology with people with dementia or parkinson's disease. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*, NordiCHI '18, pages 609–621, New York, NY, USA, September 2018. Association for Computing Machinery.

[28] Pierre Bourdieu. *The forms of capital*. Routledge, 2018.

[29] Glenn A Bowen. Grounded theory and sensitizing concepts. *International Journal of Qualitative Methods*, 5(3):12–23, September 2006.

[30] Danah Boyd and Kate Crawford. CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication and Society*, 15(5):662–679, 2012.

[31] Shannon T Brady, Geoffrey L Cohen, Shoshana N Jarvis, and Gregory M Walton. A brief social-belonging intervention in college improves adult outcomes for black americans. *Science advances*, 6(18):eaay3689, May 2020.

[32] John D Bransford, Ann L Brown, and Rodney R Cocking, editors. *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. National Academy Press, 2000.

[33] Christopher Brooks, René F Kizilcec, and Nia Dowell. Designing inclusive learning environments. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*, L@S '20, pages 225–228, New York, NY, USA, August 2020. Association for Computing Machinery.

[34] Timothy A Brown. *Confirmatory Factor Analysis for Applied Research, Second Edition*. Guilford Publications, December 2014.

[35] Peter Bruce and Andrew Bruce. *Practical Statistics for Data Scientists: 50 Essential Concepts*. "O'Reilly Media, Inc.", May 2017.

[36] Emeline Brulé and Katta Spiel. Negotiating gender and disability identities in participatory design. In *Proceedings of the 9th International Conference on Communities & Technologies - Transforming Communities*, C&T '19, pages 218–227, New York, NY, USA, June 2019. Association for Computing Machinery.

[37] Joy Adowaa Buolamwini. *Gender shades : intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. PhD thesis, Massachusetts Institute of Technology, 2017.

[38] Kenneth M Burke. Distributed leadership and shared governance in post-secondary education. *Management in Education*, 24(2):51–54, April 2010.

[39] Margaret Burnett, Anicia Peters, Charles Hill, and Noha Elarief. Finding Gender-Inclusiveness software issues with GenderMag: A field investigation. pages 2586–2598. ACM Press, 2016.

[40] Sandra L Calvert, Bonnie L Strong, and Lizann Gallagher. Control as an engagement feature for young children's attention to and learning of computer content. *The American behavioral scientist*, 48(5):578–589, January 2005.

[41] Center for Educational Leadership. Creating a theory of action for improving teaching and learning. Technical report, University of Washington, 2014.

[42] R Philip Chalmers. mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, Articles*, 48(6):1–29, 2012.

[43] Hua-Hua Chang. Understanding computerized adaptive testing. *The Sage handbook of quantitative methods for the social sciences*, pages 117–133, 2004.

[44] Hua-Hua Chang. Psychometrics behind computerized adaptive testing. *Psychometrika*, 80(1):1–20, March 2015.

[45] Peter Checkland and Sue Holwell. Action research: Its nature and validity. *Systemic Practice and Action Research*, 11(1):9–21, February 1998.

[46] Xingliang Chen, Antonija Mitrovic, and Moffat Mathews. Investigating the effect of agency on learning from worked examples, erroneous examples and problem solving. *International Journal of Artificial Intelligence in Education*, 2019.

[47] Parmit K Chilana, Celena Alcock, Shruti Dembla, Anson Ho, Ada Hurst, Brett Armstrong, and Philip J Guo. Perceptions of non-CS majors in intro programming: The rise of the conversational programmer. In *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 251–259, October 2015.

[48] Sin Wang Chong. Reconsidering student feedback literacy from an ecological perspective. *Assessment & Evaluation in Higher Education*, 46(1):92–104, January 2021.

[49] Michael Clancy and Marcia C Linn. *Designing Pascal Solutions: A Case Study Approach*. Computer Science Press, 1992.

[50] Beth A Clark, Jaimie F Veale, Marria Townsend, Hélène Frohard-Dourlent, and Elizabeth Saewyc. Non-binary youth: Access to gender-affirming primary health care. *International Journal of Transgenderism*, 19(2):158–169, April 2018.

[51] Doug Clow. The learning analytics cycle: Closing the loop effectively. In *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 134–138, New York, NY, USA, 2012. ACM.

[52] Code.org Curriculum Team. CS discoveries 2019-2020. `https://curriculum.code.org/csd-19/`, 2019. Accessed: 2021-1-17.

[53] Derrick Coetzee, Armando Fox, Marti A Hearst, and Björn Hartmann. Should your MOOC forum use a reputation system? In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, CSCW '14, pages 1176–1187, New York, NY, USA, February 2014. Association for Computing Machinery.

[54] Ira J Cohen. Structuration theory and social praxis. In Ira J Cohen, editor, *Structuration Theory: Anthony Giddens and the Constitution of Social Life*, pages 9–55. Macmillan Education UK, London, 1989.

[55] Kevin Cokley and Germine H Awad. In defense of quantitative methods: Using the "master's tools" to promote social justice. *Journal for Social Action in Counseling and Psychology*, 5(2), 2013.

[56] James S Coleman. Social theory, social research, and a theory of action. *American Journal of Sociology*, 91(6):1309–1335, 1986.

[57] Latoya O Coleman, Philip Gibson, Shelia R Cotten, Michael Howell-Moroney, and Kristi Stringer. Integrating computing across the curriculum: The impact of internal barriers and training intensity on computer integration in the elementary school classroom. *Journal of Educational Computing Research*, 54(2):275–294, April 2016.

[58] Christina Convertino. Nuancing the discourse of underrepresentation: a feminist post-structural analysis of gender inequality in computer science education in the US. *Gender and education*, 32(5):594–607, July 2020.

[59] Alison Cook-Sather. Listening to equity-seeking perspectives: how students' experiences of pedagogical partnership can inform wider discussions of student success. *Higher education research & development*, 37(5):923–936, July 2018.

[60] Alison Cook-Sather. Respecting voices: how the co-creation of teaching and learning can support academic staff, underrepresented students, and equitable practices. *Higher Education*, 79(5):885–901, May 2020.

[61] Arthur H Copeland. A reasonable social welfare function. Seminar on Applications of Mathematics to Social Sciences, 1951.

[62] Gemma Corbalan, Liesbeth Kester, and Jeroen J G van Merriënboer. Selecting learning tasks: Effects of adaptation and shared control on learning efficiency and task involvement. *Contemporary educational psychology*, 33(4):733–756, 2008.

[63] Gemma Corbalan, Liesbeth Kester, and Jeroen J G van Merriënboer. Combining shared control with variability over surface features: Effects on transfer test performance and task involvement. *Computers in human behavior*, 25(2):290–298, 2009.

[64] Gemma Corbalan, Liesbeth Kester, and Jeroen J G van Merriënboer. Learner-controlled selection of tasks with different surface and structural features: Effects on transfer and efficiency. *Computers in human behavior*, 27(1):76–81, January 2011.

[65] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, December 1994.

[66] Diana I Cordova and Mark R Lepper. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of educational psychology*, 1996.

[67] Lindsay L Cornelius and Leslie Rupert Herrenkohl. Power in the classroom: How the classroom environment shapes students' relationships with each other and with concepts. *Cognition and instruction*, 22(4):467–498, December 2004.

[68] Sasha Costanza-Chock. *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press, March 2020.

[69] National Research Council. *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*. The National Academies Press, Washington, DC, 2012.

[70] Kimberlé Williams Crenshaw. Mapping the margins: Intersectionality, identity politics, and violence against women of color. In Martha Albertson Fineman and Rixanne Mykitiuk, editors, *The Public Nature of Private Violence*, pages 93–118. Routledge, 1994.

[71] Daniela R Crişan, Jorge N Tendeiro, and Rob R Meijer. Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied psychological measurement*, 41(6):439–455, September 2017.

[72] Terry L Cross, Barbara J Bazron, Karl W Dennis, and Mareasa R Isaacs. Towards a culturally competent system of care: A monograph on effective services for minority children who are severely emotionally disturbed. Technical report, Georgetown University, March 1989.

[73] Karol Danutama and Inggriani Liem. Scalable autograder and LMS integration. *Procedia Technology*, 11:388–395, January 2013.

[74] Ali Darvishi, Hassan Khosravi, and Shazia Sadiq. Employing peer review to evaluate the quality of student generated content at scale: A trust propagation approach. In *Proceedings of the Eighth ACM Conference on Learning @ Scale*, L@S '21, pages 139–150, New York, NY, USA, June 2021. Association for Computing Machinery.

[75] Yossi Ben David, Avi Segal, and Ya'akov (kobi) Gal. Sequencing educational content in classrooms using bayesian knowledge tracing. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 354–363. ACM, April 2016.

[76] Matt J Davidson, Brett Wortzman, Min Li, and Amy J Ko. Investigating item bias in a CS1 exam with differential item functioning. In *Proceedings of the ACM Technical Symposium on Computer Science Education (SIGCSE), Research Track*. ACM, 2021.

