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 AHNP (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 features and student identity, the broader curriculum, and differing uses for assessments. We interpreted these findings in the broader context of using data on assessment bias to inform domain-experts' efforts to design more equitable learning experiences.

Author Keywords

computing education, differential item functioning, assessment interpretation and use, validity, item response theory, test bias

L@S '21, June 22–25, 2021, Virtual Event, Germany.

ACM ISBN 978-1-4503-8215-1/21/06 ...$15.00.
https://dx.doi.org/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 and personalized 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 [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-

mentation and measurement of change that is understandable and actionable [4]. Large-scale analysis can reveal patterns not easily seen at a micro-level by individuals [24], such as through analysis of intersectional identities [53]. Data can provide evidence to support disruption of the status quo [26].

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 [8, 36, 62]. 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 [24]. Improving equity in learning experiences is a complex and iterative process that requires a multitude of methods and stakeholders' expertise [60].

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 [49, 47]. 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 [42, 15].

DIF analyses suggests potential bias, but judgement is required to interpret and act on DIF results [73]. 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.

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 [63]. 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 [14, 42, 16]) perform similarly on a given test question. Developers of concept inventories (tests used as a standardized measure of conceptual understanding [40]) often use DIF methods to understand how fair an inventory is (e.g. [42, 16]). 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 [74, 17]. Researchers have used DIF to detect potential bias in summative evaluations (e.g. final exams) in large courses, such as introductory CS [14]. 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 [17]. 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 [75]. 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 [75].

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 [75]:

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 [15, 71], 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 [76].

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 [30]. 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 [14].

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 [74]. 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 [73]. 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 [35]. 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.

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 [10]. 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*"[10].

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 2 describes the items.

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.

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 1 shows reported demographics for students.

Table 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

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 the *NR* column of Table 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 [1], shown in Table 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 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 [1, 15]. 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 [39, 48]. Removing any of the 17 items resulted in a decrease in α ($\Delta\alpha < 0$), so we analyzed all items.

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* [15]. 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 [13]. 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 [57]. Confirmatory analysis with one factor showed a strong model fit (RMSEA = 0.018, CFI = 0.916, TLI = 0.903) [6].[1]

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 with an additional parameter to account for guessing. We compared model fit using the Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) [15]. 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.

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 [7] 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).

[1] EFA was conducted using R with `psych::fa()` [51] using a maximum likelihood factor analysis with a varimax oblique transformation. CFA was conducted with `lavaan::cfa()` [54] with fixed residual variances (`std.lv=T`) and full information maximum likelihood (FIML) approach for handling missing data.

Table 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

To adjust for multiple comparisons, we used a Benjamini-Hochberg p-value correction [3], an adjustment that maximizes power while controlling the false discovery rate to the nominal value (in this case, 5%) [42]. While more advanced DIF methods enable comparison of a reference group to multiple focal groups (e.g. [69, 68]), 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[2] (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 [33, 19, 77, 61, 21], and WA racial groups tended to be dominant [65, 52, 41, 20].

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) [43]. Because these are dichotomous items scored as 0 or 1, we can interpret SIDS to be the average 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 [18].

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 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 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 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 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

[2]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 [67].

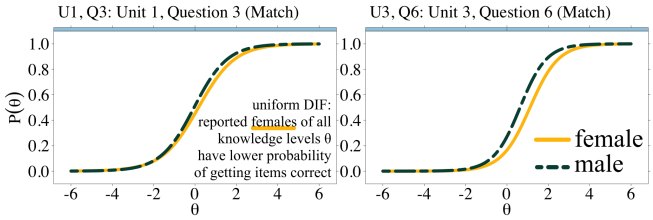

Figure 1. Traces for items that exhibited (uniform) gender-based DIF of medium or large effect. (Items w/ • in blue rows of Table 3)

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 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 3), we found that one item disadvantaged students who reported as male (*U1, Q4*, medium e.s.).

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. 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 2 shows the trace plots for items that exhibited uniform DIF with large e.s..

