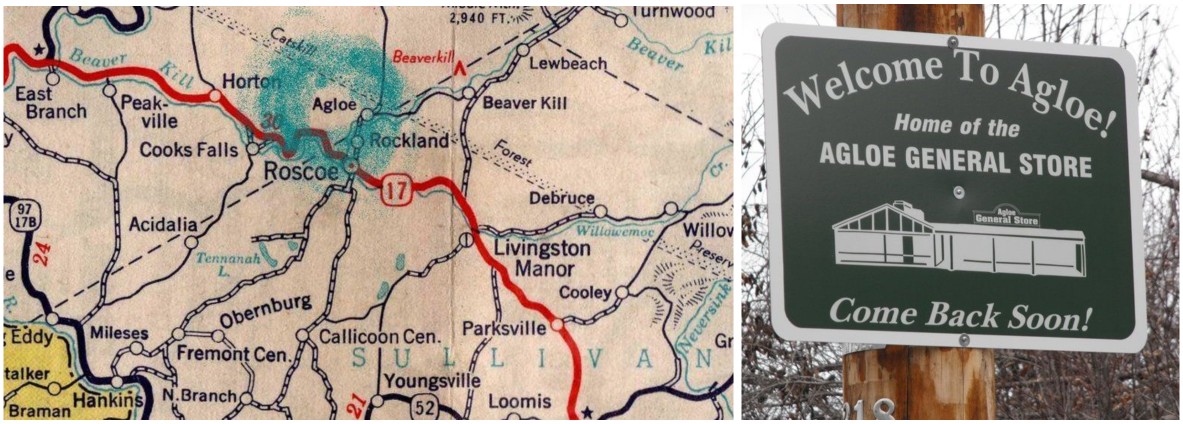
Xie, Benjamin. “How Data Can Support Equity in Computing Education.” *XRDS* 27, no. 2 (December 14, 2020): 48–52. https://doi.org/10.1145/3433136

How Data Can Support Equity in Computing Education

Benjamin Xie (he/him)
University of Washington, Seattle
bxie@uw.edu
@benjixie

Data has historically been a tool of oppression. But if we consider how its interpretations and uses affect minoritized groups, data-driven tools could support diversity, equity, and inclusion in computing education and beyond.



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. But then it became a reality. (*Booklist/ American Library, Joyce Conroy*)

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. To prevent competing companies from copying their maps, they created the fictitious place of "Agloe". 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 famous 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. 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 [1]. 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 and trans people from being considered in government decision-making; facial recognition datasets are predominantly of white men, resulting in diminished classification accuracy for darker skin and the false arrest of a Black man in Detroit; 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 of colors.

Within the context of computer science (CS) 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, intelligent tutoring systems that use data from other students' performance to personalize and adapt learning experiences, and auto-graders to make evaluating assessments more efficient. 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). 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.

As part of my doctoral studies at the University of Washington Seattle, I research how data affect students in minoritized groups and how we can design interactions with data for more equitable learning. I frame equity as access to and successful participation in education set in the context of economic, social, cultural, and political considerations of a time and place. That is to say that differences in learning outcomes should not be attributed to differences of identity, wealth, income, power, or possessions. Equity often involves a corrective measure to adjust for aggregate historical social inequities [8].