[77] Mark H Davis. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113–126, 1983.

[78] R J De Ayala. *The theory and practice of item response theory*. Methodology in the social sciences. Guilford Press, New York, 2009.

[79] Alain Desrosières. *The Politics of Large Numbers: A History of Statistical Reasoning*. Harvard University Press, 2002.

[80] R D Dietz, R H Pearson, M R Semak, C W Willis, N Sanjay Rebello, Paula V Engelhardt, and Chandralekha Singh. Gender bias in the force concept inventory? AIP, 2012.

[81] Betsy DiSalvo, Mark Guzdial, Charles Meadows, Ken Perry, Tom McKlin, and Amy Bruckman. Workifying games: successfully engaging african american gamers with computer science. In *Proceeding of the 44th ACM technical symposium on Computer science education*, SIGCSE '13, pages 317–322, New York, NY, USA, March 2013. Association for Computing Machinery.

[82] E J Dixon-Román, J L Jackson, and others. Reconceptualizing the Quantitative-Qualitative divide: Toward a new empiricism. *Handbook of the*, 2020.

[83] Neil J Doran. Tests as contests, overview of NCME fairness volume, and a welcome challenge, 2017.

[84] Neil J Dorans. Contributions to the quantitative assessment of item, test, and score fairness. In *Advancing human assessment*, pages 201–230. Springer, Cham, 2017.

[85] Neil J Dorans and Edward Kulick. Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement*, 23(4):355–368, 1986.

[86] Remy Dou, Karina Bhutta, Monique Ross, Laird Kramer, and Vishodana Thamotharan. The effects of computer science stereotypes and interest on middle school boys' career intentions. *ACM Transactions on Computing Education*, 20(3):1–15, June 2020.

[87] Steven P Dow. *Understanding User Engagement in Immersive and Interactive Stories*. PhD thesis, Georgia Institute of Technology, 2008.

[88] Ruth Dunn. *Minority Studies*. LibreTexts, 2021.

[89] Carol S Dweck. *Mindset: The New Psychology of Success*. Ballantine Books, 2008.

[90] R Edward Freeman. *Strategic Management: A Stakeholder Approach*. Cambridge University Press, March 2010.

[91] Eric Eide, Dominic J Brewer, and Ronald G Ehrenberg. Does it pay to attend an elite private college? evidence on the effects of undergraduate college quality on graduate school attendance. *Economics of education review*, 17(4):371–376, October 1998.

[92] Hamid Ekbia, Michael Mattioli, Inna Kouper, G Arave, Ali Ghazinejad, Timothy Bowman, Venkata Ratandeep Suri, Andrew Tsou, Scott Weingart, and Cassidy R Sugimoto. Big data, bigger dilemmas: A critical review: Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*, 66(8):1523–1545, 2015.

[93] David L Eng and Shinhee Han. A dialogue on racial melancholia. *Psychoanalytic dialogues*, 10(4):667–700, August 2000.

[94] Barbara Ericson and Mark Guzdial. Measuring demographics and performance in computer science education at a nationwide scale using AP CS data. In *Proceedings of the 45th ACM technical symposium on Computer science education*, SIGCSE '14, pages 217–222, New York, NY, USA, March 2014. Association for Computing Machinery.

[95] Karl Anders Ericsson and Herbert Alexander Simon. *Protocol Analysis: Verbal Reports as Data Revised Edition*. The MIT Press, 1993.

[96] Martha Escobar, Jeff Gray, Kathleen Haynie, Mohammed A Qazi, Yasmeen Rawajfih, Pamela McClendon, Donnita Tucker, and Wendy Johnson. Engaging black female students in a Year-Long preparatory experience for AP CS principles. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, SIGCSE '21, pages 706–724, New York, NY, USA, March 2021. Association for Computing Machinery.

[97] Y Fan, L J Shepherd, E Slavich, D Waters, M Stone, R Abel, and E L Johnston. Gender and cultural bias in student evaluations: Why representation matters. *PloS one*, 14(2):e0209749, February 2019.

[98] Ronald F Ferguson. *Toward Excellence with Equity: An Emerging Vision for Closing the Achievement Gap*. Harvard Education Press, December 2007.

[99] Tracy Ferne and André A Rupp. A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language assessment quarterly*, 4(2):113–148, July 2007.

[100] Rob Filback and Alan Green. New directions for diversity at USC rossier. *Futures in Urban Ed, the Magazine of the USC Rossier School of Education*, August 2013.

[101] Sally A Fincher and Anthony V Robins. *The Cambridge Handbook of Computing Education Research*. Cambridge University Press, February 2019.

[102] Christian Fischer, Zachary A Pardos, Ryan Shaun Baker, Joseph Jay Williams, Padhraic Smyth, Renzhe Yu, Stefan Slater, Rachel Baker, and Mark Warschauer. Mining big data in education: Affordances and challenges. *Review of Research in Education*, 44(1):130–160, March 2020.

[103] Gail E FitzSimons. A framework for evaluating quality and equity in Post-Compulsory mathematics education. In Bill Atweh, Mellony Graven, Walter Secada, and Paola Valero, editors, *Mapping Equity and Quality in Mathematics Education*, pages 105–121. Springer Netherlands, Dordrecht, 2011.

[104] Julie Flapan, Jean J Ryoo, and Roxana Hadad. Building systemic capacity to scale and sustain equity in computer science through Multi-Stakeholder professional development. In *Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*. IEEE, 2020.

[105] Carol L Fletcher and Jayce R Warner. CAPE: a framework for assessing equity throughout the computer science education ecosystem. *Communications of the ACM*, 64(2):23–25, January 2021.

[106] Terri Flowerday and Gregory Schraw. Teacher beliefs about instructional choice: A phenomenological study. *Journal of educational psychology*, 92(4):634–645, December 2000.

[107] Michel Foucault. *The Foucault Reader*. Pantheon Books, 1984.

[108] Floyd J Fowler, Jr. and Thomas W Mangione. *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. SAGE, 1990.

[109] Peter Francis, Christine Broughan, Carly Foster, and Caroline Wilson. Thinking critically about learning analytics, student outcomes, and equity of attainment. *Assessment & Evaluation in Higher Education*, pages 1–11, December 2019.

[110] Mark Freeman, Paul Blayney, and Paul Ginns. Anonymity and in class learning: The case for electronic response systems. *Australasian Journal of Educational Technology*, 22(4), November 2006.

[111] Peter A Frensch, Axel Buchner, and Jennifer Lin. Implicit learning of unique and ambiguous serial transitions in the presence and absence of a distractor task. *Journal of experimental psychology. Learning, memory, and cognition*, 20(3):567–584, 1994.

[112] Batya Friedman and David G Hendry. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press, May 2019.

[113] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information and System Security*, 14(3):330–347, July 1996.

[114] John P Fry. Interactive relationship between inquisitiveness and student control of instruction. *Journal of educational psychology*, 63(5):459–465, October 1972.

[115] Nichole M Garcia, Nancy López, and Verónica N Vélez. QuantCrit: rectifying quantitative methods through critical race theory. *Race Ethnicity and Education*, 21(2):149–157, March 2018.

[116] Brie Gertler. Self-Knowledge. `https://plato.stanford.edu/archives/spr2020/entries/self-knowledge/`, 2020. Accessed: 2021-10-13.

[117] Google Inc. and Gallup Inc. Images of computer science: Perceptions among students, parents and educators in the U.S. Technical report, Gallup, 2015.

[118] Google Inc. and Gallup Inc. Diversity gaps in computer science: Exploring the underrepresentation of girls, blacks and hispanics. Technical report, Gallup, 2016.

[119] Jamie Gorson and Eleanor O'Rourke. Why do CS1 students think they're bad at programming? investigating self-efficacy and self-assessments at three universities. In *Proceedings of the 2020 ACM Conference on International Computing Education Research*, ICER '20, pages 170–181, New York, NY, USA, August 2020. Association for Computing Machinery.

[120] Emily Gould and Rachael Swell. Managing labels. `https://docs.github.com/en/github/managing-your-work-on-github/managing-labels`, 2011. Accessed: 2021-4-2.

[121] Colin Gray, Seda Yilmaz, Shanna Daly, Colleen Seifert, and Richard Gonzalez. Idea generation through empathy: Reimagining the 'cognitive walkthrough'. In *2015 ASEE Annual Conference and Exposition Proceedings*, Industrial Design Conference Presentations, Posters and Proceedings. ASEE Conferences, 2015.

[122] A G Greenwald and M R Banaji. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4–27, January 1995.

[123] Pamela Grimm. Social desirability bias. In Jagdish Sheth and Naresh Malhotra, editors, *Wiley International Encyclopedia of Marketing*, volume 50, page 537. John Wiley & Sons, Ltd, Chichester, UK, December 2010.

[124] Philip Guo. Silent technical privilege. *Slate*, 2014.

[125] David Hammer and Leema K Berland. Confusing claims for data: A critique of common practices for presenting qualitative research on learning. *Journal of the Learning Sciences*, 23(1):37–46, January 2014.

