Hi
name + pronouns
I'm a PhD candidate at the University of Washington Seattle,
a university which acknowledges the Coast Salish peoples of this land, the land which touches the shared waters of all tribes and bands within the Duwamish (doo-amish), Puyallup (pee-all-up), Suquamish, Tulalip (too-lay-lip) and Muckleshoot nations.
And I'm EXCITED to share work that I've done with colleagues from the UW Information School, College of Education, DUB group, as well as with nonprofit code.org

Tests are not perfect measurement instruments

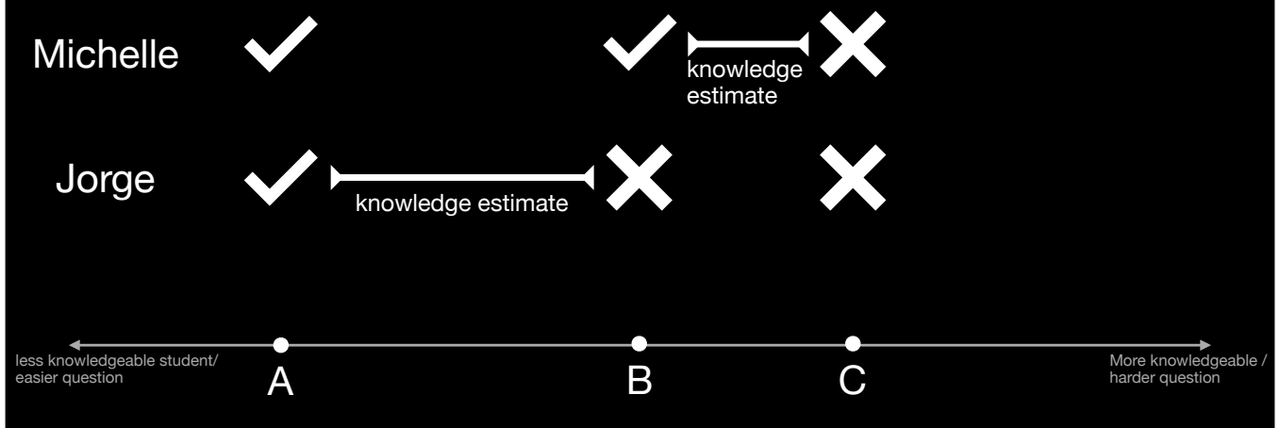
@benjixie
bxie@uw.edu

We use test scores for a lot of things.
University use scores for test such as the AP CS exams to determine if a student should be accepted into a university or major.
Teachers use tests for summative purposes such as grading.
Students use tests to self-assess what they know, and the results can affect their self-efficacy and sense-of belongingness.
How do we know how good our tests are?

How different people interpret and use test scores is important, but tests are imperfect measures of knowledge.

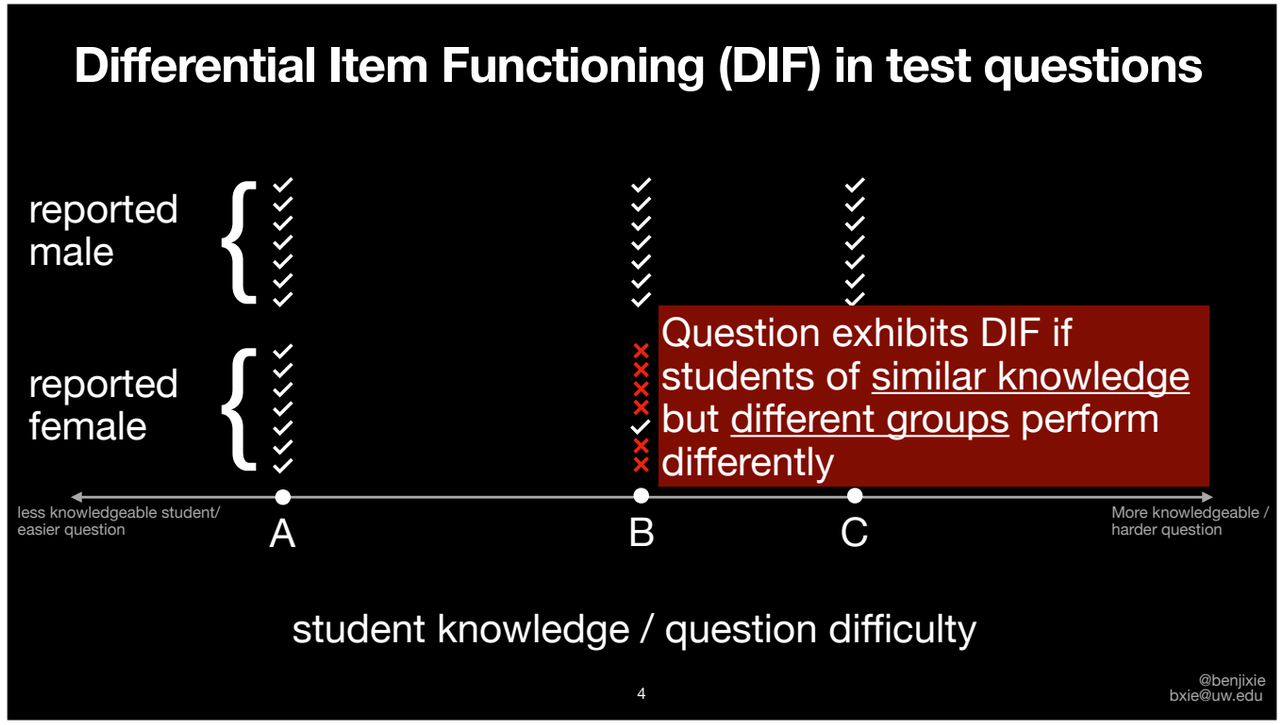
Modeling students & questions (w/ IRT)

Student knowledge & question difficulty share continuous dimension



@benjixie
bxie@uw.edu

To understand how good a test is, we have to make a few assumptions.
Following Item Response Theory (IRT), we can assume that student knowledge and question difficulty are on the same continuous dimension.
Say we have questions A, B, C, where A is the least difficult question and C is the most difficult.
Say two students, Michelle and Jorge, answer these questions.
Based on their responses to these questions to estimate their knowledge levels. Michelle's is between B and C because she got B correct but C incorrect.
By a similar logic, Jorge's knowledge is between A and B.