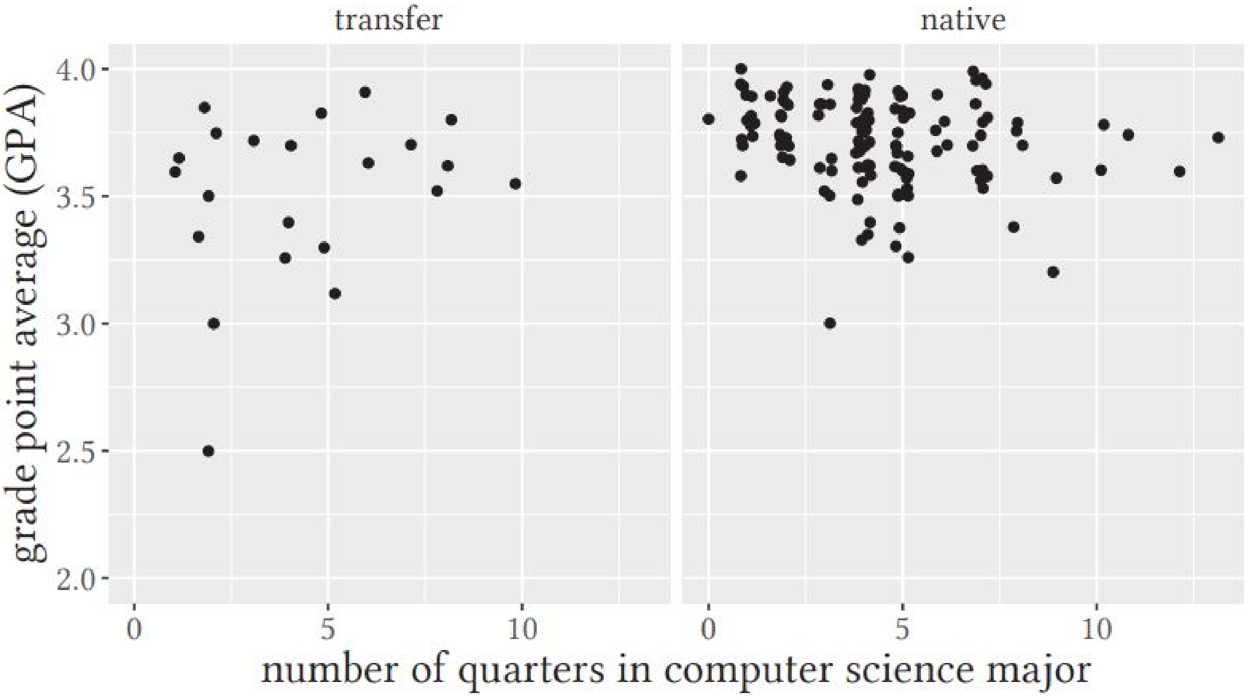
My research has two axioms: 1) data are artificial constructs that reflect decisions and biases of people who created and use them; and 2) equity is too complex of a goal to achieve through complete automation. With those truths in mind, I have explored using data to support equity in computing education by understanding minoritized students as people and not just data, considering data as well as context to address inequities, designing human-centered AI to amplify minoritized voices while maintaining student privacy. In this article, I will share what I have learned with a goal of providing others ideas, approaches, and language for creating a more equitable and just computing community.

Equity by Understanding, Empowering End-Users

People are more than the data we produce about them and use to represent them. For example, if we perceive students only through the decontextualized data we collect from them, we risk defining them by their perceived problems or shortcomings, which would result in a deficit framing. A deficit framing would frame low pass rates for Black and Hispanic students for AP Computer Science exams as a shortcoming of students of color. But doing so is both erroneous and unconstructive. To enable equitable learning, we must first understand end-users as complex human beings embedded in social contexts. Equitable teaching involves framing how the knowledge and skills of culturally diverse students are strengths (asset framing [7]), how students bring with them different prior experiences and prior knowledge (preparatory privilege [6]), and how ethnicity, gender, and other aspects of identity intersect to shape people's lived experiences (intersectional identity [2]). To do so, we must understand students not as data but as people who are often minoritized by exclusive societal norms and structures.