[126] Jean Hardy, Susan Wyche, and Tiffany Veinot. Rural HCI research: Definitions, distinctions, methods, and opportunities. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW):1–33, November 2019.

[127] Erik Harpstead, J Elizabeth Richey, Huy Nguyen, and Bruce M McLaren. Exploring the subtleties of agency and indirect control in digital learning games. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, LAK '19. ACM, 2019.

[128] William J Hawkins, Neil T Heffernan, and Ryan S J D Baker. Learning bayesian knowledge tracing parameters with a knowledge heuristic and empirical probabilities. In *Intelligent Tutoring Systems*, pages 150–155. Springer International Publishing, 2014.

[129] John F Helliwell and Robert D Putnam. Education and social capital. May 1999.

[130] Kay Hempsall. Developing leadership in higher education: perspectives from the USA, the UK and australia. *Journal of Higher Education Policy and Management*, 36(4):383–394, July 2014.

[131] Katie Hendrickson, Liz Gauthier, Maggie Osorio Glennon, Alexis Menocal Harrigan, Hannah Weissman, Carol Fletcher, Sarah Dunton, Jake Baskin, and Janice Mak. 2021 state of computer science: Empowering action through advocacy. Technical report, Code.org, 2021.

[132] Joseph Henrich, Steven J Heine, and Ara Norenzayan. Most people are not WEIRD. *Nature*, 466(7302):29, July 2010.

[133] Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *The Behavioral and brain sciences*, 33(2-3):61–83; discussion 83–135, June 2010.

[134] An-Li Herring. Residents raise concerns about bias in allegheny county's automated decision-making tools. *WITF*, March 2020.

[135] Charles G Hill, Maren Haag, Alannah Oleson, Chris Mendez, Nicola Marsden, Anita Sarma, and Margaret Burnett. Gender-Inclusiveness personas vs. stereotyping: Can we have it both ways? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 6658–6671, New York, NY, USA, May 2017. Association for Computing Machinery.

[136] Anna Lauren Hoffmann. Data violence and how bad engineering choices can damage society. `https://medium.com/s/story/data-violence-and-how-bad-engineering-choices-can-damage-society-39e44150e1d4`, April 2018. Accessed: 2021-7-14.

[137] Anna Lauren Hoffmann. Terms of inclusion: Data, discourse, violence. *New Media & Society*, page 1461444820958725, September 2020.

[138] Paul W Holland, Howard Wainer, and Educational Testing Service. *Differential Item Functioning*. Psychology Press, 1993.

[139] Jeffrey D Holmes. The bad Test-Taker identity. *Teaching of psychology*, page 0098628320979884, December 2020.

[140] Meredith I Honig, Michael A Copland, Lydia Rainey, Juli Anna Lorton, and Morena Newton. Central office transformation for district-wide teaching and learning improvement. Technical report, University of Washington, 2010.

[141] M Horvath, A M Ryan, and S L Stierwalt. The influence of explanations for selection test use, outcome favorability, and Self-Efficacy on Test-Taker perceptions. *Organizational behavior and human decision processes*, 83(2):310–330, November 2000.

[142] Roya Hosseini, I-Han Hsiao, Julio Guerra, and Peter Brusilovsky. What should I do next? adaptive sequencing in the context of open social student modeling. In *Design for Teaching and Learning in a Networked World*, pages 155–168. Springer International Publishing, 2015.

[143] Aleata Hubbard Cheuoua. Confronting inequities in computer science education: A case for critical theory. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, SIGCSE '21, pages 425–430, New York, NY, USA, March 2021. Association for Computing Machinery.

[144] David Hutchison, Takeo Kanade, Josef Kittler, Jon M Kleinberg, Friedemann Mattern, John C Mitchell, Moni Naor, Oscar Nierstrasz, C Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y Vardi, and Gerhard Weikum. *Intelligent Tutoring Systems*, volume 3220 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[145] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, New York, NY, 8 edition, 2013.

[146] Magdalena Jara and Harvey Mellar. Quality enhancement for e-learning courses: The role of student feedback. *Computers & education*, 54(3):709–714, April 2010.

[147] Ben Jee, Jennifer Wiley, and Thomas Griffin. Expertise and the illusion of comprehension. In *Proceedings of the Annual Conference of the Cognitive Science Society*, pages 387–392, 2006.

[148] Michael Kane. Articulating a validity argument. In *The Routledge Handbook of Language Testing*. Routledge, November 2009.

[149] Michael Kane. Validity and fairness. *Language Testing*, 27(2):177–182, April 2010.

[150] Michael T Kane. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73, March 2013.

[151] Jussi Kasurinen and Uolevi Nikula. Estimating programming knowledge with bayesian knowledge tracing. In *Proceedings of the 14th Annual ACM SIGCSE Conference on Innovation and Technology in Computer Science Education*, ITiCSE '09, pages 313–317, New York, NY, USA, 2009. ACM.

[152] Stephen Kemmis. Participatory action research and the public sphere. *Educational Action Research*, 14(4):459–476, December 2006.

[153] Stephen Kemmis and Lindsay Fitzclarence. *Curriculum theorising: Beyond reproduction theory*. UNSW Press, 1986.

[154] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[155] Maurice G Kendall. *Rank Correlation Methods*. Griffin, 1970.

[156] Mohammad M Khajah, Yun Huang, José P González-Brenes, Michael C Mozer, and Peter Brusilovsky. Integrating knowledge tracing and item response theory: A tale of two frameworks. In *CEUR Workshop Proceedings*, volume 1181, pages 7–15. University of Pittsburgh, January 2014.

[157] Daniel B King and Shivani Joshi. Gender differences in the use and effectiveness of personal response devices. *Journal of science education and technology*, 17(6):544–552, December 2008.

[158] Päivi Kinnunen and Beth Simon. CS majors' self-efficacy perceptions in CS1: Results in light of social cognitive theory. In *Proceedings of the Seventh International Workshop on Computing Education Research*, ICER '11, pages 19–26, New York, NY, USA, 2011. ACM.

[159] Rob Kitchin. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):205395171452848, July 2014.

[160] René F Kizilcec and Andrew J Saltarelli. Can a diversity statement increase diversity in MOOCs? In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*, number Article 2 in L@S '19, pages 1–8, New York, NY, USA, June 2019. Association for Computing Machinery.

[161] Michael S Knapp, Michael A Copland, Meredith I Honig, Margaret L Plecki, and Bradley S Portin. Learning-Focused leadership and leadership support: Meaning and practice in urban systems. Technical report, University of Washington, 2010.

[162] Amy J Ko. Computing education research FAQ. `https://faculty.washington.edu/ajko/cer`, 2021. Accessed: 2021-5-20.

[163] Amy J Ko, Alannah Oleson, Neil Ryan, Yim Register, Benjamin Xie, Mina Tari, Matthew Davidson, Stefania Druga, and Dastyni Loksa. It is time for more critical CS education. *Communications of the ACM*, 63(11):31–33, November 2020.

[164] Mary Ellen Kondrat. Actor-Centered social work: Re-visioning "Person-in-Environment" through a critical theory lens. *The Social worker*, 47(4):435–448, October 2002.

[165] Sean Kross and Philip J Guo. Students, systems, and interactions: synthesizing the first four years of learning@scale and charting the future. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, number Article 2 in L@S '18, pages 1–10, New York, NY, USA, June 2018. Association for Computing Machinery.

[166] Chinmay Kulkarni. Design perspectives of learning at scale: Scaling efficiency and empowerment. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*, L@S '19, pages 18:1–18:11, New York, NY, USA, 2019. ACM.

[167] Chinmay Kulkarni, Julia Cambre, Yasmine Kotturi, Michael S Bernstein, and Scott R Klemmer. Talkabout: Making distance matter with small groups in massive classes. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1116–1128, New York, NY, USA, February 2015. Association for Computing Machinery.

[168] Harrison Kwik, Benjamin Xie, and Amy J Ko. Experiences of computer science transfer students. In *Proceedings of the 2018 ACM Conference on International Computing Education Research*, ICER '18, pages 115–123. ACM Press, 2018.

[169] Charles E Lance, Marcus M Butts, and Lawrence C Michels. The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, 9(2):202–220, 2006.

[170] Sarah Levine, Danielle Keifert, Ananda Marin, and Noel Enyedy. Hybrid argumentation in literature and science for K–12 classrooms. In *Handbook of the cultural foundations of learning*, pages 141–159. Routledge, 2020.

[171] Catherine Lewis. What is improvement science? do we need it in education? *Educational researcher*, 44(1):54–61, January 2015.

[172] Colleen M Lewis, Niral Shah, and Katrina Falkner. Equity and diversity. In Sally A Fincher and Anthony V Robins, editors, *The Cambridge Handbook of Computing Education Research*, pages 481–510. Cambridge University Press, 2019.

