Investigating item bias in a CS1 exam with differential item functioning

Matt J. Davidson
University of Washington
College of Education
Seattle, Washington, USA
mattjd@uw.edu

Brett Wortzman
University of Washington
Paul G. Allen School of Computer Science & Engineering
Seattle, Washington, USA
brettwo@cs.washington.edu

Amy J. Ko
University of Washington
The Information School, DUB Group
Seattle, Washington, USA
ajko@uw.edu

Min Li
University of Washington
College of Education
Seattle, Washington, USA
minli@uw.edu

ABSTRACT

Reliable and valid exams are a crucial part of both sound research design and trustworthy assessment of student knowledge. Assessing and addressing item bias is a crucial step in building a validity argument for any assessment instrument. Despite calls for valid assessment tools in CS, item bias is rarely investigated. What kinds of item bias might appear in conventional CS1 exams? To investigate this, we examined responses to a final exam in a large CS1 course. We used differential item functioning (DIF) methods and specifically investigated bias related to binary gender and year of study. Although not a published assessment instrument, the exam had a similar format to many exams in higher education and research: students are asked to trace code and write programs, using paper and pencil. One item with significant DIF was detected on the exam, though the magnitude was negligible. This case study shows how to detect DIF items so that future researchers and practitioners can do these analyses.

CCS CONCEPTS

• **Social and professional topics → Student assessment**.

KEYWORDS

psychometrics, equity, validity, differential item functioning, CS1

ACM Reference Format:
Matt J. Davidson, Brett Wortzman, Amy J. Ko, and Min Li. 2021. Investigating item bias in a CS1 exam with differential item functioning. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE '21), March 13–20, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3408877.3432397

SIGCSE '21, March 13–20, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8062-1/21/03…$15.00
https://doi.org/10.1145/3408877.3432397

1 INTRODUCTION

Imagine a student from an underrepresented group in CS, excited to see how they did on the final exam for their first university-level CS course. Despite doing well on most of the questions, they got a low score on one of the items. They think about how hard they worked that term, the hours put in to assignments and studying. They may conclude that CS is simply not for them. Unfortunately, their performance on that one question may not reflect their actual understanding of the concepts, but could be a result of unaddressed bias in the question.

The APA/NCME/AERA Standards for educational and psychological testing list fairness as an important consideration for any test, as important as validity and reliability [1]. Fairness is also intrinsically desirable: both teachers and students would like to be using and taking assessments that they believe to be fair. In addition, it is a necessary precondition for a strong validity argument [19]. An item is considered fair to the extent that it primarily measures the construct being assessed and performance is not systematically related to other factors, such as an examinee's gender or race [40]. Any differences in responses that is not due to the targeted knowledge weakens the argument that exam scores indicate understanding of the targeted knowledge.

Item fairness can be assessed operationally by investigating whether different groups of examinees are measured similarly for each item [23]. An unfair item is one that measures systematically differently for groups of examinees who are assumed to have the same underlying true ability. Researchers must carefully consider whether differences in measurement, when detected, are instances of bias. For example, consider an exam measuring knowledge of Java. Examinees with more knowledge of Java will likely score higher, which provides evidence for construct validity. However, if examinees who identify as male, regardless of their knowledge of Java, score higher than examinees who identify as any non-male category, there is bias. This would weaken the exam's validity argument, because there is bias in the scores that is unrelated to the concept being measured.

There have been many calls for more validity work on CS instruments (e.g. Tew & Dorn [37]). And indeed, much validity research has been done on instruments used in the CSEd community. That

research has explored validity through content review with domain experts [28, 38], thinkalouds [18, 28, 38], comparison with a more established measure [5, 7, 10, 28, 38], item statistics [16, 27, 34, 41], and the presentation of formal validity arguments [26] as described by Kane [19]. Test-level fairness has been investigated for some instruments; for example, Decker [10] compared the total scores of a number of different subgroups. However, we found no examples of studies that used item responses to examine bias in items on an exam or instrument.

In the psychometrics community, item fairness is investigated with differential item functioning (DIF) analysis [40]. DIF analysis is an umbrella term for a set of statistical techniques that can be used to compare groups of test-takers by matching on either total test score or estimated ability, and then seeing whether each item measures similarly in all groups [2, 9].

Much of the literature on DIF is focused on the development and exploration of DIF methods. For example, studies use simulations to determine sample size requirements [3] or compare new DIF methods to established ones [21]. Some papers report the the results of DIF analyses, but few move step by step through the analytical process of preparing and analyzing data with DIF methods. One exception is Martinková et al [23], who provide clear descriptions of DIF methods and apply them to real data, but do not explain preliminary steps to ensure that data are suitable for use with the method.