Understanding minoritized groups can occur at various levels of depth and scale. More in-depth studies can help us better understand how and why this minoritization occurs. As an example, I worked with a then-undergrad researcher on understanding academic and social experiences of transfer students [5]. Transfer students, or students who transferred from one post-secondary institution to another (e.g. from community college to four-year institution) make up about half of recent computer and information sciences graduates surveyed by the NSF in 2010. Yet little is known about the experiences of these students who tend to be more ethnically and gender diverse, older, have more financial and familial responsibilities, and live further away from campus. Prior researchers had observed "transfer shock" where new transfer students would face challenges adjusting to a new academic and social environment, often resulting in an initial drop in GPA. Through surveys and interviews with transfer students and non-transfer students at the University of Washington, we learned who these transfer students were, what their transfer process was like, how they found support, and what challenges they still faced. This more nuanced understanding provided explanations to factors contributing to "transfer shock" and what institutions may be able to do to better support their diverse transfer students. So to use data for equity, we must understand not just the data we produce and use but also the people who we expect this data to represent.

Xie, Benjamin. "How Data Can Support Equity in Computing Education." *XRDS* 27, no. 2 (December 14, 2020): 48–52. https://doi.org/10.1145/3433136



GPA of transfer and non-transfer ("native") students. While there was a statistically significant difference in GPA between the groups, we chose not to report it because doing so would invite misinterpretation. We found GPA was too confounded by differences in non-CS courses between groups. Fig 1. from [5].

With a deeper understanding of minoritized students and their needs, we may be able to design systems to support equitable learning by empowering students to take ownership over their own learning experience. Rather than have a standardized one-size fits all online learning experience or one where a data-driven system prescribes the next thing to do, I explored the effect of affording learners the agency to make decisions on what to learn next. My colleagues and I built Codeitz, an online learning tool to teach programming which affords learners the agency to decide for themselves what to learn next while also providing information to inform their decision [10]. We theorized that a learner with low self-efficacy (had little belief in their capacity to take actions to learn) could follow the recommendations we provided, while a learner who was more confident could deviate from the recommendations and explore as they wanted. We found that providing learners multiple pathways and information to help them decide how to navigate them did not translate to improved learning outcomes. This may have been because taking ownership of a learning experience deviated from expectations of students who were more familiar with having experiences defined for them. Furthermore, students were skeptical of the adaptive recommendations from our statistical model. Trust in data-driven outputs must be earned, especially by end-users!

So equity by using data to understand diverse students and afford them multiple pathways of engagement is an ongoing investigation. But understanding people is just the beginning. Understanding the context of data use is the next step.

Equity by interrogating the content and context of data

Because data are representations created through a series of decisions made by a select group, using them for equity requires us to interrogate not just the content but also the context.

I am interrogating and analyzing data to understand potential equity issues as an intern with Code.org, "a nonprofit dedicated to expanding access to computer science in schools and increasing participation by young women and students from other underrepresented groups." The goal of my research with Code.org is to understand how well their middle school CS Discoveries curriculum serves the 550,000+ students and 2,500 teachers from around the world who rely on it to learn computing. Ideally, this curriculum would be able to serve students from minoritized groups (young women, students of color) at least as effectively as it does for other students. Statistical analysis of assessment data to identify patterns which may suggest validity problems or bias in a test [9] is a reasonable first step. But analyzing learning outcomes just shows a decontextualized endpoint.

To really understand how equitable Code.org's curriculum is, we are interrogating the content as well as the context of data. Data are part facts, part artifacts, come with in-built intentions, and are created in pursuit of specific goals and purposes [3]. Understanding how to interpret and use data also involves understanding it through interrogation of the decisions and values latent in its production and analysis. This involves not just judging assessment data as right or wrong, but also looking at what knowledge the assessment intends to measure and what may confound that measurement.

For example, imagine if data indicated that a minoritized group got a test question wrong more frequently. That could be because the test question was inaccessible for students with color blindness, because it assessed knowledge beyond what the instructional materials covered, because the instructional material included culturally presumptuous language that students from outside of the western world may find confusing, because teachers had less access to professional development, or something else entirely. This understanding involves partnerships with domain experts (instructional designers, instructors, students) to interrogate not just the results in the data but the context that may help us interpret the results. Data can help us identify potential sources of inequities, but to validate and address these issues involves a deep understanding of the context surrounding it.