Now say we wanted to look at a group of students who reported as male and a group who reported as female. All students got questions A and C correct.
We would expect them to get question B correct as well. And say all the students who reported as male do get B correct.
CLICK
But say we observe that most students who report as female get question B wrong.
CLICK
This is a toy example that demonstrates Differential Item Functioning (DIF)
CLICK
where students oh similar knowledge levels but different groups (genders in this case) perform differently on an item, question B in this case.

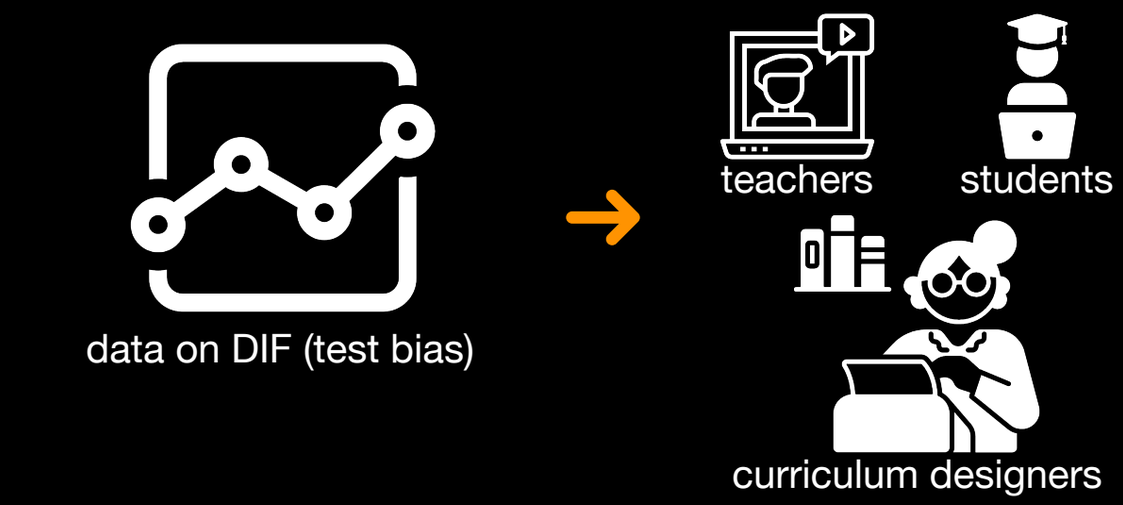
DIF is a technique to identify potential bias in test questions.

What if DIF could signal opportunities for better pedagogy?

@benjixie
bxie@uw.edu

Traditionally, DIF has been used as a filter…
what if we could use it to improve equitably we teach?

Connecting data on DIF w/ domain experts



@benjixie
bxie@uw.edu

Data on DIF can help identify or substantiate nuanced patterns of disparities or bias. But we need the domain expertise of stakeholders such as students, teachers, and curriculum designers to interpret and use these findings to address inequities.

How do domain-experts use data on test bias by gender and race for equity?

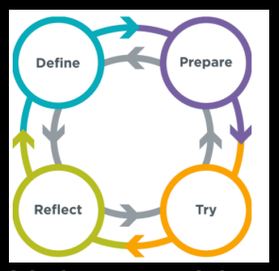
@benjixie
bxie@uw.edu

So this paper explores how domain-experts (curriculum designers) might be able to interpret and use gender and race-based DIF for equity related goals.

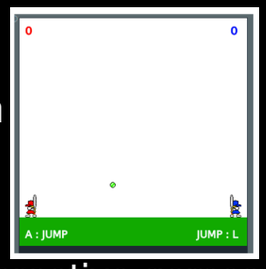
Code.org CS Discoveries (CSD) ’19-‘20

- Computing as creative form of expression
- 19,617 students (most 11 - 16 yrs old)
- 17 questions for formative use

Unit 2: web dev



Unit 1: problem solving



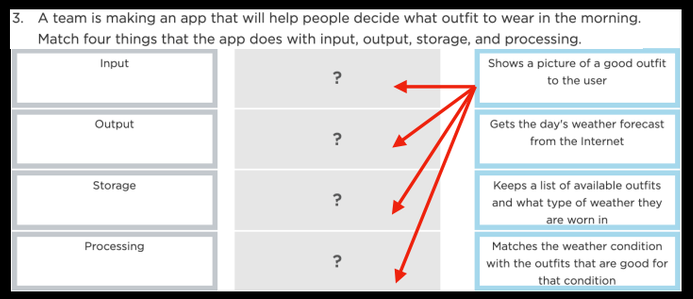
Unit 3: interactive games

@benjixie
bxie@uw.edu

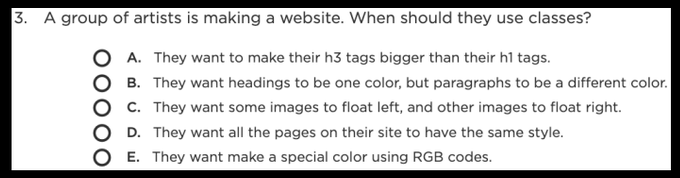
For this study, I partnered with Code.org, a nonprofit dedicated to inclusive. Computing education
We analyzed responses from ~20,000 students for middle school CS Discoveries (CSD) curriculum