This paper will address the gaps in CS instrument validity studies and accessible literature on DIF by presenting a case study of applying DIF to an exam used in a large CS1 course. By describing the process of preparing the data, choosing DIF methods, and interpreting the results, we strive to lower barriers to applying DIF to CSEd instruments. In the process, we address the following research question: did any items on this CS exam favor students based on either their binary gender or their year of study? Substantive results from the analysis will show the type of fairness and bias issues that DIF analysis can help reveal in CS assessments, as well as what test designers or instructors can and should do with results of DIF analyses.

2 DIFFERENTIAL ITEM FUNCTIONING

All DIF methods work by comparing how two or more groups perform on each item in an exam, after matching examinees on a criterion. That criterion can be the total score on the exam, an item response theory estimated ability, or some other variable. As originally developed, DIF analysis methods compared two groups only; newer methods exist to compare more than two groups (e.g. generalized logistic regression [14, 35]), but they are not yet adopted as widely as two group methods. Regardless, the underlying statistics are similar whether comparing two or more than two groups. In the two group case, the group that we are concerned about items being biased against is called the *focal group*, and the other group is the *reference group*. If bias is found that is related to a higher probability of correct response for one group, that group is said to be "favored" by the item. The motivating question is: *for each item, did examinees from the focal group perform similarly to those of the same ability in the reference group?*

In the rest of this section, we present our data from a CS exam in one large introductory programming course, examining basic psychometric properties, surfacing potential bias, and considering whether our data fit modeling assumptions.

2.1 Data from a CS1 Final Exam

The data for this study are from a final exam for a CS1 course. The course is one quarter (10 weeks) long, taught in Java, and intended for both majors and non-majors. Less than 10% of enrolled students are currently CS majors, though many more may be prospective majors and/or students in related majors. Enrollment is typically between 500 and 1,000 students per quarter, consisting primarily of undergraduates, but with small populations of graduate and non-matriculated students as well. Though the course has no prerequisites and no prior knowledge is assumed, roughly half of students self-report some level of previous programming experience. Students with significant experience are encouraged to enroll directly in the related CS2 course, but some choose to take CS1 anyway.

The final exam we analyzed is a proctored, one hour and fifty minute, written exam given at the end of the course. It is a summative assessment for the entire quarter, but with extra emphasis given to material introduced in the last 40% of the course. The exam has a standard format—three code tracing questions and seven code writing questions—and this format is known to students ahead of time. Students are given access to a database of previous exam questions (without solutions) for practice, and questions from this database are often reused on the exam. Because the database is large and solutions are not provided, it is unlikely students would have memorized a solution to a problem that appears on the actual exam, though they may have completed the same problem during practice or preparation.

The three code tracing items required students to provide specific values based on a given code snippet. The code writing items required students to write code based on a given specification. For this analysis, each value provided on code tracing items was marked correct or incorrect, and the course instructor determined how many correct values indicated sufficient understanding of the concept(s) being assessed. Writing items were graded on a rubric and correctness was similarly determined, meaning that each of the 10 items on the exam was marked as right or wrong.

Choosing focal and reference groups for this study was limited in two ways: first, only certain demographic data was available, and second, focal groups cannot be too small (sample size is discussed in Section 3). Student records provided two different groupings that were of substantive interest and had sufficient sample size: binary gender and year of study. Gender is a complex, fluid aspect of a person's identity, with far more than two discrete possibilities. Despite the limitations of binary gender as a grouping variable, it was chosen because female-identifying students often do not feel they belong in CS [15]. Year of study was chosen because older students may have more experience taking exams or more refined study habits. Year of study was dichotomized as first-year or greater than first-year student, the latter group including graduate and non-matriculated students. The total sample included responses from

939 students. The two focal groups considered were 360 students reported as female, and 246 students beyond first year.

2.1.1 Basic psychometric properties. DIF is a method for item response data, and cannot be used if only total scores are available. The test data must have a score for each examinees' response to each item on the test (though some missing values can be accommodated). In addition, basic psychometric properties of the test and items should be investigated: *reliability*, *difficulty*, and *discrimination*. Reliability indicates how much scores randomly vary: it is a necessary, but not sufficient, condition for any argument about an exam's validity. It is pointless to make validity arguments of any kind unless the exam in question has been demonstrated to measure students reliably. Cronbach's α is often used for reliability.[1] In the classical test theory (CTT) approach [2], item difficulty can be assessed by calculating the proportion of examinees who answered the item correctly. CTT discrimination, or how well an item distinguishes between lower- and higher-scoring examinees, is often assessed by the adjusted item-total correlation, which is the relationship between scores on that item with total scores. For difficulty, lower values mean the item is more difficult, while for discrimination, higher values mean the item distinguishes more sharply. For our CS1 exam, Table 1 reports both, along with α-drop, which shows how α would change if each item were dropped from the exam.