[173] Sarah Lewthwaite and David Sloan. Exploring pedagogical culture for accessibility education in computing science. In *Proceedings of the 13th International Web for All Conference*, number Article 3 in W4A '16, pages 1–4, New York, NY, USA, April 2016. Association for Computing Machinery.

[174] Julie Libarkin. Concept inventories in higher education science. In *National Research Council Promising Practices in Undergraduate STEM Education Workshop*, volume 13, page 14, 2008.

[175] Sarah Lichtenstein and Paul Slovic. *The Construction of Preference*. Cambridge University Press, August 2006.

[176] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. Facing the cold start problem in recommender systems. *Expert systems with applications*, 41(4, Part 2):2065–2073, March 2014.

[177] Alex Lishinski, Aman Yadav, Jon Good, and Richard Enbody. Learning to program: Gender differences and interactive effects of students' motivation, goals, and Self-Efficacy on performance. In *Proceedings of the 2016 ACM Conference on International Computing Education Research*, ICER '16, pages 211–220, New York, NY, USA, 2016. ACM.

[178] Samuel A Livingston and Michael J Zieky. *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Educational Testing Service, 1982.

[179] Samuel A Livingston and Michael J Zieky. A comparative study of Standard-Setting methods. *Applied Measurement in Education*, 2(2):121–141, April 1989.

[180] Dastyni Loksa, Benjamin Xie, Harrison Kwik, and Amy J Ko. Investigating novices' in situ reflections on their programming process. In *Proceedings of the ACM Technical Symposium on Computer Science Education (SIGCSE), Research Track*. ACM, 2020.

[181] Yanjin Long and Vincent Aleven. Supporting students' Self-Regulated learning with an open learner model in a linear equation tutor. In *Artificial Intelligence in Education*, Lecture Notes in Computer Science, pages 219–228. Springer, Berlin, Heidelberg, July 2013.

[182] Yanjin Long and Vincent Aleven. Gamification of joint Student/System control over problem selection in a linear equation tutor. In *Intelligent Tutoring Systems*, pages 378–387. Springer International Publishing, 2014.

[183] Stephanie Lunn, Leila Zahedi, Monique Ross, and Matthew Ohland. Exploration of intersectionality and computer science demographics: Understanding the historical context of shifts in participation. *ACM Trans. Comput. Educ.*, 21(2):1–30, March 2021.

[184] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM transactions on knowledge discovery from data*, 1(1):3–es, March 2007.

[185] Monika Mandl, Alexander Felfernig, Erich Teppan, and Monika Schubert. Consumer decision making in knowledge-based recommendation. *Journal of intelligent information systems*, 37(1):1–22, August 2011.

[186] Martin N Marger. *Race and Ethnic Relations: American and Global Perspectives, 10th Edition*. Cengage, 2015.

[187] Jane Margolis. Unlocking the clubhouse: a decade later and now what? In *Proceeding of the 44th ACM technical symposium on Computer science education*, SIGCSE '13, pages 9–10, New York, NY, USA, March 2013. Association for Computing Machinery.

[188] Herbert W Marsh and Lawrence A Roche. Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist Association*, 52(11), January 1997.

[189] James D Marshall. Michel foucault: Educational research as problematisation. In *Why Foucault?: New Directions in Educational Research*, pages 15–28. Peter Lang, 2007.

[190] Patrícia Martinková, Adéla Drabinová, Yuan-Ling Liaw, Elizabeth A Sanders, Jenny L McFarland, and Rebecca M Price. Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *Cell biology education*, 16(2):rm2, 2017.

[191] Viktor Mayer-Schoenberger and K Cukier. The rise of big data: How it's changing the way we think about the world. *Foreign affairs*, 92(3):28–40, 2013.

[192] Robert McCartney, Jonas Boustedt, Anna Eckerdal, Kate Sanders, and Carol Zander. Folk pedagogy and the geek gene: Geekiness quotient. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, SIGCSE '17, pages 405–410, New York, NY, USA, March 2017. Association for Computing Machinery.

[193] Nora McDonald, Karla Badillo-Urquiola, Morgan G Ames, Nicola Dell, Elizabeth Keneski, Manya Sleeper, and Pamela J Wisniewski. Privacy and power: Acknowledging the importance of privacy research and design for vulnerable populations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–8, New York, NY, USA, April 2020. Association for Computing Machinery.

[194] Nora McDonald and Andrea Forte. The politics of privacy theories: Moving from norms to vulnerabilities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '20. ACM, 2020.

[195] Peter McLaren. *Critical Pedagogy: A Look at the Major Concepts*. Routledge/Falmer Press, 2002.

[196] Adam W Meade. A taxonomy of effect size measures for the differential functioning of items and scales. *The Journal of applied psychology*, 95(4):728–743, July 2010.

[197] Paola Medel and Vahab Pournaghshband. Eliminating gender bias in computer science education materials. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education - SIGCSE '17*, pages 411–416, Seattle, Washington, USA, 2017. ACM Press.

[198] C Mendez. The InclusiveMag method: A start towards more inclusive software for diverse populations. 2020.

[199] Christopher Mendez, Lara Letaw, Margaret Burnett, Simone Stumpf, Anita Sarma, and Claudia Hilderbrand. From GenderMag to InclusiveMag: An inclusive design Meta-Method. In *2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 97–106. IEEE, October 2019.

[200] Sarah Mercer. The complexity of learner agency. *Apples - Journal of Applied Language Studies*, 2012.

[201] Janet Metcalfe, Teal S Eich, and David B Miele. Metacognition of agency: proximal action and distal outcome. *Experimental brain research. Experimentelle Hirnforschung. Experimentation cerebrale*, 229(3):485–496, September 2013.

[202] Janet Metcalfe and Herbert S Terrace. *Agency and Joint Attention*. OUP USA, September 2013.

[203] Daniel C Moos and Roger Azevedo. Self-regulated learning with hypermedia: The role of prior domain knowledge. *Contemporary educational psychology*, 33(2):270–298, April 2008.

[204] Margaret E Morris, Kevin S Kuehn, Jennifer Brown, Paula S Nurius, Han Zhang, Yasaman S Sefidgar, Xuhai Xu, Eve A Riskin, Anind K Dey, Sunny Consolvo, and Jennifer C Mankoff. College from home during COVID-19: A mixed-methods study of heterogeneous experiences. *PloS one*, 16(6):e0251580, June 2021.

[205] Laurie Murphy, Sue Fitzgerald, Raymond Lister, and Renée McCauley. Ability to 'explain in plain english' linked to proficiency in computer-based programming. In *Proceedings of the Ninth Annual International Conference on International Computing Education Research*, ICER '12, pages 111–118, New York, NY, USA, 2012. ACM.

[206] Christof Nachtigall, Ulf Kröhne, Ulrike Enders, and Rolf Steyer. Causal effects and fair comparison: Considering the influence of context variables on student competencies. In Johannes Hartig, Eckhard Klieme, and Detlev Leutner, editors, *Assessment of Competencies in Educational Contexts*, pages 315–336. Hogrefe Publishing, October 2008.

[207] Kevin L Nadal. *The SAGE Encyclopedia of Psychology and Gender*. SAGE Publications, April 2017.

[208] National Academies of Sciences and Medicine. *Assessing and Responding to the Growth of Computer Science Undergraduate Enrollments*. The National Academies Press, Washington, DC, 2018.

[209] National Academies of Sciences, Engineering, and Medicine. *How People Learn II: Learners, Contexts, and Cultures*. National Academies Press, Washington, D.C., September 2018.

[210] Timothy Neate, Aikaterini Bourazeri, Abi Roper, Simone Stumpf, and Stephanie Wilson. Co-Created personas: Engaging and empowering users with diverse needs within the design process. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12. Association for Computing Machinery, New York, NY, USA, May 2019.

[211] Thomas O Nelson, John Dunlosky, Aurora Graf, and Louis Narens. Utilization of metacognitive judgments in the allocation of study during multitrial learning. *American Psychological Society*, 5(4), 1994.

[212] Huy Nguyen, Erik Harpstead, Yeyu Wang, and Bruce M McLaren. Student agency and Game-Based learning: A study comparing low and high agency. In *Artificial Intelligence in Education*, pages 338–351. Springer International Publishing, 2018.

[213] NPR. With 'paper towns,' author john green reopens search for agloe, N.Y. *NPR*, July 2015.

[214] Jum C Nunnally. *Psychometric theory*. McGraw-Hill series in psychology. McGraw-Hill, New York, 2d ed. edition, 1978.

[215] Jaclyn Ocumpaugh, Ryan Baker, Sujith Gowda, Neil Heffernan, and Cristina Heffernan. Population validity for educational data mining models: A case study in affect detection. *British journal of educational technology: journal of the Council for Educational Technology*, 45(3):487–501, May 2014.

[216] Tor Ole B Odden and Rosemary S Russ. Defining sensemaking: Bringing clarity to a fragmented theoretical construct. *Science education*, 103(1):187–205, January 2019.