Just as we want a test score to represent knowledge of a subject, we want our data to be valid representations of some more complex phenomena. But validity is a fragile construct that is

Xie, Benjamin. “How Data Can Support Equity in Computing Education.” *XRDS* 27, no. 2 (December 14, 2020): 48–52. https://doi.org/10.1145/3433136

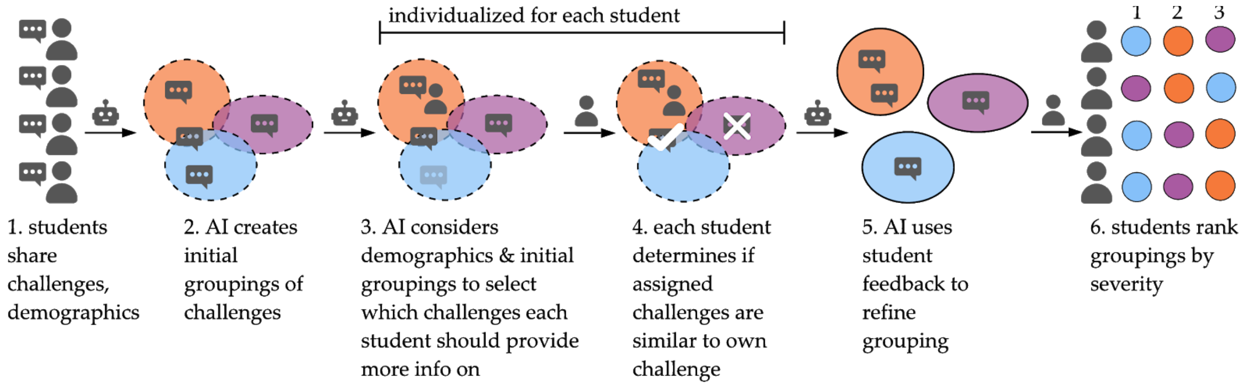
dependent on how we plan to interpret and use our data [4]. So if we want to use data to support equity, we must constantly interrogate our processes to identify how our decisions potentially exclude and minoritize those the data are supposed representa.

Using data to help people identify where and how to investigate

We may also be able to support equity through organized and annotated data helping people make sense of a larger phenomena, such as student experiences.

In large CS courses of up to 600 students, lone instructors often struggle to understand the experiences of all their students. Furthermore, students in minoritized groups often face unique challenges that their peers, who vastly outnumber them, do not face. For example, our pilot studies identified a Black student not being able to consistently attend lectures because he had to work three jobs, and a sexual assault survivor who found required course readings triggering to their PTSD. After being made aware of these challenges, instructors took concrete actions to accommodate these students’ challenges. Leaving these challenges hidden and unaddressed widens disparities in learning experiences for minoritized students [6], but informing instructors of inequities can help them address disparities.

To anonymously amplify minoritized voices, I'm investigating how an interactive machine learning system could inform instructors by contextualizing student feedback with demographic information and student perceptions. This system is detailed in the figure below. At a high level, this system asks students to share challenges getting in the way of their learning as well as demographic information (Step 1). It then uses unsupervised machine learning to attempt to group similar challenges (Step 2). Using demographic information and initial groupings, it then asks students to comment on how similar the challenge they shared is to their challenges of peers with similar demographics (step 4). This information will help refine groupings (step 5). Finally, students will rank groupings of challenges based on how severely they think those challenges disrupt learning (step 6). The output is a report of challenges students face with each challenge having contextual information including k-anonymous demographic labels and student perceptions of challenge severity.


individualized for each student
1 2 3
1. students share challenges, demographics
2. AI creates initial groupings of challenges
3. AI considers demographics & initial groupings to select which challenges each student should provide more info on
4. each student determines if assigned challenges are similar to own challenge
5. AI uses student feedback to refine grouping
6. students rank groupings by severity

The output will be a report an instructor could use to identify what challenges affect which student groups, as well as how severe students perceive different groupings of challenges to be. I am currently investigating how this information can help instructors make changes to support more equitable learning in their classes as well as how these reports may foster broader conversations on addressing systemic injustices.