Table 1: Item statistics for a CS1 final exam. CTT difficulty (proportion correct), discrimination (adjusted item-total correlations), and reliability (α-drop) values are reported for each item. Cronbach's α for the overall exam was .77.

	Difficulty	Discrimination	Reliability
RefMyst	0.65	0.50	0.75
ArraySim	0.66	0.45	0.76
InheritMys	0.71	0.40	0.76
switch1	0.28	0.32	0.77
switch2	0.45	0.55	0.74
filter	0.82	0.55	0.75
isFiblike	0.53	0.67	0.73
Critters	0.45	0.44	0.76
delta	0.35	0.55	0.74
numWord	0.33	0.54	0.75

Table 1 shows that items have a variety of difficulties (which is preferable, helping the test discriminate between students with varying abilities), and most item-total correlations are above .40, which is a loose floor for item discrimination [2]. In all cases dropping an item would not increase α, so there is reason to include each in the analysis and in any future administrations of the exam. Acceptable reliability values vary depending on the intended use of test scores [2]. Even within the same intended use, there is disagreement on what values are required [36]. Cronbach's α can be interpreted as a correlation: squaring the value and subtracting it from one gives the percentage of variance unaccounted for. It may be reasonable, then, to argue that reliability must at least be .70, since this would mean that about half of the variance in scores was unexplained. Using that logic, the value of .77 for this exam is therefore acceptable. Because reliability, difficulty, and discrimination are acceptable, it makes sense to move on to doing DIF analysis.

2.1.2 Item Response Theory assumptions. In addition to basic psychometric properties, some DIF methods may involve additional requirements of the data. For example, item response theory (IRT)-based DIF methods require that the exam and responses meet the assumptions of IRT models: *local independence*, *unidimensionality*, and *functional form*. Local independence assumes that an examinee's responses to any two items are independent, except for that examinee's underlying ability. This can be assessed by analyzing the test design; for example, if examinees read a passage and respond to a set of questions about that passage, responses to those items may be dependent on some property of the passage. Unidimensionality is the assumption that item responses are largely driven by a single underlying skill or trait. This can be examined with factor analysis [33]. Functional form is the assumption that the data follow whatever function is specified by the IRT model. For example, some IRT models include an estimate for guessing: this might make sense for data collected from a multiple choice exam, but might not for free response items. This assumption can be assessed by considering the exam format and model parameters, as well as by examining model fit indices, with good model fit providing evidence that the functional form is acceptable for the data.

For this final exam, local independence likely holds since there were no items that had shared passages or prompts. Unidimensionality was examined with exploratory and confirmatory factor analysis. Exploratory analysis suggested a single factor according to the eigenvalues > 1 criterion [33], and a confirmatory model was fit using the lavaan package [31] in *R* [29]. With a single common factor the model fit was excellent, with RMSEA = .036, CFI = .97, TLI = .96.[2] The functional form assumption was examined by fitting three IRT models to the data with one (1PL), two (2PL), and three parameters (3PL). The 1PL assumes that all items have the same discrimination value but vary in difficulty; the 2PL assumes items vary in discrimination and difficulty; the 3PL is a 2PL with a parameter for guessing. Model fit was compared using the Bayes Information Criterion (BIC), which can be used to compare models applied to the same data. The lowest value of BIC indicates the best fitting model. The 1PL had the highest BIC at 10605.06, the 3PL next highest at 10601.99, and the lowest was the 2PL with 10533.54. This shows that the 2PL is the best fitting of the three; the 2PL also showed excellent fit, with RMSEA = .022, CFI = .99, TLI = .99, which suggests that the functional form is a good fit to the data.

3 CHOOSING A DIF METHOD

Having checked the basic psychometric properties and IRT assumptions, we can now consider what DIF methods to use. There are at least three considerations when choosing DIF methods: how items

[1] Although commonly used, Cronbach's α is almost always an *underestimation* of reliability for a test. This is because many tests do not meet the assumptions inherent in the α calculation [39]. A good alternative is McDonald's ω or Guttman's lower bound, also known as *glb*. Further discussion of the limitations of α and alternatives can be found in Sijtsma [32] and a response to Sijtsma by Revelle and Zinbarg [30].

[2] These are commonly reported fit measures for confirmatory factor models. The values all point to excellent fit of a one dimensional model to the data. More on model fit indices and values can be found in Kline [20].

were scored, type(s) of DIF to be detected, and sample size. In general, items can be scored either dichotomously (right or wrong) or polytomously (e.g. with partial credit), and different DIF methods are suited to each response type. In addition, two types of DIF are possible: *uniform* and *non-uniform*. Uniform DIF is when all the examinees of one group are favored on an item. Non-uniform DIF exists when the group that is favored changes based on the total score. For example, reference group members who score below the average are favored, while those scoring above average are not favored. The final consideration is sample size. A contingency table-based DIF method, like the Mantel-Haenszel statistic, generally has lower sample size requirements, but can only reliably detect uniform DIF [2, 9]. The logistic regression method works well in detecting both DIF types with as few as 200 examinees per group [2], while IRT-based methods generally require larger samples, around 700 examinees total and about 300 in the focal group [44]. Recent research has found that samples as small as 25 to 50 per group may be enough to find uniform DIF for both the logistic regression method and IRT-based methods [3].