[217] Andy Olesko. Man falsely arrested because of facial recognition software error sues detroit. *Courthouse News Service*, April 2021.

[218] Alannah Oleson, Christopher Mendez, Zoe Steine-Hanson, Claudia Hilderbrand, Christopher Perdriau, Margaret Burnett, and Amy J Ko. Pedagogical content knowledge for teaching inclusive design. In *Proceedings of the 2018 ACM Conference on International Computing Education Research*, ICER '18, pages 69–77, New York, NY, USA, August 2018. Association for Computing Machinery.

[219] Alannah Oleson, Meron Solomon, and Amy J Ko. Computing students' learning difficulties in HCI education. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–14, New York, NY, USA, April 2020. Association for Computing Machinery.

[220] Mark Olssen * and Michael A Peters. Neoliberalism, higher education and the knowledge economy: from the free market to knowledge capitalism. *Journal of Education Policy*, 20(3):313–345, January 2005.

[221] Robert Palmer and Marybeth Gasman. "it takes a village to raise a child": The role of social capital in promoting academic success for african american men at a black college. *Journal of college student development*, 49(1):52–70, January 2008.

[222] Abelardo Pardo and George Siemens. Ethical and privacy principles for learning analytics. *British journal of educational technology: journal of the Council for Educational Technology*, 45(3):438–450, May 2014.

[223] Zachary A Pardos and Neil T Heffernan. KT-IDEM: Introducing item difficulty to the knowledge tracing model. In *User Modeling, Adaption and Personalization*, pages 243–254. Springer Berlin Heidelberg, 2011.

[224] Miranda C Parker, Mark Guzdial, and Shelly Engleman. Replication, validation, and use of a language independent CS1 knowledge assessment. In *Proceedings of the 2016 ACM Conference on International Computing Education Research*, ICER '16, pages 93–101, New York, NY, USA, 2016. ACM.

[225] Harold Pashler, Patrice M Bain, Brian A Bottge, Arthur Graeser, Kenneth Koedinger, Mark McDaniel, and Janet Metcalfe. Organizing instruction and study to improve student learning. Technical Report NCER 2007-2004, U.S. Department of Education, 2007.

[226] Elizabeth Patitsas, Jesse Berlin, Michelle Craig, and Steve Easterbrook. Evidence that computer science grades are not bimodal. In *Proceedings of the 2016 ACM Conference on International Computing Education Research*, ICER '16, pages 113–121, New York, NY, USA, 2016. ACM.

[227] Michael Quinn Patton. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*. SAGE Publications, November 2014.

[228] John W Payne, James R Bettman, and Eric J Johnson. *The Adaptive Decision Maker*. Cambridge University Press, May 1993.

[229] Evan M Peck, Sofia E Ayuso, and Omar El-Etr. Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12. Association for Computing Machinery, New York, NY, USA, May 2019.

[230] Radek Pelánek. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User modeling and user-adapted interaction*, 27(3-5):313–350, 2017.

[231] Heather E Price. Large-Scale datasets and social justice: Measuring inequality in opportunities to learn. In Kamden K Strunk and Leslie Ann Locke, editors, *Research Methods for Social Justice and Equity in Education*, pages 203–215. Springer International Publishing, Cham, 2019.

[232] Paul Prinsloo and Sharon Slade. Student vulnerability, agency and learning analytics: An exploration. *Journal of Learning Analytics*, 3(1):159–182, April 2016.

[233] Cynthia Putnam, Maria Dahman, Emma Rose, Jinghui Cheng, and Glenn Bradford. Teaching accessibility, learning empathy. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '15, pages 333–334, New York, NY, USA, October 2015. Association for Computing Machinery.

[234] Stephen M Quintana, A Wade Boykin, Andrew Fuligni, Sandra Graham, Samuel Ortiz, and Frank C Worrell. Ethnic and racial disparities in education: Psychology's contributions to understanding and reducing disparities. Technical report, American Psychological Association, 2012.

[235] Vennila Ramalingam, Deborah LaBelle, and Susan Wiedenbeck. Self-efficacy and mental models in learning to program. In *Proceedings of the 9th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education*, ITiCSE '04, pages 171–175, New York, NY, USA, 2004. ACM.

[236] Vennila Ramalingam and Susan Wiedenbeck. Development and validation of scores on a computer programming Self-Efficacy scale and group analyses of novice programmer Self-Efficacy. *Journal of Educational Computing Research*, 19(4):367–381, 1998.

[237] Yolanda A Rankin, Jakita O Thomas, and Sheena Erete. Real talk: Saturated sites of violence in CS education. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, SIGCSE '21, pages 802–808, New York, NY, USA, March 2021. Association for Computing Machinery.

[238] Justin Reich. *Failure to Disrupt: Why Technology Alone Can't Transform Education*. Harvard University Press, 2020.

[239] Justin Reich and Mizuko Ito. From good intentions to real outcomes: Equity by design in learning technologies. Technical report, Digital Media and Learning Research Hub, 2017.

[240] William Revelle. psych: Procedures for psychological, psychometric, and personality research. `https://CRAN.R-project.org/package=psych`, December 2020. Accessed: 2021-2-4.

[241] John T E Richardson. Instruments for obtaining student feedback: a review of the literature. *Assessment & Evaluation in Higher Education*, 30(4):387–415, August 2005.

[242] Jessica Roberts and Leilah Lyons. Examining spontaneous perspective taking and fluid Self-to-Data relationships in informal Open-Ended data exploration. *Journal of the Learning Sciences*, 29(1):32–56, January 2020.

[243] Judy Robertson and Maurits Kaptein. *Modern Statistical Methods for HCI*. Springer, 2016.

[244] Kevin Robinson, Keyarash Jahanian, and Justin Reich. Using online practice spaces to investigate challenges in enacting principles of equitable computer science teaching. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, SIGCSE '18, pages 882–887, New York, NY, USA, February 2018. Association for Computing Machinery.

[245] Kimberley Rogers, Isabel Dziobek, Jason Hassenstab, Oliver T Wolf, and Antonio Convit. Who cares? revisiting empathy in asperger syndrome. *Journal of autism and developmental disorders*, 37(4):709–715, April 2007.

[246] Ido Roll, Eliane Stampfer Wiese, Yanjin Long, Vincent Aleven, and Kenneth R Koedinger. Tutoring self-and co-regulation with intelligent tutoring systems to help students acquire better learning skills. *Design recommendations for intelligent tutoring systems*, 2:169–182, 2014.

[247] Karen Rosenblum and Toni-Michelle Travis. *The Meaning of Difference: American Constructions of Race and Ethnicity, Sex and Gender, Social Class, Sexuality, and Disability*. February 2015.

[248] Monique Ross, Zahra Hazari, Gerhard Sonnert, and Philip Sadler. The intersection of being black and being a woman: Examining the effect of social computing relationships on computer science career choice. *ACM Trans. Comput. Educ.*, 20(2):1–15, February 2020.

[249] Yves Rosseel. lavaan: An R package for structural equation modeling. *Journal of Statistical Software, Articles*, 48(2):1–36, 2012.

[250] Jonathan P Rowe, Lucy R Shores, Bradford W Mott, James C Lester, and North Carolina. Integrating learning, problem solving, and engagement in Narrative-Centered learning environments. *International Journal of Artificial Intelligence in Education*, 21:115–133, 2011.

[251] Richard M Ryan, C Scott Rigby, and Andrew Przybylski. The motivational pull of video games: A Self-Determination theory approach. *Motivation and emotion*, 30(4):344–360, December 2006.

[252] Donald G Saari and Vincent R Merlin. The copeland method. *Economic theory*, 8(1):51–76, February 1996.

[253] Robert Sawyer, Andy Smith, Jonathan Rowe, Roger Azevedo, and James Lester. Is more agency better? the impact of student agency on Game-Based learning. In *Artificial Intelligence in Education*, pages 335–346. Springer International Publishing, 2017.

[254] Jesse Schell. *The Art of Game Design: A Book of Lenses, Second Edition*. A K Peters/CRC Press, 2014.

[255] Toni Schmader and Michael Johns. Converging evidence that stereotype threat reduces working memory capacity. *Journal of personality and social psychology*, 85(3):440–452, 2003.

[256] Toni Schmader, Michael Johns, and Chad Forbes. An integrated process model of stereotype threat effects on performance. *Psychological review*, 115(2):336–356, April 2008.

[257] Kimberly A Scott, Kimberly M Sheridan, and Kevin Clark. Culturally responsive computing: a theory revisited. *Learning, media and technology*, 40(4):412–436, October 2015.

[258] Kimberly A Scott and Mary Aleta White. COMPUGIRLS' standpoint: Culturally responsive computing and its effect on girls of color. *Urban Education*, 48(5):657–681, September 2013.

[259] Niral Shah and Colleen M Lewis. Amplifying and attenuating inequity in collaborative learning: Toward an analytical framework. *Cognition and instruction*, 37(4):423–452, October 2019.