17 CSP items: multiple choice, matching

Matching (Unit 1, Q4)



Multiple-choice (U2, Q3)



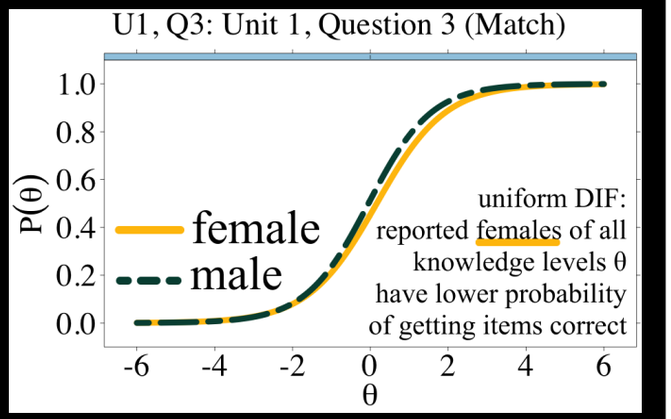
@benjixie
bxie@uw.edu

The 17 questions I analyzed were either matching questions or multiple choice questions.
Matching questions required students to place options in their correct locations.
Multiple choice questions required students to choose one or two options.
All questions scored dichotomously (entirely correct or entirely wrong)

Found evidence of bias by gender, race

Question **biased** if >5% difference in probability that students w/ same knowledge, different groups get question correct ($p < 0.001$, medium or large effect size)

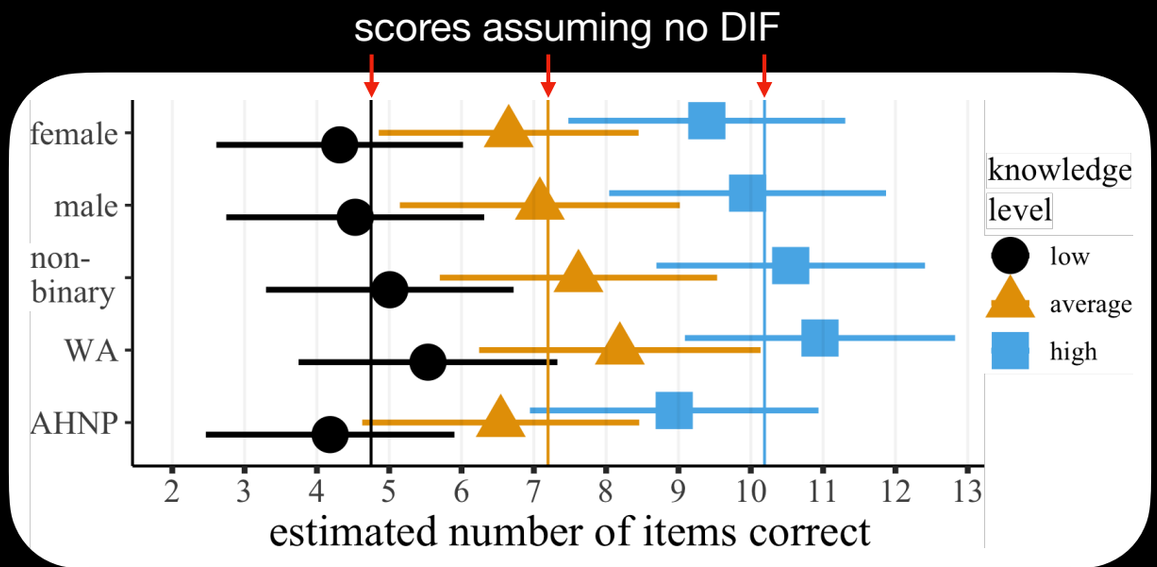
- 2 questions disadvantaged reported **females** vs reported **male**
- 13 questions disadvantaged **AHNP** (African/Black, Hispanic/ Latinx, Native American/Alaskan Native, and Pacific Islander) vs **WA** (white, Asian)



@benjixie
bxie@uw.edu

Our quantitative analysis focused on checking for potential test bias by gender and race.
We say a question is biased if on average, a student from a disadvantaged group is at least 5% less likely to get that question correct compared to a student of similar knowledge from the other group. This is equivalent to checking for statistical significance with a medium or large effect size.
We found two questions disadvantaged students who reported as female compared to students who reported as male.
CLICK and describe trace plot
Most test questions disadvantaged AHNP students (African/Black, Hispanic/Latinx, Native/Alaskan Native, and Pacific Islander) compared to WA students (white, Asian)

Students of equivalent knowledge levels would have performed differently

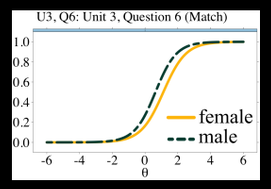


@benjixie
bxie@uw.edu

Put together, we can say that students of equivalent knowledge but different genders or races would score differently on the CSD assessments.

So as a whole, this test disadvantages AHNP and reported female students the most, and advantages WA students the most.