We chose to use both the logistic regression and likelihood ratio test methods, for a few reasons. First, we wanted to be able to detect both uniform and non-uniform DIF. In addition, we needed a method that worked for dichotomous item responses. We dichotomized the answers because polytomous DIF methods, and polytomous IRT models, involve estimating numerous additional parameters, which would both require a larger sample and increase the chances of overfitting the data. Finally, we wanted to use two different methods: a DIF item detected by multiple methods is more likely to truly be a DIF item.

4 ANALYSIS

This section will describe the DIF methods we applied to the CS1 final exam data. We begin by explaining likelihood ratio tests in general, and then provide details about how we implemented our chosen DIF methods.

4.1 Likelihood ratio tests

Likelihood ratio tests (LRTs) can be used to compare how well two models fit the same data. The likelihood of a model with more parameters is compared to a model with fewer parameters to see whether the additional parameters significantly improve the fit. The ratio of likelihoods for the two models is distributed as a χ^2 statistic with degrees of freedom equal to the difference in the number of parameters between the two models. If that χ^2 test is significant, it indicates that the additional parameter(s) significantly improves model fit. Typically in the DIF context, the difference between the the two models being compared is a variable for group membership. If the ratio exceeds the critical value, then the model including group membership fits better, suggesting that group membership is a significant predictor of score.

4.2 Logistic regression DIF

Logistic regression DIF methods match participants based on their total scores, and test whether the relationship between overall score and the probability of a correct answer is the same for both groups [42]. First, we estimated a baseline model for each item with the probability of correct response based only on total exam scores. Then we estimated two separate models to detect uniform and non-uniform DIF. For uniform DIF, we added a variable to indicate binary gender or year of study, and for non-uniform DIF, we also added an interaction term that interacts group and total score. Each model is then compared to the baseline model using a LRT. Because the models differ from the baseline model in one parameter only, if either fits the data significantly better than the baseline model, then that item is potentially biased.

After finding evidence of significant DIF for an item, we examined the effect size of the DIF. Most items show some amount of DIF, and sometimes even statistically significant DIF can have a small effect size. For logistic regression DIF, we used the Jodoin and Gierl [17] effect size, based on the change in Nagelkerke's R^2, an R^2 statistic for logistic regression. Like other effect sizes, the authors provide guidelines for what constitutes a negligible, moderate, and large effect of DIF.

4.3 Likelihood ratio test DIF

LRT DIF is an IRT-based DIF detection method, introduced by Thissen, Steinberg, and Wainer [8]. LRT DIF uses LRTs to compare IRT models: the hypothesis being tested is whether the item parameters for a given item, like discrimination and difficulty, are the same for each group. We estimated a baseline model that constrained the item parameters to be the same for both groups (i.e. assuming no DIF), and a second model that allowed parameters to be different for each group (i.e. assuming DIF). If the LRT is significant, there is evidence of DIF for the item being tested. For example, when testing for binary gender DIF, we fit a baseline model that assumes the difficulty is the same for both binary genders, while a second model allowed the difficulty to be estimated separately. If the second model fits the data significantly better than the first model, meaning that the item's difficulty was different for each binary gender, then there is evidence of uniform DIF for that item.

For LRT DIF there is not a standard effect size measure. Meade [24] has proposed a taxonomy of effect sizes for LRT DIF, which uses item parameters to calculate changes in expected scores as a result of DIF. We use two of those measures: signed in-sample differences (SIDS) and unsigned in-sample differences (UIDS). SIDS is the average difference in score for examinees in the focal group, which allows non-uniform effects to cancel out. It represents how much scores change as a result of DIF, on average. UIDS is the difference in expected score if DIF favored a single group across the whole ability scale, meaning it summarizes the magnitude of DIF both in favor of and against the focal group. These effect sizes can be interpreted similarly to the standardized mean difference in probabilities in Dorans and Kulick [11], who provide guidelines for negligible, possible, and large effects.

5 RESULTS

To illustrate the use of these DIF methods, we analyzed responses from the CS1 final exam we described in Section 2.1 to see whether items were biased based on examinees' binary gender or year of study. This section will present and interpret results from logistic regression and LRT DIF methods using that response data. All analyses were conducted using the R statistical software [29]. Logistic

regression DIF was run with *difR* package [22], while LRT DIF used the *mirt* package [6]. Results are discussed first for binary gender, and then year of study.