[260] Simone G Shamay-Tsoory, Judith Aharon-Peretz, and Daniella Perry. Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain: a journal of neurology*, 132(Pt 3):617–627, March 2009.

[261] George Siemens and Ryan S J d Baker. Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 252–254, New York, NY, USA, April 2012. Association for Computing Machinery.

[262] Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118, February 1955.

[263] Sharon Slade and Paul Prinsloo. Learning analytics: Ethical issues and dilemmas. *The American behavioral scientist*, 57(10):1510–1529, October 2013.

[264] Suzanne L Slocum-Gori and Bruno D Zumbo. Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social indicators research*, 102(3):443–461, 2011.

[265] Erica L Snow, Laura K Allen, Matthew E Jacovina, and Danielle S McNamara. Does agency matter?: Exploring the impact of controlled behaviors within a game-based environment. *Computers & education*, 82:378–392, 2015.

[266] Oddny Judith Solheim. The impact of reading Self-Efficacy and task value on reading comprehension scores in different item formats. *Reading psychology*, 32(1):1–27, January 2011.

[267] Katta Spiel, Oliver Haimson, and Danielle Lottridge. How to do better with gender on surveys: A guide for HCI researchers. *ACM Interactions*, 26(4), 2019.

[268] James Spillane and David Miele. Evidence in practice: A framing of the terrain. *Teachers College record*, 109(13):46–73, 2007.

[269] Frances K Stage. Answering critical questions using quantitative data. *New directions for institutional research*, 2007(133):5–16, 2007.

[270] Frances K Stage and Ryan S Wells. Critical quantitative inquiry in context. *New directions for institutional research*, 2013(158):1–7, June 2014.

[271] Philip Stark, Richard Freishtat, and Carol Lauer. An evaluation of course evaluations. *ScienceOpen Research*, September 2014.

[272] Claude Steele. Stereotype threat and African-American student achievement. In *The Inequality Reader*, pages 276–281. Routledge, 2 edition, 2011.

[273] Kamden K Strunk and Jasmine S Betties. Using critical theory in educational research. In Kamden K Strunk and Leslie Ann Locke, editors, *Research Methods for Social Justice and Equity in Education*, pages 71–79. Springer International Publishing, Cham, 2019.

[274] Kamden K Strunk and Leslie Ann Locke, editors. *Research Methods for Social Justice and Equity in Education*. Palgrave Macmillan, 2019.

[275] S Stumpf, A Peters, S Bardzell, M Burnett, D Busse, J Cauchard, and E Churchill. Gender-Inclusive HCI research and design: A conceptual review. *Foundations and Trends in Human–Computer Interaction*, 13(1):1–69, March 2020.

[276] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5), 2002.

[277] Huib K Tabbers and Bastiaan de Koeijer. Learner control in animated multimedia instructions. *Instructional Science*, 38(5):441–453, September 2010.

[278] Burçin Tamer and Jane Stout. Recruitment and retention of undergraduate students in computing: Patterns by gender and Race/Ethnicity. Technical report, Computing Research Association, 2016.

[279] Michelle Taub, Robert Sawyer, Andy Smith, Jonathan Rowe, Roger Azevedo, and James Lester. The agency effect: The impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. *Computers & education*, 147:103781, April 2020.

[280] Heather Thiry and Sarah T Hug. Sustaining student engagement and equity in computing departments during the COVID-19 pandemic. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, SIGCSE '21, pages 987–993, New York, NY, USA, March 2021. Association for Computing Machinery.

[281] Phil Turner and Susan Turner. Is stereotyping inevitable when designing with personas? *Design Studies*, 32(1):30–44, January 2011.

[282] Suraj Uttamchandani. Equity in the learning sciences: Recent themes and pathways. In *13th International Conference of the Learning Sciences (ICLS)*. International Society of the Learning Sciences (ISLS), 2018.

[283] Sepehr Vakil. Ethics, identity, and political vision: Toward a Justice-Centered approach to equity in computer science education. *Harvard educational review*, 88(1):26–52, March 2018.

[284] Sepehr Vakil. "i've always been scared that someday i'm going to sell out": Exploring the relationship between political identity and learning in computer science education. *Cognition and instruction*, 38(2):87–115, April 2020.

[285] Richard R Valencia. *The Evolution of Deficit Thinking: Educational Thought and Practice*. Routledge, November 2012.

[286] José Van Dijck. Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & society*, 12(2):197, 2014.

[287] András Vargha and Harold D Delaney. A critique and improvement of the CL common language effect size statistics of McGraw and wong. *Journal of educational and behavioral statistics: a quarterly publication sponsored by the American Educational Research Association and the American Statistical Association*, 25(2):101–132, June 2000.

[288] Nanette Veilleux, Rebecca Bates, Cheryl Allendoerfer, Diane Jones, Joyous Crawford, and Tamara Floyd Smith. The relationship between belonging and ability in computer science. In *Proceeding of the 44th ACM technical symposium on Computer science education*, SIGCSE '13, pages 65–70, New York, NY, USA, March 2013. Association for Computing Machinery.

[289] Ana María Villegas and Tamara Lucas. Preparing culturally responsive teachers: Rethinking the curriculum. *Journal of teacher education*, 53(1):20–32, January 2002.

[290] Cindy M Walker. What's the DIF? why differential item functioning analyses are an important part of instrument development and validation. *Journal of psychoeducational assessment*, 29(4):364–376, August 2011.

[291] David A Walker. JMASM9: converting kendall's tau for correlational or meta-analytic analyses. *Journal of modern applied statistical methods: JMASM*, 2003.

[292] Erin Walker, Nikol Rummel, and Kenneth R Koedinger. Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity. *International Journal of Computer-Supported Collaborative Learning*, 6(2):279–306, 2011.

[293] Margaret Walsh, Crystal Hickey, and Jim Duffy. Influence of item content and stereotype situation on gender differences in mathematical problem solving. *Sex roles*, 41(3-4):219–240, August 1999.

[294] Gregory M Walton and Geoffrey L Cohen. A question of belonging: race, social fit, and achievement. *Journal of personality and social psychology*, 92(1):82–96, January 2007.

[295] Noah Wardrip-Fruin, Michael Mateas, Steven Dow, and Serdar Sali. Agency reconsidered. *DiGRA Conference*, 2009.

[296] Jayce R Warner, Joshua Childs, Carol L Fletcher, Nicole D Martin, and Michelle Kennedy. Quantifying disparities in computing education: Access, participation, and intersectionality. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, SIGCSE '21, pages 619–625, New York, NY, USA, March 2021. Association for Computing Machinery.

[297] Alicia Nicki Washington. When twice as good isn't enough: The case for cultural competence in computing. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, SIGCSE '20, pages 213–219, New York, NY, USA, February 2020. Association for Computing Machinery.

[298] William Watson and Sunnie Lee Watson. An argument for clarity: What are learning management systems, what are they not, and what should they become. *TechTrends*, 2007.

[299] Max Weber. *From Max Weber: Essays in Sociology*, volume 33. Routledge, 1948.

[300] S Christian Wheeler and Richard E Petty. The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological bulletin*, 127(6):797–826, November 2001.

[301] Tiffani L Williams. 'underrepresented minority' considered harmful, racist language. *Communications of the ACM*, June 2020.

[302] Carol M Woods. Evaluation of MIMIC-Model methods for DIF testing with comparison to Two-Group analysis. *Multivariate behavioral research*, 44(1):1–27, January 2009.

[303] Carol M Woods, Li Cai, and Mian Wang. The Langer-Improved wald test for DIF testing with multiple groups: Evaluation and comparison to Two-Group IRT. *Educational and psychological measurement*, 73(3):532–547, June 2013.

[304] Beverly Park Woolf. *Building intelligent interactive tutors: student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann Publishers/Elsevier, Amsterdam ; Boston, 2009.

[305] Benjamin Xie. How data can support equity in computing education. *XRDS: Crossroads, The ACM Magazine for Students*, 27(2):48–52, December 2020.

[306] Benjamin Xie, Matt J Davidson, Baker Franke, Emily McLeod, Min Li, and Amy J Ko. Domain experts' interpretations of assessment bias in a scaled, online computer science curriculum. In *Proceedings of the Eight ACM Conference on Learning @ Scale*, volume 29 of *L@S 2021*. ACM, 2021.

[307] Benjamin Xie, Matthew J Davidson, Min Li, and Amy J Ko. An item response theory evaluation of a Language-Independent CS1 knowledge assessment. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, SIGCSE '19, pages 699–705. ACM, 2019.

[308] Benjamin Xie, Dastyni Loksa, Greg L Nelson, Matthew J Davidson, Dongsheng Dong, Harrison Kwik, Alex Hui Tan, Leanne Hwa, Min Li, and Amy J Ko. A theory of instruction for introductory programming skills. *Computer Science Education*, pages 1–49, January 2019.

[309] Benjamin Xie, Greg L Nelson, Harshitha Akkaraju, William Kwok, and Amy J Ko. The effect of informing agency in Self-Directed online learning environments. In *Proceedings of the Seventh (2020) ACM Conference on Learning @ Scale*, L@S 2020, pages 77–89. ACM, 2020.