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-

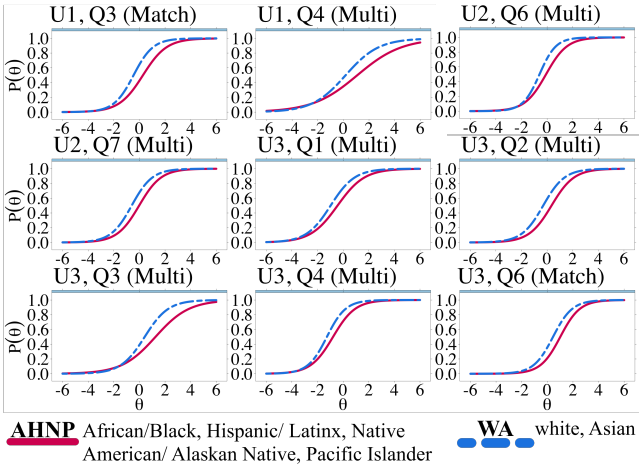

Figure 2. Traces for items that exhibited (uniform) race-based DIF with large effect size. (Items with •• in Table 4)

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[14], 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 [3] 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 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.

QUALITATIVE RESULTS: DESIGNERS' INTERPRETATION

To understand how domain experts interpreted and used DIF results, we conducted a workshop with seven curriculum de-

Table 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 $\geq$ 0.05), •• for large (SIDS $\geq$ 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

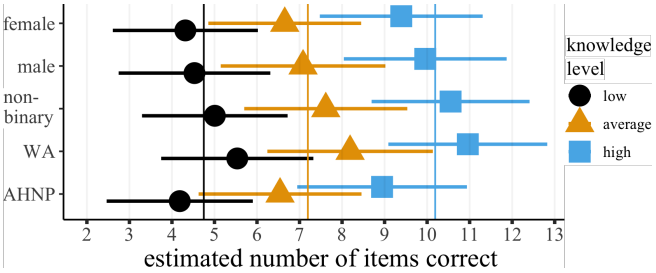

Figure 3. Expected number of items a student would get correct (out of 17) by gender and racial groups for three different knowledge levels. Knowledge levels were calculated with an IRT model assuming no DIF, where *average* is the median knowledge level in our sample ($\theta = -0.07$), *low* is a standard 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.

signers 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.

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. 1 and 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.

Table 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 $\geq$ 0.05), •• for large (SIDS $\geq$ 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

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.

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).

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].*"

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 3 and 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 [59, 64, 55] and test-taker self-efficacy [58, 31, 32].

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 [75]. 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 [29]. 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 [46], and a majority of non-binary students may have been assigned female at birth [9]. 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.

ACKNOWLEDGMENTS

This study was approved by the UW Institutional Review Board (IRB). We thank Code.org curriculum designers for their contributions. Code.org affiliates reviewed this publication and proposed changes. 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.

REFERENCES

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

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

[3] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* 57, 1 (Jan. 1995), 289–300.

[4] Marie Bienkowski, Mingyu Feng, and Barbara Means. 2012. *Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief*. Technical Report. U.S. Department of Education.

[5] Christopher Brooks, René F Kizilcec, and Nia Dowell. 2020. Designing Inclusive Learning Environments. In *Proceedings of the Seventh ACM Conference on Learning @ Scale (L@S '20)*. Association for Computing Machinery, New York, NY, USA, 225–228.

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

[7] R Philip Chalmers. 2012. mirt: A Multidimensional Item Response Theory Package for the R Environment [version 1.31]. *Journal of Statistical Software, Articles* 48, 6 (2012), 1–29.

[8] Peter Checkland and Sue Holwell. 1998. Action Research: Its Nature and Validity. *Systemic Practice and Action Research* 11, 1 (Feb. 1998), 9–21.

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

[10] Code.org Curriculum Team. 2019. CS Discoveries 2019-2020. https://curriculum.code.org/csd-19/. (2019). Accessed: 2021-1-17.