The chosen DIF methods compare two models for each of the 10 items on the exam. Any situation where multiple comparisons are made requires that p-values be made more conservative, to decrease the chances of false positives. Therefore a p-value adjustment method must be specified, and we chose the Benjamini-Hochberg adjustment [4] because it provides the most statistical power.

5.1 Binary gender

Results for binary gender DIF are presented in Table 2. To see whether any items displayed DIF when comparing binary gender groupings, the χ^2 tests should be examined. Because none of the tests are significant (i.e. there are no p-values below .05), no items were found to exhibit binary gender-based DIF. Note that, despite not finding evidence of DIF, items do not measure exactly the same for each binary gender: if items measured exactly the same, the χ^2 tests would all be zero. This highlights the importance of establishing (and controlling) a significance criterion.

Table 2: DIF Results for Gender. Because no DIF items were found, effect sizes are not reported.

	Logistic Regression				LRT			
	Uniform		Non-uni		Uniform		Non-uni	
	χ^2	p	χ^2	p	χ^2	p	χ^2	p
RefMyst	2.16	.26	0.85	.46	0.81	.61	0.19	.74
ArraySim	0.39	.59	0.70	.46	0.01	.91	0.79	.68
InheritMys	2.36	.26	1.11	.46	0.26	.68	1.69	.68
switch1	2.82	.26	1.82	.46	3.66	.21	0.60	.68
switch2	3.02	.26	1.79	.46	3.47	.21	0.50	.68
filter	0.55	.59	0.31	.58	0.33	.68	0.26	.74
isFiblike	2.03	.26	0.67	.46	2.34	.32	0.01	.92
Critters	4.78	.26	1.43	.46	3.48	.21	1.58	.68
delta	0.41	.59	2.39	.46	0.96	.61	4.39	.36
numWord	0.00	.99	1.16	.46	0.63	.61	0.77	.68

5.2 Year of study

Table 3 shows results from the logistic regression and LRT DIF methods for year of study. The logistic regression results showed no evidence of uniform DIF, since none of the χ^2 tests have p-values less than .05. There was some evidence of borderline non-uniform DIF for the **isFiblike** item (indicated by the p-value of .08). Although this suggests possible non-uniform DIF, it is unlikely the item would need to be revised or removed because the effect size is "negligible", according to the Jodoin and Gierl guidelines [17].

The right of Table 3 shows results from the LRT method. Recall that this method compared IRT-estimated item parameters across groups: the estimated parameters (difficulty and discrimination) are included in the table, along with the associated χ^2 test. Results from the LRT DIF procedure are similar to those for logistic regression. A notable difference is that LRT method found that *isFiblike* did exhibit non-uniform DIF, indicated by the significant χ^2 value.

Both the SIDS and UIDS effect sizes indicate expected change in scores on each item for beyond first year students. Because items were scored dichotomously, the scores are either zero or one. SIDS therefore shows that greater than first year students had on average a .038 lower score on *isFiblike* than first-year students with identical ability. This negligible effect is consistent with no uniform DIF being found for this item. UIDS shows the non-uniform effect by calculating expected change in scores if greater than first year students were favored across the entire ability scale. It shows that there is an average difference of .099 in scores for *isFiblike*, which is a borderline value that indicates the item should be inspected for a possible effect [11], which we do in Section 6.

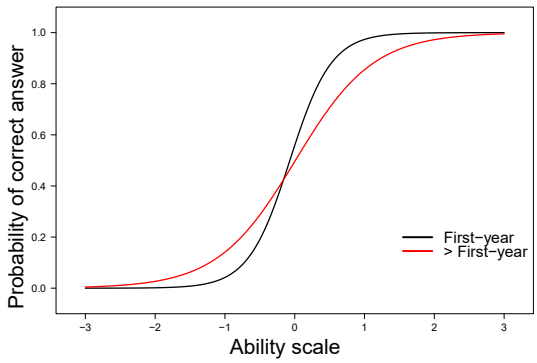

Figure 1: Item traces for the *isFiblike* item.

Figure 1 shows the *item traces* for *isFiblike*. The horizontal axis is the ability scale (with average ability at zero), and the vertical axis is the probability of a correct response for an examinee at that point on the ability scale. The IRT estimated difficulty of the item is the point on the x-axis where the probability of a correct response is 50%. There is one trace for first-year students, and one trace for beyond first year. The non-uniform nature of the DIF can be seen because the traces cross. This shows that *isFiblike* discriminates more sharply between ability levels for first-year students than for beyond first-year. It also indicates that *isFiblike* was easier for beyond first-year students with lower estimated ability, and more difficult for those with higher estimated ability.

6 DISCUSSION

Zumbo [43] argues that, when looking for explanations of DIF, we should take into account the testing situation as well as aspects of the items, and how both contribute something to scores that might not be relevant to the underlying ability of interest. Depending on the grouping variable(s) used, DIF items may indicate an opportunity for learning, rather than inherent bias within the items. It is crucial that DIF results be interpreted in the context of the test, test-takers, and score use.