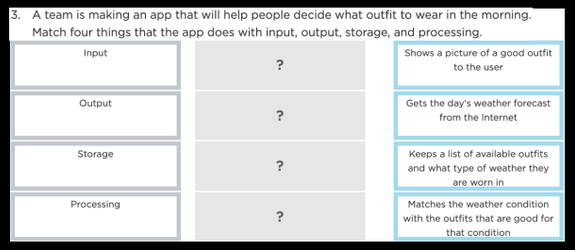
How curriculum designers interpreted DIF



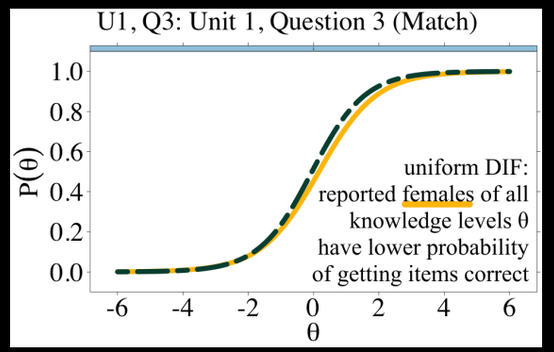
@benjixie
bxie@uw.edu

DIF does not tell us the cause of this bias or what to do about it.
So to understand that, we conducted a remote workshop where 7 Code.org curriculum designers interpreted 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.
Here are a few high level takeaways, but I point you to the paper to read more about our findings.

Considering question properties relative to student identities



"Female students are performing lower on matching [type] questions…"

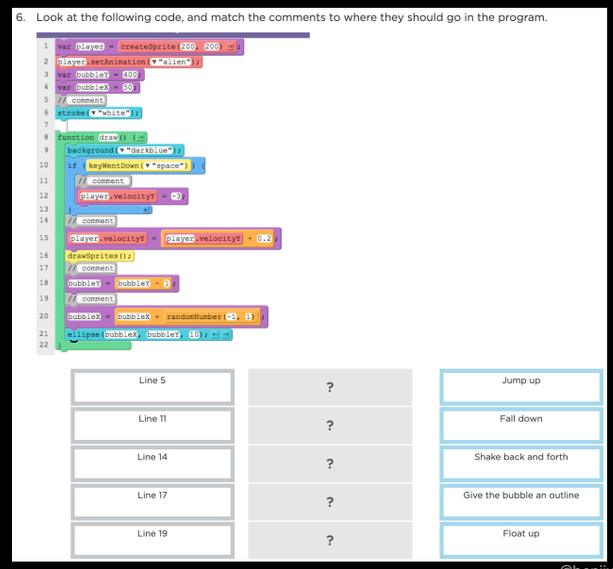


@benjixie
bxie@uw.edu

Curriculum designers considered how test design may have introduced bias, with some identifying how matching type questions disadvantaged students who reported as female.

Alignment between assessment and curriculum

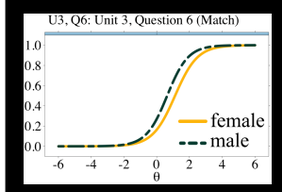
"Comments are not very well emphasized in CS Discoveries… this may be the very first [time] that students are seeing this idea of putting a comment to a block of code."



@benjixie
bxie@uw.edu

Curriculum designers also considered how the curriculum may or may not have prepared students for the test questions.
So in one case, curriculum designers acknowledged that commenting code was a skill worth learning, but may not have been well taught prior to this test question.
Considering how specific aspects test and curriculum design may contribute to bias is a potential first step to making changes that support more equitable learning experiences.

Implications: Equitable action by contextualizing data w/ domain expertise



data on DIF (test bias)