[11] Lindsay L Cornelius and Leslie Rupert Herrenkohl. 2004. Power in the Classroom: How the Classroom Environment Shapes Students' Relationships With Each Other and With Concepts. *Cogn. Instr.* 22, 4 (Dec. 2004), 467–498.

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

[13] Daniela R. Crişan, Jorge N. Tendeiro, and Rob R. Meijer. 2017. Investigating the Practical Consequences of Model Misfit in Unidimensional IRT Models. *Applied Psychological Measurement* 41, 6 (2017), 439–455. DOI:http://dx.doi.org/10.1177/0146621617695522 PMID: 28804181.

[14] Matt J Davidson, Brett Wortzman, Min Li, and Amy J Ko. 2021. 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.

[15] R J De Ayala. 2009. *The theory and practice of item response theory*. Guilford Press, New York.

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

[17] Neil J. Dorans. 2017. *Contributions to the Quantitative Assessment of Item, Test, and Score Fairness*. Springer International Publishing, Cham, 201–230. DOI: http://dx.doi.org/10.1007/978-3-319-58689-2_7

[18] Neil J Dorans and Edward Kulick. 1986. Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test. *Journal of Educational Measurement* 23, 4 (1986), 355–368.

[19] Remy Dou, Karina Bhutta, Monique Ross, Laird Kramer, and Vishodana Thamotharan. 2020. The Effects of Computer Science Stereotypes and Interest on Middle School Boys' Career Intentions. *ACM Transactions on Computing Education* 20, 3 (June 2020), 1–15.

[20] Ruth Dunn. 2021. *Minority Studies*. LibreTexts.

[21] Barbara Ericson and Mark Guzdial. 2014. 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)*. Association for Computing Machinery, New York, NY, USA, 217–222.

[22] Ronald F Ferguson. 2007. *Toward Excellence with Equity: An Emerging Vision for Closing the Achievement Gap*. Harvard Education Press.

[23] Rob Filback and Alan Green. 2013. New Directions for Diversity at USC Rossier. *Futures in Urban Ed, the Magazine of the USC Rossier School of Education* (Aug. 2013).

[24] Christian Fischer, Zachary A Pardos, Ryan Shaun Baker, Joseph Jay Williams, Padhraic Smyth, Renzhe Yu, Stefan Slater, Rachel Baker, and Mark Warschauer. 2020. Mining Big Data in Education: Affordances and Challenges. *Review of Research in Education* 44, 1 (March 2020), 130–160.

[25] Gail E FitzSimons. 2011. A Framework for Evaluating Quality and Equity in Post-Compulsory Mathematics Education. In *Mapping Equity and Quality in Mathematics Education*, Bill Atweh, Mellony Graven, Walter Secada, and Paola Valero (Eds.). Springer Netherlands, Dordrecht, 105–121.

[26] Julie Flapan, Jean J Ryoo, and Roxana Hadad. 2020. 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.

[27] Floyd J Fowler, Jr. and Thomas W Mangione. 1990. *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. SAGE.

[28] Pamela Grimm. 2010. Social Desirability Bias. In *Wiley International Encyclopedia of Marketing*, Jagdish Sheth and Naresh Malhotra (Eds.). Vol. 50. John Wiley & Sons, Ltd, Chichester, UK, 537.

[29] Anna Lauren Hoffmann. 2020. Terms of inclusion: Data, discourse, violence. *New Media & Society* (Sept. 2020), 1461444820958725.

[30] Paul W Holland, Howard Wainer, and Educational Testing Service. 1993. *Differential Item Functioning*. Psychology Press.

[31] Jeffrey D Holmes. 2020. The Bad Test-Taker Identity. *Teach. Psychol.* (Dec. 2020), 0098628320979884.

[32] M Horvath, A M Ryan, and S L Stierwalt. 2000. The Influence of Explanations for Selection Test Use, Outcome Favorability, and Self-Efficacy on Test-Taker Perceptions. *Organ. Behav. Hum. Decis. Process.* 83, 2 (Nov. 2000), 310–330.

[33] Aleata Hubbard Cheuoua. 2021. 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)*. Association for Computing Machinery, New York, NY, USA, 425–430.