For this CS exam, *isFibLike* required students to write a Java method to traverse a one-dimensional array of integers and inspect overlapping sets of three consecutive elements (i.e. elements 0, 1, and 2; elements 1, 2, and 3; etc.), checking if each element was equal to the sum of the previous two. The final result of the method was a boolean indicating whether or not all elements had the desired property. The problem was intended to assess students' ability to

Table 3: Year of study DIF results. ΔR^2 is the change in Nagelkerke's R^2, with values from 0 to 0.035 "negligible", 0.035 to 0.07 "moderate", and .07 to 1 "large" DIF effects [17]; SIDS and UIDS values from 0 to .05 are negligible, .05 to .10 intermediate, and > .10 large effects [11]; 1st yr is for first-year students, while > 1st yr is for students beyond first year.

	Logistic Regression						Likelihood Ratio Test (LRT)									
	Uniform			Non-Uniform			Uniform				Non-Uniform				Effect Sizes	
							Difficulty		Sig. Test		Discrimination		Sig. Test			
	χ^2	p	ΔR^2	χ^2	p	ΔR^2	1st yr	> 1st yr	χ^2	p	1st yr	> 1st yr	χ^2	p	SIDS	UIDS
RefMyst	0.48	.86	.0005	2.24	.47	.0022	-0.58	-0.82	0.00	.99	1.39	0.97	2.51	.38	0.01	0.05
ArraySim	4.36	.37	.0047	0.92	.85	.0010	-0.79	-0.53	4.91	.27	1.18	0.95	0.92	.63	-0.07	0.07
InheritMys	0.35	.86	.0004	0.07	.88	.0001	-1.14	-0.95	0.18	.99	0.92	1.02	0.18	.96	-0.02	0.02
switch1	0.04	.86	.0000	0.02	.88	.0000	1.29	1.27	0.01	.99	0.84	0.83	0.00	.99	0.003	0.003
switch2	0.85	.86	.0007	2.17	.47	.0020	0.23	0.14	0.04	.99	1.59	2.24	2.50	.38	0.02	0.06
filter	0.12	.86	.0001	0.24	.88	.0003	-1.14	-1.06	1.34	.95	2.29	3.22	1.53	.54	0.01	0.03
isFiblike	**1.23**	**.86**	**.0009**	**6.95**	**.08**	**.0052**	**-0.07**	**0.01**	**0.91**	**.95**	**3.36**	**1.79**	**8.25**	**.04**	**-0.038**	**0.099**
Critters	0.03	.86	.0000	0.06	.88	.0001	0.25	0.29	0.05	.99	1.13	1.09	0.02	.99	-0.01	0.01
delta	0.24	.86	.0002	0.17	.88	.0002	0.52	0.54	0.77	.95	1.82	2.23	0.77	.63	-0.02	0.03
numWord	0.07	.86	.0001	0.09	.88	.0001	0.58	0.64	0.01	.99	1.96	1.85	0.06	.99	-0.01	0.01

traverse an array, compare nearby elements, and track the combined truth or falsehood of a condition across all elements of an array.

The steeper slope of the trace for first-year students in Figure 1 shows that *isFiblike* provided more information about the ability of first-year students than it did for beyond first-year students. As for why, we can offer only reasoned speculation. The concept measured in this problem was somewhat sophisticated, and it may have been more difficult for students to understand what their solution should do (as opposed to how they should do it). Beyond first-year students may have compensatory factors, like more experience taking exams or skills for reading and understanding complex prompts. As a result, performance of first-year students may have been more directly impacted by their underlying ability, leading to sharper discrimination. Additional data and analysis could shed light on this result, like grouping students by prior exam experience. Thinkaloud interviews could also reveal some required skill or knowledge for a correct answer that had been overlooked.

Because the magnitude of the DIF for *isFiblike* was negligible, no revisions would need to be made to address bias related to binary gender or year in school. What to do with DIF items depends on the intended use of the assessment. For a research instrument, any items with significant DIF should be thrown out or revised, and new items written and tested for to ensure that no items have DIF. For a course exam there may be more options. If exam items go into a bank of items to be reused, DIF items should be revised or not put into the bank. An instructor may also choose to throw out DIF items when grading the exam. Looking at how DIF shows up in items over multiple administrations of similar exams may also reveal patterns to item bias that may call into question the (re)use of certain item types.

One reason that DIF may not have been found in this case is that the sample was not random. The students who happened to take this CS1 course may or may not be representative of the population of CS1 students. Therefore, it is possible that these results would not replicate with another term's sample of CS1 students. If random samples are not available for DIF analysis, then DIF should continue to be checked with each sample of students. On the other hand, if an exam can be administered to a random sample of students, any DIF items will likely also be DIF items with another sample.