Designing this system required considerations that include how to consider intersectionality of identity while ensuring privacy, how to establish trust in data-driven outputs through transparency (e.g. conveying uncertainty), and how to ensure minoritized perspectives would not be lost in the aggregation and reduction that comes in data analysis.

Conclusion: Equity is about Augmenting Human Intelligence

Equity is a complex and evolving goal. Data from a bygone past can help us by calling attention to unusual patterns which may suggest equity issues. But it's up to humans to interpret and use this information appropriately. Unfortunately, those with the expertise to understand the context are often not the ones designing our coded gazes. But if we can design tools that provide context by enabling people to qualify any quantifications of them, we can provide people the information they need to enact equitable change.

If our goal is truly to have more equity in the world, to understand, empower, and amplify students who have historically minoritized, then we must consider more than the content of data and what we do with it. We must also consider the context they will exist in and how domain experts interpret and use these data. With careful consideration of the context as well as the content of data and clear understandings of how our goals translate to design decisions, we can make steps towards using data to support a more equitable and just computing community.

Xie, Benjamin. "How Data Can Support Equity in Computing Education." *XRDS* 27, no. 2 (December 14, 2020): 48–52. https://doi.org/10.1145/3433136

Bio

Benjamin Xie is a PhD candidate at the University of Washington Information School and a research intern at Code.org. At UW, he is advised by Prof. Amy J. Ko in the Code & Cognition Lab. Benjamin's research interest is in designing interactive tools that support equity in computing education by affording learners the agency to advocate for their own learning experiences. He engages in the fields of computing education, human-computer interaction, and learning at scale. He is a National Science Foundation (NSF) Graduate Research Fellow. He received his bachelor's and master's degrees in computer science at MIT, researching with Prof. Hal Abelson and MIT App Inventor as a MIT EECS-Google Research and Innovation Scholar.

References

[1] Buolamwini, J.A. 2017. *Gender shades : intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. Massachusetts Institute of Technology.

[2] Crenshaw, K.W. 1994. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *The Public Nature of Private Violence*. M.A. Fineman and R. Mykitiuk, eds. Routledge. 93–118.

[3] Hardy, L. et al. 2020. From Data Collectors to Data Producers: Shifting Students' Relationship to Data. *Journal of the Learning Sciences*. 29, 1 (Jan. 2020), 104–126.

[4] Kane, M. 2010. Validity and fairness. *Language Testing*. 27, 2 (Apr. 2010), 177–182.

[5] Kwik, H. et al. 2018. Experiences of Computer Science Transfer Students. *Proceedings of the 2018 ACM Conference on International Computing Education Research* (2018), 115–123.

[6] Margolis, J. 2010. *Stuck in the Shallow End: Education, Race, and Computing*. MIT Press.

[7] Robinson, K. et al. 2018. Using Online Practice Spaces to Investigate Challenges in Enacting Principles of Equitable Computer Science Teaching. *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (New York, NY, USA, Feb. 2018), 882–887.

[8] Strunk, K.K. and Locke, L.A. eds. 2019. *Research Methods for Social Justice and Equity in Education*. Palgrave Macmillan.

[9] Xie, B. et al. 2019. An Item Response Theory Evaluation of a Language-Independent CS1 Knowledge Assessment. *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (2019), 699–705.

[10] Xie, B. et al. 2020. The Effect of Informing Agency in Self-Directed Online Learning Environments. *Proceedings of the Seventh (2020) ACM Conference on Learning @ Scale* (2020)