Identify nuanced patterns, bias

curriculum designers

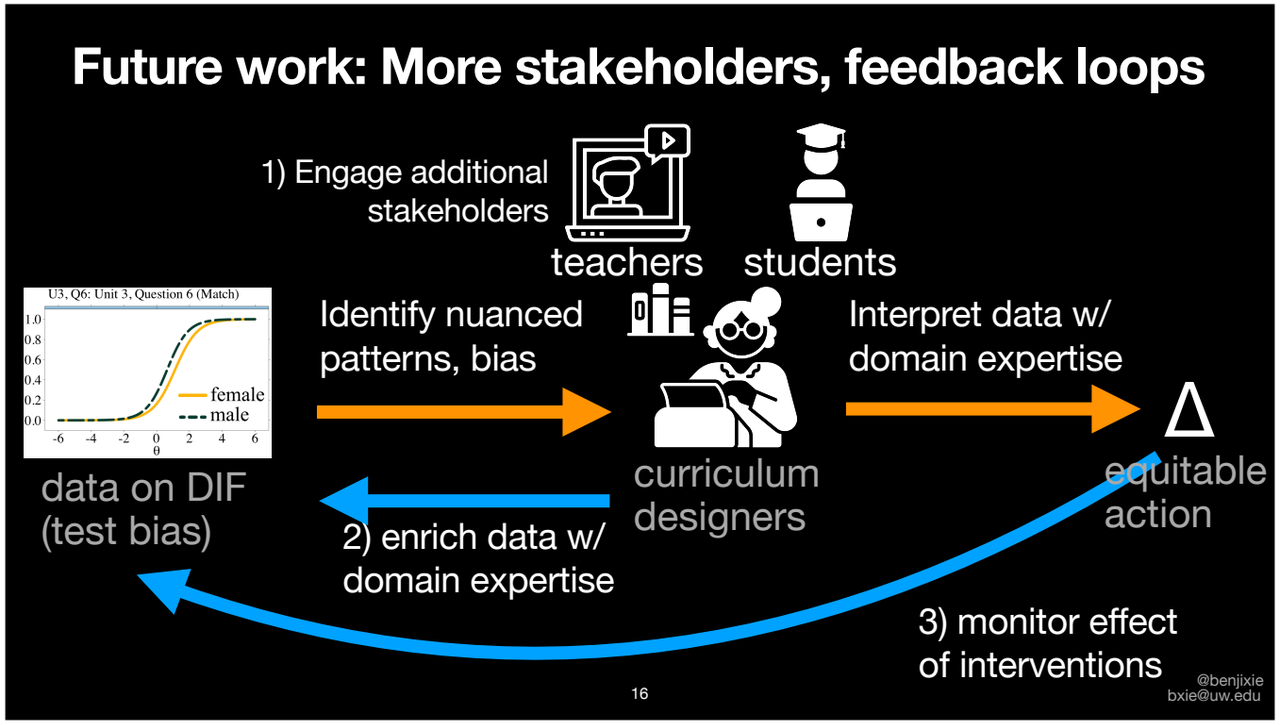
Interpret data w/ domain expertise

Δ

equitable action

@benjixie
bxie@uw.edu

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



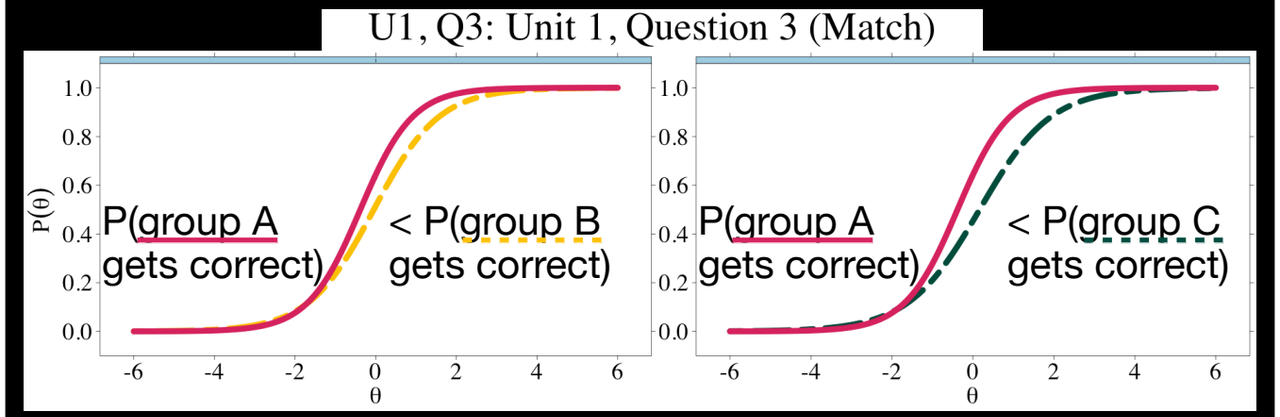
Some potential future work includes engaging additional stakeholders,
Using human-centered AI techniques to enrich data with domain expertise,
And monitoring the effect of interventions

friendly reminder: data w/o context can be misleading

@benjixie
bxie@uw.edu

(time check)
In the spirit of pride month, I did want to share one finding which also serves as yet another reminder that data without context invites misinterpretation

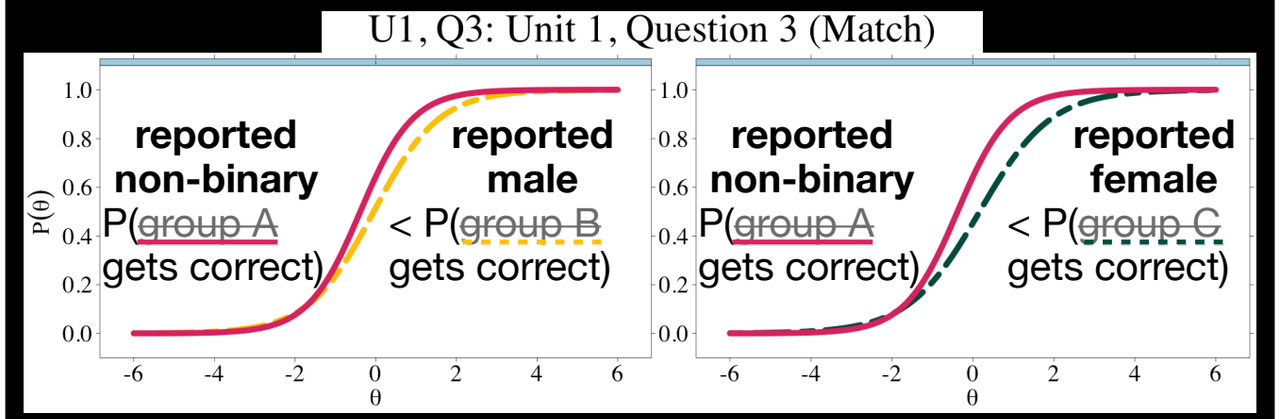
data-driven: question provides **group A** unfair advantage



@benjixie
bxie@uw.edu

Here are two trace plots which show the probably that learners from different groups will get a test question correct given different latent knowledge levels (theta). These two plots show that this question unfairly advantages group A, the solid red line. So if you had three students of the same knowledge level, the student from group A would still perform better compared to students from groups B and C.

~~data-driven: question provides reported non-binary students unfair advantage~~



evidence-driven: data on non-binary students may not reflect reality

@benjixie
bxie@uw.edu

These three groups are actually reported genders students.
So the data-driven result is that students who report as non-binary have an advantage for this item and 4 others, and that we must make changes to better support students who report as male or female.
If you have a feeling that something isn't quite right, we did too. Because there is a mountain of computing and STEM education literature detailing the minoritization of non-binary students.
So we consulted with learning scientists and critical scholars and found out that coming out as non-binary in primary school is a constant and challenging process and that a majority of non-binary students may have been assigned female at birth. So the evidence driven result is that our data and labels may not reflect identity of students.