7 DIF IN CS EDUCATION

If, as a community, we care about fairness and establishing strong arguments for the validity of our assessment instruments, DIF analysis must become a standard component of validation research. When DIF methods are applied to large enough samples, they provide rich and actionable data about whether and how items are biased for subgroups of students. However, DIF methods require experience and statistical knowledge to use and interpret. This paper seeks to address that with this analysis of a CS1 final exam. Our hope is that making the methods accessible will lead to wider adoption of DIF analyses as standard practice for validation work on instruments for research. While also applicable to course exams, additional tools will likely be required to facilitate DIF analysis by instructors. Future work should apply DIF to widely used instruments and questions, such as AP CS A (an exam with demonstrated differential results for female [13] and non-white [12] students), the SCS1 [28], as well as promising newly developed instruments like the Programming Self-Efficacy survey [34].

Finding bias in exam items is important not only so that we can be confident that scores are accurate measures of the targeted knowledge or skills. As Messick [25] argues: "These issues are critical for...all educational and psychological assessment – because validity, reliability, comparability, and fairness are not just measurement principles, they are *social values* that have meaning and force outside of measurement." We must do whatever we can to ensure that our assessments are aligned with our values, and DIF is one important way that we can do that work.

8 ACKNOWLEDGEMENTS

The authors thank anonymous reviewers for their valuable comments and suggestions. This material is based upon work supported by the National Science Foundation under Grant No. 1735123.

REFERENCES

[1] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for educational and psychological testing.* American Educational Research Association.

[2] Deborah L. Bandalos. 2018. *Measurement theory and applications for the social sciences.* Guilford Press.

[3] William C. M. Belzak. 2019. Testing Differential Item Functioning in Small Samples. *Multivariate Behavioral Research* (Oct 2019), 1–26. https://doi.org/10.1080/00273171.2019.1671162

[4] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[5] Ryan Bockmon, Stephen Cooper, Jonathan Gratch, and Mohsen Dorodchi. 2019. (Re)Validating Cognitive Introductory Computing Instruments. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education - SIGCSE '19.* ACM Press, 552–557. https://doi.org/10.1145/3287324.3287372

[6] R. Philip Chalmers. 2012. mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software* 48, 6 (2012), 1–29. https://doi.org/10.18637/jss.v048.i06

[7] Nick Cheng and Brian Harrington. 2017. The Code Mangler: Evaluating Coding Ability Without Writing any Code. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education - SIGCSE '17.* ACM Press, 123–128. https://doi.org/10.1145/3017680.3017704

[8] Howard Wainer David Thissen, Lynne Steinberg. 1988. Use of item response theory in the study of group differences in trace lines. *Test validity* (1988), 147.

[9] R. J. De Ayala. 2009. *The theory and practice of item response theory.* Guilford Press.

[10] Adrienne Decker. 2007. *How Students Measure Up: an Assessment Instrument for Introductory Computer Science.* Ph.D. Dissertation. State University of New York.

[11] 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.

[12] Barbara Ericson, Shelly Engelman, Tom McKlin, and Ja'Quan Taylor. 2014. Project Rise up 4 CS: Increasing the Number of Black Students Who Pass Advanced Placement CS A. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education* (Atlanta, Georgia, USA) *(SIGCSE '14).* Association for Computing Machinery, New York, NY, USA, 439–444. https://doi.org/10.1145/2538862.2538937

[13] Barbara J. Ericson, Miranda C. Parker, and Shelly Engelman. 2016. Sisters Rise Up 4 CS: Helping Female Students Pass the Advanced Placement Computer Science A Exam. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education* (Memphis, Tennessee, USA) *(SIGCSE '16).* Association for Computing Machinery, New York, NY, USA, 309–314. https://doi.org/10.1145/2839509.2844623

[14] W. Holmes Finch. 2016. Detection of Differential Item Functioning for More Than Two Groups: A Monte Carlo Comparison of Methods. *Applied Measurement in Education* 29, 1 (Jan 2016), 30–45. https://doi.org/10.1080/08957347.2015.1102916

[15] Google and Gallup. 2015. Images of Computer Science: Perceptions Among Students, Parents, and Educators in the U.S. (2015).

[16] Geoffrey L. Herman, Craig Zilles, and Michael C. Loui. 2014. A psychometric evaluation of the digital logic concept inventory. *Computer Science Education* 24, 4 (Oct 2014), 277–303. https://doi.org/10.1080/08993408.2014.970781

[17] Michael G Jodoin and Mark J Gierl. 2001. Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied measurement in education* 14, 4 (2001), 329–349.