[34] Ben Jee, Jennifer Wiley, and Thomas Griffin. 2006. Expertise and the illusion of comprehension. In *Proceedings of the Annual Conference of the Cognitive Science Society*. 387–392.

[35] Michael Kane. 2010. Validity and fairness. *Language Testing* 27, 2 (April 2010), 177–182.

[36] Stephen Kemmis. 2006. Participatory action research and the public sphere. *Educational Action Research* 14, 4 (Dec. 2006), 459–476.

[37] René F Kizilcec and Andrew J Saltarelli. 2019. Can a diversity statement increase diversity in MOOCs?. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale (L@S '19)*. Association for Computing Machinery, New York, NY, USA, 1–8.

[38] Sean Kross and Philip J Guo. 2018. 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 (L@S '18)*. Association for Computing Machinery, New York, NY, USA, 1–10.

[39] Charles E Lance, Marcus M Butts, and Lawrence C Michels. 2006. The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? *Organizational Research Methods* 9, 2 (2006), 202–220.

[40] Julie Libarkin. 2008. Concept Inventories in Higher Education Science. In *National Research Council Promising Practices in Undergraduate STEM Education Workshop*, Vol. 13. 14.

[41] Martin N Marger. 2015. *Race and Ethnic Relations: American and Global Perspectives, 10th Edition*. Cengage.

[42] Patrícia Martinková, Adéla Drabinová, Yuan-Ling Liaw, Elizabeth A Sanders, Jenny L McFarland, and Rebecca M Price. 2017. Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments. *Cell Biol. Educ.* 16, 2 (2017), rm2.

[43] Adam W Meade. 2010. A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology* 95, 4 (2010), 728.

[44] Daniel C Moos and Roger Azevedo. 2008. Self-regulated learning with hypermedia: The role of prior domain knowledge. *Contemp. Educ. Psychol.* 33, 2 (April 2008), 270–298.

[45] Christof Nachtigall, Ulf Kröhne, Ulrike Enders, and Rolf Steyer. 2008. Causal effects and fair comparison: Considering the influence of context variables on student competencies. In *Assessment of Competencies in Educational Contexts*, Johannes Hartig, Eckhard Klieme, and Detlev Leutner (Eds.). Hogrefe Publishing, 315–336.

[46] Kevin L Nadal. 2017. *The SAGE Encyclopedia of Psychology and Gender*. SAGE Publications.

[47] National Academies of Sciences, Engineering, and Medicine. 2018. *How People Learn II: Learners, Contexts, and Cultures*. National Academies Press, Washington, D.C.

[48] Jum C Nunnally. 1978. *Psychometric theory* (2d ed. ed.). McGraw-Hill, New York.

[49] Harold Pashler, Patrice M Bain, Brian A Bottge, Arthur Graeser, Kenneth Koedinger, Mark McDaniel, and Janet Metcalfe. 2007. *Organizing Instruction and Study to Improve Student Learning*. Technical Report NCER 2007-2004. U.S. Department of Education.

[50] Heather E Price. 2019. Large-Scale Datasets and Social Justice: Measuring Inequality in Opportunities to Learn. In *Research Methods for Social Justice and Equity in Education*, Kamden K Strunk and Leslie Ann Locke (Eds.). Springer International Publishing, Cham, 203–215.

[51] William Revelle. 2020. psych: Procedures for Psychological, Psychometric, and Personality Research. `https://CRAN.R-project.org/package=psych`. (Dec. 2020). version 1.9.12.31.

[52] Karen Rosenblum and Toni-Michelle Travis. 2015. *The Meaning of Difference: American Constructions of Race and Ethnicity, Sex and Gender, Social Class, Sexuality, and Disability*.

[53] Monique Ross, Zahra Hazari, Gerhard Sonnert, and Philip Sadler. 2020. 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 (Feb. 2020), 1–15.

[54] Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, Articles* 48, 2 (2012), 1–36. version 0.6-5.