[310] Benjamin Xie, Greg L Nelson, and Amy J Ko. An explicit strategy to scaffold novice program tracing. In *2018 ACM SIGCSE Technical Symposium on Computer Science Education*, SIGCSE '18, New York, NY, USA, 2018. ACM.

[311] Benjamin Xie, Alannah Oleson, Jayne Everson, and Amy J Ko. Surfacing equity issues in large computing courses with Peer-Ranked, Demographically-Labeled student feedback. *Proceedings of the ACM on Human-Computer Interaction*, 2022. Accepted to CSCW 2022.

[312] Saijing Zheng, Mary Beth Rosson, Patrick C Shih, and John M Carroll. Understanding student motivation, behaviors and perceptions in MOOCs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1882–1895, New York, NY, USA, February 2015. Association for Computing Machinery.

[313] Michael Zieky. Practical questions in the use of DIF statistics in test development. In Paul W Holland and Howard Wainer, editors, *Differential Item Functioning*, pages 337–347. Erlbaum, 1993.

[314] Michael Zieky. A DIF primer. Technical report, Educational Testing Service, 2003.

[315] Daniel Zingaro. Peer instruction contributes to self-efficacy in CS1. In *Proceedings of the 45th ACM technical symposium on Computer science education*, pages 373–378. ACM, March 2014.

[316] Tukufu Zuberi. *Thicker Than Blood: How Racial Statistics Lie*. U of Minnesota Press, 2001.

[317] Bruno D Zumbo. Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly*, 4(2):223–233, 2007.

[318] Bruno D Zumbo and Michaela N Gelin. A matter of test bias in educational policy research: Bringing the context into picture by investigating Sociological/Community moderated (or mediated) test and item bias. *Journal of Educational Research & Policy Studies*, 5(1):1–23, 2005.

[319] Stuart Zweben and Betsy Bizot. Taulbee survey. Technical report, Computing Research Association, 2019.

Appendix A

SUPPLEMENTAL INFORMATION FOR CODEITZ STUDY

A.1 Post-test with solutions

A.1.1 Q1

Write down all values printed as output after this code runs.

```
x = 2
y = 5
z = 3

if (y % x == 1):
    print("a")
    x = x * x;
elif (y % x == 2):
    print("b")
    z = z * z
else:
    print("c")
    y = y * y
if (y / x == 1):
    print("g")
    x = x + 3
else:
    y = y * 2
```

```
    print("h")

print(x)
print(y)
print(z)
```

Answer:

```
a
h
4
10
3
```

Scoring: 4 points maximum

- 1 pt for first 2 lines (-0.5 for each incorrect, additional, or missing line)
- 3 pts for last 3 lines
  - -1 pt if var name included w/ correct number (e.g. x = 2)
  - -1 total if lines begin with "x=", "y=", "z="
  - -0.5 for each additional line
- -1 total if no new lines
- -0 if quotes around strings
- -1 total if quotes around numbers

Justification: This exercise assesses knowledge of variable updates, conditional

A.1.2 Q2

What is printed as a result of this code segment?

```
name = "james"
time = "night"

print("hi")
if(time != "day" and name == "Alice"):
    print("hi alice")
elif(time != "day"):
    print("hello")
    print("name")
else:
    print("good day to you")
print("done")
```

Answer:

```
hi
hello
name
done
```

Scoring: 2.5 points maximum

- 0.5 pt for each line except 3rd (1.5 total)
- 1 if "name" line missing, 0.5 if "james"
- -0.25 lines 1 and 2 merged ("good night name"), -0.5 if 1 and 2 merged with var name ("good night james")

- -1 (total) if lines have additional info on them (e.g. ‘output = "hello"‘)

Notes:

- Expect 3rd line to be common error (variable vs literal).
- No points off if new lines missing?
- Never actually asks to print var value

A.1.3 Q3

For the next three questions (3A-3C), consider the following code. The code below assumes that the variables a, b, and c all store numbers (integers or floats).

```
x = -1
y = -1

if(a < b and a < c):
    print(1)
    x = a
elif(b < c):
    print(2)
    x = b
else:
    print(3)
    x = c

if(a > b and a > c):
    print(4)
    y = a
```

```
elif(b > c):
    print(5)
    y = b
else:
    print(6)
    y = c

val = y - x

if(val > 0):
    print("THE VALUE:")
    print(val)
```

Given the variable values a = 1.1, b = 5, c = 2, determine the output of the code and write the output below:

Answer:

```
1
5
THE VALUE:
3.9
```

Scoring: 3 points maximum

- -0.5 pt each for first 3 lines (-0.25 for each additional line, e.g. var update)
- 1.5 pts for last line correct (-0.5 if "THE VALUE:" and "3.9" on same line, so 1.5 total for "THE VALUE: 3.9")
- -0.5 for each line > 4

In the box below, summarize in plain English what the code does.

Example answer:

```
Finds difference between max and min of 3 numbers.
```

Scoring: 3 points maximum

- 2 pt for mentioning max/min (1 pt for max, 1 pt for min)
- 1 pt for mentioning finding difference between max and min
- -50% if describing code line by line (e.g. mentioning every variable specifically)

If 'OUTPUT:' was not printed, what is the relationship between the variables a, b, and c?

Example answer:

```
Output not printed when a, b, c all <= 0 or all equal to each other.
```

Scoring: 3 points maximum

- 1.5 pts for saying output not printed when a, b, c all <= 0 (-0.25 if they say < 0)
- 1.5 pts for saying output not printed when a, b, and c all same (a==b, b==c)
- 1 pt if only mention when conditional false

A.1.4 Q4

The code below assumes that the variables *a*, *b*, and *c* all store integers.

```
x = a%2==0
y = b%2==0
z = c%2==0
```

```
u = 0

if(x):
  u = u + 1
if(y):
  u = u + 1
if(z):
  u = u + 1

print(u)
```

Given the variable values a = -2, b = 3, c = 4, determine the output of the code and write the output in the box below:

Answer:

```
2
```

Scoring: 2 points maximum

- 2 pts for right answer (-1 pt if write "u" or "u =")
- -0.5 for each additional line (0 pts total if last line does not have "2" in it)

In the box below, summarize in plain English what the code does.

Example answer:

```
Prints the number of even numbers stored in variables
```

Scoring: 2 points maximum

- 2 points if mentions "even" (or "divisible by 2")
  - -1 if say "number of variables where

  - -1.5 if just mentioning conditional ("if statement is true")

- 1 pt for mentioning "counting" or "how many"

  - 0.5 pt for mentioning updating ‘u‘

- Only 1 if only mentions conditionals or boolean ("how many statements are true")

- -50% if describing code line by line (e.g. mentioning every variable specifically)

What would variables a, b, and c have to be for 0 to be printed?

Example answer:

```
a,b, and c would have to all be odd numbers.
```

Scoring: 2 points maximum

Scoring: 2 points maximum

- -0.5 if don’t mention even/odd or divisible by 2 and instead say "when a,b, and c %2 does not equal 0"

- -1-1.5 if says only subset of variables must be odd

- -1 if additional constraint added (e.g. a, b, and c must be absolute value, positive)

- -1 if think it must be even number or when a, b, and c %2 !=0

- -1.5 if only mentions when u equals 0

- -1.5 if only mentions if statements (e.g. "when all if statements invalid")

- -1.5 if give specific valid example (e.g. a = 1, b = 3, c = 5)

A.1.5 Q5

Two friends regularly play chess against each other and they want to keep track of who was the last person to the win and how many previous games in a row they won. To do so, they ask you write some code to help them.

Predefined Variables:

- Four variables have already been defined:
- The variable leader has the name of the person who won the previous game(s).
- The variable follower contains the name of the person who lost the previous game.
- The variable current_streak contains the number of consecutive games that have been won by leader.
- The variable winner contains the name of the person who just won a game.

Code Instructions:

They ask you to write code to do the following:

1. If winner is equal to follower, then there is a new champion.
   (a) Swap the names stored in leader and follower to reflect this change.
   (b) Reset current_streak to 0.
   (c) Print "new leader"
2. If winner is equal to leader, then the person who won the previous game has won another one
   (a) Update current_streak by adding 1 to the previous value.
   (b) Print "same leader"

3. If winner is not equal to follower or leader, then there is an unknown player.

    (a) Print "unknown player"

Example Execution

Here are a few examples of what how the code would execute:

- If the variable winner was set to "Luca" and the variable follower was also set to "Luca", the values stored in leader and follower would swap, current_streak would be set to 0, and "new leader" would be printed.

- If the variable winner was set to "Abby", the variable leader was also set to "Abby", and the variable current_streak were set to 4, then current_streak would be updated to 5 and "same leader" would be printed.

- If the variable winner was set to "Kim", the variable leader was set to "Juan", and the variable follower were set to "Olaf", then "unknown player" would be printed.