[18] Lisa C. Kaczmarczyk, Elizabeth R. Petrick, J. Philip East, and Geoffrey L. Herman. 2010. Identifying student misconceptions of programming. In *Proceedings of the 41st ACM technical symposium on Computer science education - SIGCSE '10.* ACM Press, 107. https://doi.org/10.1145/1734263.1734299

[19] Michael T. Kane. 2013. Validating the Interpretations and Uses of Test Scores: Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement* 50, 1 (Mar 2013), 1–73. https://doi.org/10.1111/jedm.12000

[20] Rex B Kline. 2015. *Principles and practice of structural equation modeling.* Guilford publications.

[21] Sunbok Lee. 2017. Detecting Differential Item Functioning Using the Logistic Regression Procedure in Small Samples. *Applied Psychological Measurement* 41, 1 (Jan 2017), 30–43. https://doi.org/10.1177/0146621616668015

[22] D. Magis, S. Beland, F. Tuerlinckx, and P. De Boeck. 2010. A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods* 42 (2010), 847–862.

[23] 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. *CBE—Life Sciences Education* 16, 2 (Jun 2017), rm2. https://doi.org/10.1187/cbe.16-10-0307

[24] Adam W. Meade. 2010. A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology* 95, 4 (Jul 2010), 728–743. https://doi.org/10.1037/a0018966

[25] Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist* 50, 9 (1995), 741.

[26] Greg L. Nelson, Andrew Hu, Benjamin Xie, and Amy J. Ko. 2019. Towards validity for a formative assessment for language-specific program tracing skills. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research.* ACM, 1–10. https://doi.org/10.1145/3364510.3364525

[27] Thomas H. Park, Meen Chul Kim, Sukrit Chhabra, Brian Lee, and Andrea Forte. 2016. Reading Hierarchies in Code: Assessment of a Basic Computational Skill. In *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education - ITiCSE '16.* ACM Press, 302–307. https://doi.org/10.1145/2899415.2899435

[28] Miranda C. Parker, Mark Guzdial, and Shelly Engleman. 2016. Replication, Validation, and Use of a Language Independent CS1 Knowledge Assessment. In *Proceedings of the 2016 ACM Conference on International Computing Education Research - ICER '16.* ACM Press, 93–101. https://doi.org/10.1145/2960310.2960316

[29] R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[30] William Revelle and Richard E. Zinbarg. 2009. Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika* 74, 1 (Mar 2009), 145–154. https://doi.org/10.1007/s11336-008-9102-z

[31] Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2 (2012), 1–36. http://www.jstatsoft.org/v48/i02/

[32] Klaas Sijtsma. 2009. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika* 74, 1 (Mar 2009), 107–120. https://doi.org/10.1007/s11336-008-9101-0

[33] Suzanne L. Slocum-Gori and Bruno D. Zumbo. 2011. Assessing the Unidimensionality of Psychological Scales: Using Multiple Criteria from Factor Analysis. *Social Indicators Research* 102, 3 (Jul 2011), 443–461. https://doi.org/10.1007/s11205-010-9682-8

[34] Phil Steinhorst, Andrew Petersen, and Jan Vahrenhold. 2020. Revisiting Self-Efficacy in Introductory Programming. In *Proceedings of the 2020 ACM Conference on International Computing Education Research.* 158–169.

[35] Dubravka Svetina and Leslie Rutkowski. 2014. Detecting differential item functioning using generalized logistic regression in the context of large-scale assessments. *Large-scale Assessments in Education* 2, 1 (Dec 2014), 4. https://doi.org/10.1186/s40536-014-0004-5

[36] Keith S. Taber. 2018. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education* 48, 6 (Dec 2018), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

[37] A. E. Tew and B. Dorn. 2013. The Case for Validated Tools in Computer Science Education Research. *Computer* 46, 9 (2013), 60–66.

[38] Allison Elliott Tew and Mark Guzdial. 2011. The FCS1: a language independent assessment of CS1 knowledge. In *Proceedings of the 42nd ACM technical symposium on Computer science education - SIGCSE '11.* ACM Press, 111. https://doi.org/10.1145/1953163.1953200

[39] Italo Trizano-Hermosilla and Jesús M. Alvarado. 2016. Best Alternatives to Cronbach's Alpha Reliability in Realistic Conditions: Congeneric and Asymmetrical Measurements. *Frontiers in Psychology* 7 (May 2016). https://doi.org/10.3389/fpsyg.2016.00769

[40] 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. https://doi.org/10.1177/0734282911406666

[41] Benjamin Xie, Matthew J. Davidson, Min Li, and Andrew 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 Press, 699–705. https://doi.org/10.1145/3287324.3287370

[42] Bruno D Zumbo. 1999. *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* 57 pages.

[43] Bruno D Zumbo. 2007. Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language assessment quarterly* 4, 2 (2007), 223–233.

[44] Rebecca Zwick. 2012. A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series* 2012, 1 (2012), i–30.