[55] Toni Schmader, Michael Johns, and Chad Forbes. 2008. An integrated process model of stereotype threat effects on performance. *Psychol. Rev.* 115, 2 (April 2008), 336–356.

[56] Niral Shah and Colleen M Lewis. 2019. Amplifying and Attenuating Inequity in Collaborative Learning: Toward an Analytical Framework. *Cogn. Instr.* 37, 4 (Oct. 2019), 423–452.

[57] Suzanne L Slocum-Gori and Bruno D Zumbo. 2011. Assessing the Unidimensionality of Psychological Scales: Using Multiple Criteria from Factor Analysis. *Soc. Indic. Res.* 102, 3 (2011), 443–461.

[58] Oddny Judith Solheim. 2011. The Impact of Reading Self-Efficacy and Task Value on Reading Comprehension Scores in Different Item Formats. *Read. Psychol.* 32, 1 (Jan. 2011), 1–27.

[59] Claude Steele. 2011. Stereotype Threat and African-American Student Achievement. In *The Inequality Reader* (2 ed.). Routledge, 276–281.

[60] Kamden K Strunk and Leslie Ann Locke (Eds.). 2019. *Research Methods for Social Justice and Equity in Education*. Palgrave Macmillan.

[61] Burçin Tamer and Jane Stout. 2016. *Recruitment and Retention of Undergraduate Students in Computing: Patterns by Gender and Race/Ethnicity*. Technical Report. Computing Research Association.

[62] Suraj Uttamchandani. 2018. 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).

[63] Cindy M. Walker. 2011. What's the DIF? Why Differential Item Functioning Analyses Are an Important Part of Instrument Development and Validation. *Journal of Psychoeducational Assessment* 29, 4 (Aug 2011), 364–376. DOI: http://dx.doi.org/10.1177/0734282911406666

[64] Margaret Walsh, Crystal Hickey, and Jim Duffy. 1999. Influence of Item Content and Stereotype Situation on Gender Differences in Mathematical Problem Solving. *Sex Roles* 41, 3-4 (Aug. 1999), 219–240.

[65] Max Weber. 1948. *From Max Weber: Essays in Sociology*. Vol. 33. Routledge.

[66] S Christian Wheeler and Richard E Petty. 2001. The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychol. Bull.* 127, 6 (Nov. 2001), 797–826.

[67] Tiffani L Williams. 2020. 'Underrepresented Minority' Considered Harmful, Racist Language. *Commun. ACM* (June 2020).

[68] Carol M Woods. 2009. Evaluation of MIMIC-Model Methods for DIF Testing With Comparison to Two-Group Analysis. *Multivariate Behav. Res.* 44, 1 (Jan. 2009), 1–27.

[69] Carol M Woods, Li Cai, and Mian Wang. 2013. The Langer-Improved Wald Test for DIF Testing With Multiple Groups: Evaluation and Comparison to Two-Group IRT. *Educ. Psychol. Meas.* 73, 3 (June 2013), 532–547.

[70] Benjamin Xie. 2020. How data can support equity in computing education. *XRDS: Crossroads, The ACM Magazine for Students* 27, 2 (Dec. 2020), 48–52.

[71] Benjamin Xie, Matthew J Davidson, Min Li, and Amy J Ko. 2019. 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)*. ACM, 699–705.

[72] Benjamin Xie, Greg L Nelson, Harshitha Akkaraju, William Kwok, and Amy J Ko. 2020. 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)*. ACM. To appear.

[73] Michael Zieky. 1993. Practical questions in the use of DIF statistics in test development. In *Differential Item Functioning*, Paul W Holland and Howard Wainer (Eds.). Erlbaum, 337–347.

[74] Michael Zieky. 2003. *A DIF Primer*. Technical Report. Educational Testing Service.

[75] Bruno D Zumbo. 2007. Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Lang. Assess. Q.* 4, 2 (2007), 223–233.

[76] Bruno D Zumbo and Michaela N Gelin. 2005. 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 (2005), 1–23.

[77] Stuart Zweben and Betsy Bizot. 2019. *Taulbee Survey*. Technical Report. Computing Research Association.