So the more accurate statement is that this data-driven result is actually misleading because of its lack of context. So let’s replace that….
CLICK
with a more contextualized evidence-driven finding that suggests that data on non-binary students may not reflect student experience. So all this is to say that, analyzing data is dangerous, so we need to consult with domain-experts to contextualize and validate our supposed findings.

Domain Experts' Interpretations of Assessment Bias in a Scaled, Online Computer Science Curriculum

Benjamin Xie
University of Washington
bxie@uw.edu

Matt J. Davidson
University of Washington
mattjd@uw.edu

Baker Franke
Code.org
baker@code.org

Emily McLeod
Code.org
emily@code.org

Min Li
University of Washington
minli@uw.edu

Amy J. Ko
University of Washington
ajko@uw.edu

ABSTRACT
Understanding inequity at scale is necessary for designing equitable online learning experiences, but also difficult. Statistical techniques like Differential Item Functioning (DIF) can help identify whether items/questions in an assessment exhibit potential bias by disadvantaging certain groups (e.g. whether item disadvantages woman vs man of equivalent knowledge). While testing companies typically use DIF to identify items to remove, we explored how domain-experts such as curriculum designers could use DIF to better understand how to design instructional materials to better serve students from diverse groups. Using Code.org's online Computer Science Discoveries (CSD) curriculum, we analyzed 139,097 responses from 19,617 students to identify DIF by gender and race in assessment items (e.g. multiple choice questions). Of the 17 items, we identified six that disadvantaged students who reported as female when compared to students who reported as non-binary or male. We also identified that most (13) items disadvantaged AIINP (African/Black, Hispanic/Latinx, Native American/Alaskan Native, Pacific Islander) students compared to WA (white, Asian) students. We then conducted a workshop and interviews with seven curriculum designers and found that they interpreted item bias relative to an intersection of item content and student identity, the broader curriculum, and differing uses for assessments. We interpreted these findings in the broader context of using data on bias to inform domain-experts' efforts to design more equitable learning experiences.

Author Keywords
computing education, differential item functioning, item response theory, test bias

L@S '21 June 22–25, 2021, Potsdam, Germany

ISBN 978-1-4503-8215-1/21/06…$15.00
DOI 10.1145/3430895.3460141

CCS Concepts
•**Social and professional topics** → **Student assessment;** *K-12 education;* •**Human-centered computing** → *Empirical studies in HCI;*

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 [25]. Equity also implies a goal of implementing corrective measures to adjust for aggregate harm from historic social inequalities [50]. This might mean providing additional support to students from minoritized groups ([2, 70, 33]), as they face unique challenges that if left unaddressed could pose serious impediments to science and technology learning [12]. Previous efforts to design equitable learning experiences include designing adaptive and personalized online environments [5, 72], adjusting environments to support inclusion [37], and enabling broader access to learning [38].

However, achieving equity is rarely straightforward: inequities in learning stem from a complex interplay between multiple structures and interactions [56]. 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 [45, 11, 23].

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 [22, 26]. Students alone cannot be responsible for identifying equity issues because their focus is on learning [44, 34] and self-advocacy may bring about burden and risks to minoritized groups including stereotype threat [66, 19] and social-desirability biases [28, 27]. Teachers have a significant role in addressing inequities [56], but they would need information that is understandable and actionable [4, 26] 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 [50]. Data can support equity by enabling rapid improvement of practices through experi-

- **Validity evidence** for most used middle school computing test
- DIF for students who reported as **non-binary**
- **Challenges** with domain expert interpretation of DIF

@benjixie
bxie@uw.edu

I encourage you to read the paper to learn more about validity evidence, why our analysis of non-binary students' responses required contextualization, and challenges with interpreting DIF.

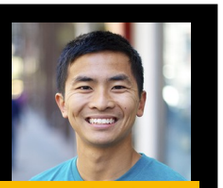
Data helps us identify existence and extent of biases.
Domain expertise helps us identify causes, take equitable action

@benjixie
bxie@uw.edu

But the main take away is this:

Domain Experts' Interpretations of Assessment Bias in a Scaled, Online Computer Science Curriculum

Benjamin Xie | bxie@uw.edu | @benjixie | benjixie.com/las21



On job market! Post-doc, research scientist.

- **RQ**: How do domain-experts use data on test bias by gender and race for equity?
- **takeaway**: Data identifies inequalities, unfairness, bias. Equitable action requires engaging domain experts.
- **design implication**: Domain experts can connect bias to causes, identify equitable action

Questions I have:

1. If minorizited groups are small, what should we do with data about them?
2. Ideally, who should be interpreting and using our findings? How do we support them?
3. How should we balance privacy and protection with checking for bias?
4. How do people qualify our labels/ representations of them?