Solution:

```
if winner==follower:
  follower = leader
  leader = winner
  current_streak = 0
  print("new leader")

elif winner == leader:
  current_streak = current_streak + 1
  print("same leader")
```

```
else:
  print("unknown player")
```

Scoring: 4

- 2.5 for if condition (-2 if swap wrong)
  - -0.5 if updates winner w/ correct swap
  - -1 if swap uses temp var but still wrong
  - -1.5 if swap w/o 3rd var
- 1 for elif
  - -0.5 if current_streak not updated correctly
- 0.5 for else
  - -0.25 if condition added to it (ok if elif with condition; not grading condition for logical correctness)
- -0 if uses 3 if statements (technically incorrect code, but instructions were unclear)
- -0.25-0.5 for minor syntax errors (logic is correct, code may need small adjustment to run correctly).
  - E.g. single = for equality check (-0.5 if done everywhere, -0.25 if only done once)
- -0.25-0.5 for minor syntax errors (logic is correct, code may need small adjustment to run correctly).
  - E.g. multiple wrong variable names used, quotes around var names (-1 if all vars)

Notes

- "Winner" and "leader" being different vars is confusing
- Logic error where if use multiple if statements (if 1st condition true, 2nd condition also true b/c of var update). That's ok b/c question miswritten.
- Focus of question around swap being used correctly as well as conditionals used effectively

A.1.6 Q6

Say you and 2 friends (a total of 3 people) split a bill. The amounts each of you paid are decimal numbers stored in the variables amt1, amt2, and amt3. You want to determine if you paid within 0.000001 (1e-6) bitcoin of the bill. The cost of the meal is stored in the decimal variable cost. You are worried that you may have underpaid or overpaid.

Write code that determines if you and your friends properly paid for the bill.

- If in total you all paid at least 0.000001 less than the cost, your code should print "underpaid" and then the amount that you underpaid on the next line.
- If in total you all paid within 0.000001 of the cost, your code should print "paid in full".
- If in total you all paid at least 0.000001 more than the cost, your code should print "overpaid" and then the amount you all overpaid on the following line.

In example, say

```
amt1 = 0.001111, and
amt2 = 0.002222, and
amt3 = 0.000033, and
cost = 0.003368.
```

The output of the code would be:

```
underpaid
0.000002
```

Solution

```
paid = amt1 + amt2 + amt3
thres = 0.0001

diff = paid - cost

if diff < 0 and abs(diff) > thres:
    print("underpaid")
    print(abs(diff))
elif abs(diff) < thres:
    print("paid in full")
else:
    print("overpaid")
    print(abs(diff))
```

Scoring: 6 points maximum

- 1 point for total paid
- 1 point for difference between cost and sum of amounts paid
- 1.5 point for 3 conditionals
  - 0.5 for having 3 conditions
  - 1 pt for having conditional statements relating to float equality
- 2 point for float equality check in conditionals w/ threshold, abs value.

  - 1 pt for correct math operation
  - 0.5 pt for threshold value
  - 0.5 for abs value function (or equivalent behavior)
  - -1 for each incorrect w/ major error (logic, major syntax).
  - -0.5 for each w/ minor syntax error

- 0.5 point for correct print statements (-0.25 if values not printed)

A.1.7 Q7

Write code that determines if the variable inp, a 4 digit integer value (between 1000-9999), is a valid passcode. inp is a valid passcode if the sum of the first 3 digits modulus 7 is equal to the last digit. If inp is valid, the code should print 'valid'. If the string is not valid, it should print 'NOT valid'.

So if inp were set to 5312, it would be a valid passcode and your code would print valid because the first 3 digits (5, 3, and 1) sum to 9 and 9 modulus 7 equals the last digit (2). 1234 would not be a valid passcode and your code would print NOT valid. Write your solution in the box below.

Assume a variable inp has already been declared and stores a 4 digit integer value (between 1000-9999).

Solution:

```
digit_4 = inp % 10
inp = inp // 10
digit_3 = inp % 10
inp = inp // 10
digit_2 = inp % 10
inp = inp // 10
digit_1 = inp %10
```

```
sum_3 = digit_1 + digit_2 + digit_3

if(sum_3 % 7 == digit_4):
    print("valid")
else:
    print("NOT valid")
```

Scoring: 7 points maximum

- 3 pts for digit processing
  - -0.5 to -1.5 if inp not truncated properly
  - -0.5 to -1.5 if digits not saved properly
- 1 pt for using // instead of /
  - Ok if use ‘int()‘ (even though we didn’t teach that. . .)
- 1 point for summing digits properly
- 1.5 point for conditional % 7 (-0.5 if 2 ifs used)
  - -1 if no %7
  - -0.5 if not comparing to 4th digit
- 0.5 point for writing correct print statements

Appendix B

SUPPLEMENTAL INFORMATION FOR DIF STUDY

B.1 Questions asked during workshop with curriculum designers

Questions asked prior to study:

1. How can instructors and/or students benefit from using assessments in Code.org?

2. For Code.org, what are challenges to designing an equitable learning experience?

Post-survey

1. What was beneficial about reviewing the data on DIF (if anything)?

2. What was difficult or challenging about reviewing the data on DIF (if anything)?

3. How could you see yourself (or Code.org more broadly) using data related to identifying unfairness?

4. Anything else you want to share?

5. How long have you worked as a curriculum manager at Code.org? (select one: less than 1 year, 1-2 years, 3-5 years, 5+ years, I'm not sure, (prefer not to disclose))

6. What work have you done with the CS Discoveries curriculum (if any)?

7. What is your gender? (select 1 or many: woman, man, non-binary, (prefer not to disclose), custom response)

8. What is your ethnicity? (select 1 or many: Asian, Black/African, Hispanic/Latinx, Native American, Pacific Islander, white, (prefer not to disclose), custom response)

9. Could a member of the research team follow-up with you? (select 1: yes, maybe, no)

10. Could a member of the research team follow-up with you? (select 1: yes, maybe, no)

Appendix C

SUPPLEMENTAL INFORMATION FOR STUDENTAMP STUDY

C.1 Prompt for Theory of Action

This prompt was followed as part of interviews with teaching teams. Adapted from [41].

C.1.1 Process

1. Develop a well-elaborated conception of the problem or situation for students and teachers that motivates their actions in the first place

2. Make leadership the core of the theory of action

3. Create evidence-based rationale for all parts of the theory

4. Identify the supports needed to make the identified changes in principal practice.

C.1.2 Steps

1. First Glance

    (a) What's going on with our students' learning experiences?

        i. Impressions and observations

        ii. What needs to change?

        iii. so that students who are _ will be able to _.

    (b) How is teachers' instruction affecting student learning? What are teachers doing (or not doing) in their instruction that's helping or hindering students' performance?

i. Impressions and observations

ii. What needs to change?

iii. If instructors _, then the course will change by _.

(c) What factors external to the course are helping or hindering students' learning?

(d) If instructors _ , then the course will change by _, so that students who are _ will be able to _.

2. Grounding initial ideas with theory of action

(a) What's going on with our students' learning?

i. Description: Which responses best captures what most concerns us about students' learning experiences? What specifically is hindering for that student?

ii. Evidence: What evidence do we have that substantiates our concerns? (within current Student Amp session, in previous sessions, from outside of Student Amp e.g. conversations with students, observations from class)

iii. What needs to change? (What aspects of student learning do we need to work on? Why are we prioritizing these particular aspects of students learning as issues? Which students would benefit from this? Could be harmed from this?)

iv. What changes in teacher practice or other instructional resources do we think will make a difference in student learning?

(b) How is teachers' instruction affecting student learning? What are teachers doing (or not doing) in their instruction that's helping or hindering students' performance?

i. Description: What is an aspect of course instruction that is an issue for students' learning experiences? What specifically concerns us? Who is affected?

ii. Evidence: What additional evidence do we have or could we collect to substantiate the problem and how it's affecting student performance?

iii. What aspects of instructional practice do we need to work on to improve student learning?

iv. Why are prioritizing these particular practices as issues?

v. What specifically do teachers need to do differently?

vi. What makes us think that teachers changing practice in these ways will improve students' learning experiences?

vii. Which students would benefit from this? Could be harmed from this?

viii. What support and/or system changes will teachers need to successfully make these changes?

3. Context

   (a) What factors external to the course are helping or hindering students' learning?

      i. What aspects external factors would need to change on to support better teaching?

      ii. Why are prioritizing these particular factors?

      iii. Who would need to get involved to address these issues?

      iv. What makes us think that getting these stakeholders involved will improve teaching?

      v. What supports and/or systemic changes will be needed to successfully make these changes?

4. If instructors _, then the course will change by _, so that students who are _ will be able to _.

5. 3 Putting it all Together (Focus on what needs to change)

   (a) What information are we missing?

   (b) How will we use our theory of action? Who do we need to engage in dialog with about our theory of action and why?

(c) What are the most important things we need to convey to these audiences about our theory of action & the need for change? In what ways do we need their support?