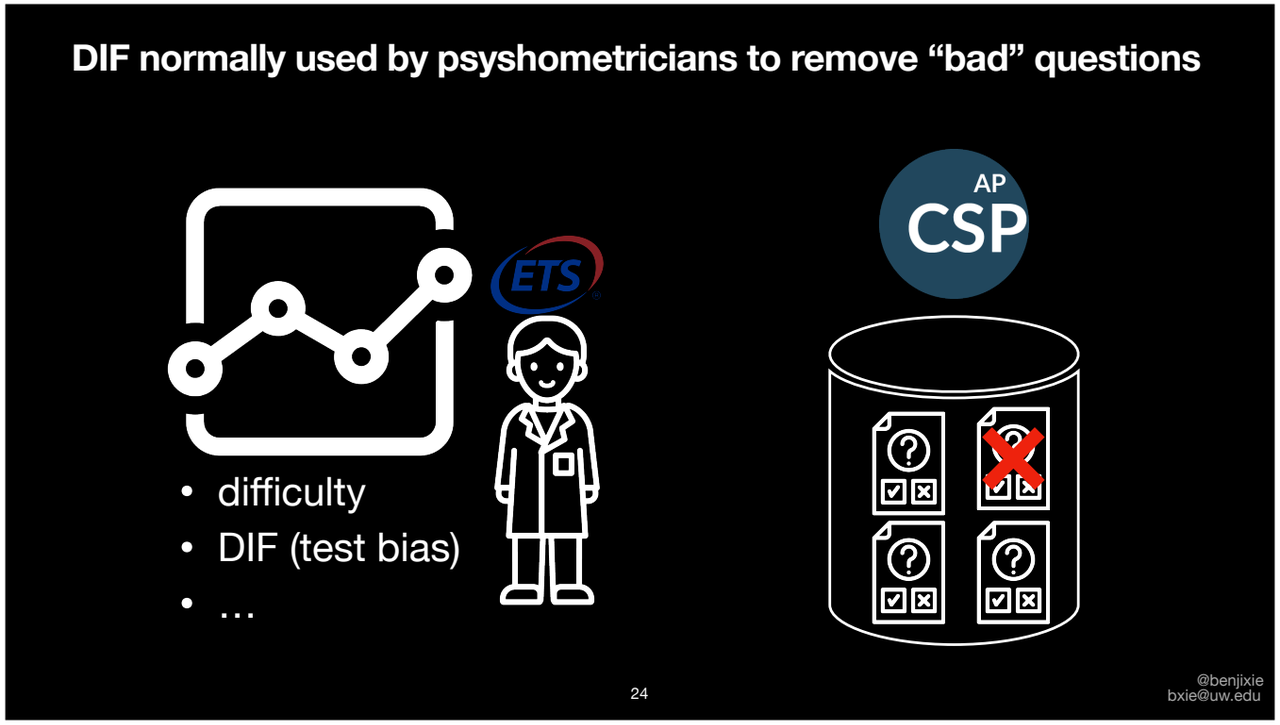
Information School UNIVERSITY of WASHINGTON | COLLEGE OF EDUCATION UNIVERSITY of WASHINGTON | C O D E | NSF

@benjixie
bxie@uw.edu

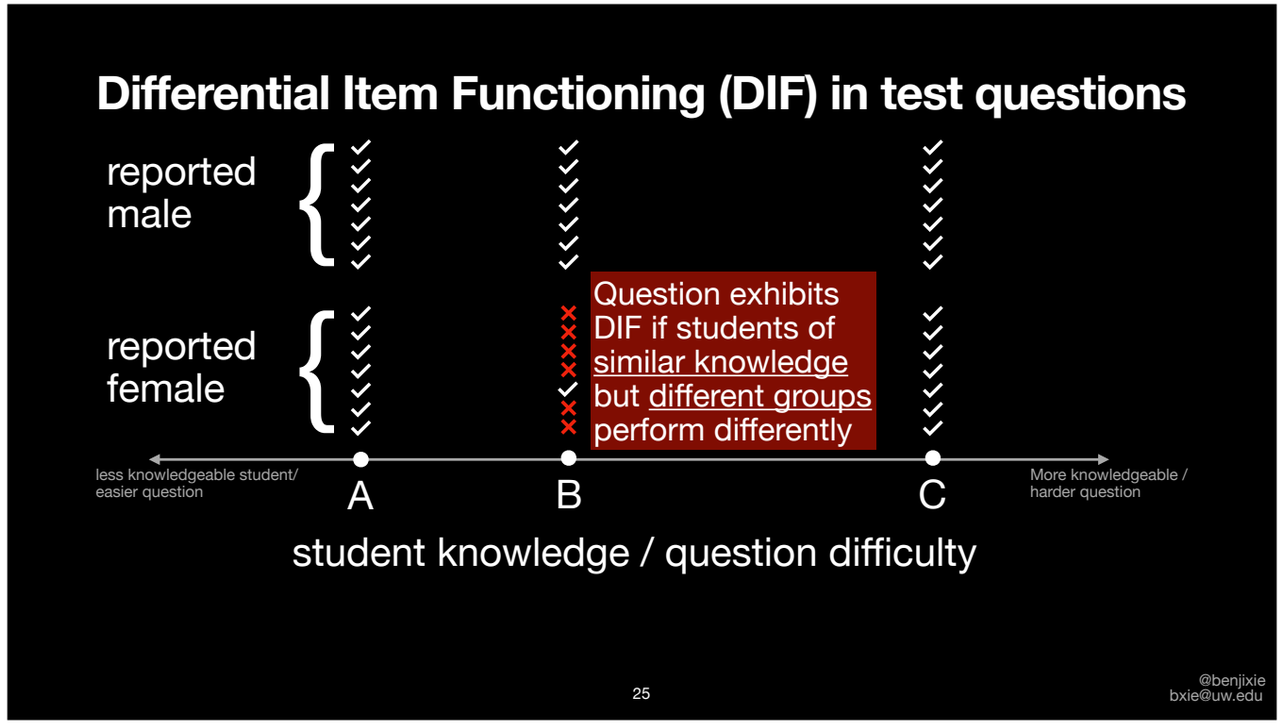
The link to read the full paper and access supplementary material is benjixie.com/las21.
I'm legally obligated to inform everyone I run into that I'm on the job market, seeking postdoc and research scientist opportunities and would love to chat.
And I'll conclude by pointing you to a list of 4 questions that keep me up at night that can inspire some discussion and ideation.

Supplementary slides

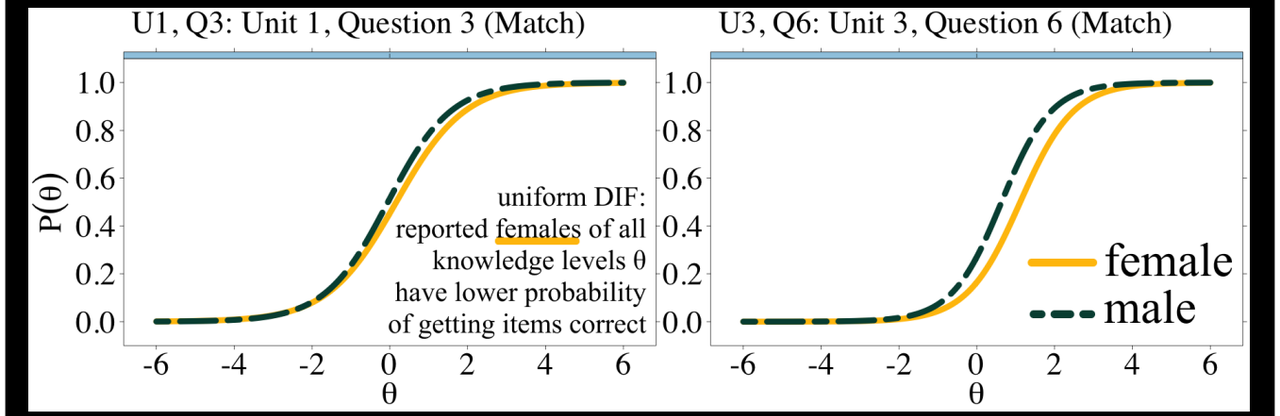
@benjixie
bxie@uw.edu



DIF is often used by educational testing companies to ensure high stakes exams are fair. For example, say someone at ETS was creating questions for this upcoming years’ AP Computer Science Principles exam. They would likely use DIF techniques to identify questions that exhibited DIF and remove them because they may disadvantage certain groups (by gender or race for example).

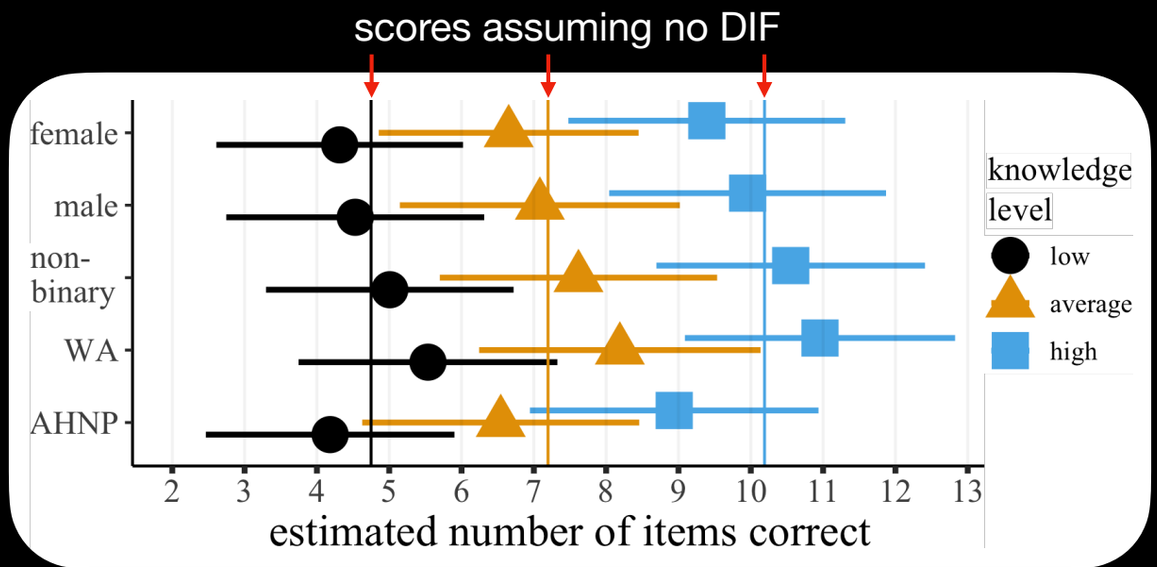


Now say we wanted to look at a group of students who reported as male and a group who reported as female. All students got questions A and C correct.
We would expect them to get question B correct as well. And say all the students who reported as male do get B correct.
But say we observe that most students who report as female get question B wrong.
This is a toy example that demonstrates Differential Item Functioning (DIF), where students oh similar knowledge levels but different groups (genders in this case) perform differently on an item, question B in this case.
DIF is a technique to identify potential bias in test questions.



@benjixie
bxie@uw.edu

Students of equivalent knowledge levels would have performed differently



@benjixie
bxie@uw.edu

Put together, we can say that students of equivalent knowledge but different genders or races would score differently on the CSD assessments.

So as a whole, this test disadvantages AHNP and reported female students the most, and advantages WA students the most.

Questions that keep me up at night

1. If minorizited groups are small,
   what should we do with data about them?
2. Ideally, who should be interpreting and using our findings?
   How do we support them?
3. How should we balance privacy and protection with checking for bias?
4. How do people qualify our labels/ representations of them?

@benjixie
bxie@uw.edu