Investigating Distractors with DIF in a College-Level Introductory Computing Assessment

Sin Yu Ciou

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

Master of Education

University of Washington

2024

Committee:

Min Li

Chun Wang

Program Authorized to Offer Degree:

Education

University of Washington

Abstract

Investigating Distractors with DIF in a College-Level Introductory Computing Assessment

Sin Yu Ciou

Chair of the Supervisory Committee:

Min Li

Department of Education

Assessing learning at scale often relies on multiple-choice exams, which consist of a stem followed by a correct option and several distractors. While careful design of all parts of a question is essential to ensure assessment validity, each component, including distractors, can introduce bias and unfairly impact specific groups of students. This study extends Differential Item Functioning (DIF) analysis to investigate bias in distractors within an introductory computing assessment for undergraduate students. Using Differential Distractor Functioning (DDF) analysis on responses from 259 students to a computer science introductory level assessment, we identified problematic distractors that exhibited differential performance patterns for male and female students. Our finding provides insights into potential biases within assessments, advancing efforts to create more equitable measures of learning for all students.

Keywords: differential distractor functioning, gender DIF, distractors analysis, assessment validity, computing assessment, assessment bias

Investigating Distractors with DIF in a College-Level Introductory Computing Assessment

The lack of female representation in computing professions has been consistently considered one critical area for improving the broader participation of all individuals in STEM fields. For example, the proportion of women with bachelor's degrees in computer science has declined in the last two decades despite a slight increase at master's and doctoral levels. Still, women are under-represented at both graduate and graduate levels (National Science Foundation 2017).

Several researchers try to associate gender differences in computer self-efficacy, stereotypes, interests, values, and classroom experiences with the disparity between genders in entering the CS field. Eccles' expectancy-value theory provides one of the most comprehensive theoretical gender differences in the STEM path selection (e.g., Eccles, 1994, 2005; Wigfield & Eccles, 2000). Drawing from that theory, women are less likely to enter math-intensive professional fields due to relatively lower math and science expectancies and values than men. Further, women tend to hold lower self-efficacy in male-dominated domains regardless of their actual abilities, skills, or performance (Jagacinski, 2013). Other explanatory theories include prior experiences that state women have less programming experience, are often exposed to computers at an older age, and express interest in CS later than men (Schmidt, 2011), perception of CS major that views CS or CS-related majors as competitive and male-dominated (Lewis et al., 2016), and CS courses that are repeated reported as less supportive and inclusive for women (Astin & Astin, 1992; Redmond et al., 2013). Along with those factors (e.g., stereotype threat, high school preparation, or classroom practices), the use of equitable assessments (or failing to do so) is also an influential factor.

The purpose of Differential Item Functioning (DIF) procedures is to identify potential biases in assessment items that may disadvantage certain groups of students based on characteristics such as gender, ethnicity, or socioeconomic background. By detecting biased items, researchers and educators can ensure that assessments fairly measure the intended constructs for all students, fostering equity in education. DIF analysis plays a critical role in improving the validity and reliability of assessments, particularly in STEM fields where equitable evaluations remain underdeveloped. Furthermore, it is insufficient to fully ensure equity only at the item level since this does not look into biases within options. For instance, one option in a multiple-choice item may favor one group over another, even if the overall item does not show DIF. To ensure true equity, it is essential to extend the analysis to the option level, which provides deeper insights and stronger evidence that the assessment is unbiased, ultimately supporting fairness for all learners.

This thesis study examines the fairness of introductory computer science assessments by employing DIF procedures to identify and analyze potentially biased items and distractors. Specifically, it investigates whether any systematic biases exist in item distractors across genders.

The first section presents the theoretical background, focusing on CS assessments and potential flaws that may be partially attributable to gender differences in assessment performance. I then continue to provide an overview of the differential distractor functioning (DDF) methods in relation to regular DIF to flag potentially biased options in assessment items. Afterward, the source of data and procedures of this DDF study are presented. After that, empirical findings are provided, followed by discussions about limitations and future research.

Literature Review

Research on Assessing Computer Science Learning

The assessment of student learning is an essential component of any educational subject. High-quality assessment instruments are available in many STEM fields, providing researchers and educators with tools to accurately measure student learning and assess curricular innovations (Tew & Guzdial, 2010). One key aspect of assessment validation is fairness, ensuring that assessments are equitable and accessible to all students, regardless of their backgrounds; students have equal opportunity to demonstrate their understanding, conditional on their ability.

Consistent with STEM assessment research, multiple studies have documented the gender gap in CS. Sibia et al. (2024) found that male students and those with prior programming experience generally exhibited greater confidence in their potential for success and faced fewer perceived obstacles. In contrast, female students and those lacking programming experience were more likely to express concerns about their intellectual abilities and view challenges less positively. Cheek (1995) brought attention to the gendered aspect of computer-based science exams, suggesting that these tests might not be gender-neutral and might adversely affect underrepresented populations. For example, males tend to show greater interest in non-academic computer use, while females are influenced by social and collaborative environments, such as single-sex settings or using computers with friends. Furthermore, females tend to use computers as tools for tasks, whereas males often explore them for their own sake.

In contrast to the reported gender gap, the application of DIF in CS education remains underexplored. Davidson et al.'s study was the first to introduce DIF to the computer science education community, showcasing how psychometric methods can be effectively applied in this field. Two notable studies in computing education research were inspired by Davidson et al.'s

work. Xie et al. (2019) analyzed data from 19,617 students on Code.org, employing DIF analysis by race and gender. They identified six items favoring male students over other groups. In contrast, Na et al. applied DIF analysis to support the validity of an early childhood computational thinking assessment. They found no evidence of DIF related to gender or age across the items. However, the study confirmed gaps in the latent ability of computational thinking skills, with boys and older children demonstrating higher levels of ability. Furthermore, Paker et al. (2024) examined DIF across intersecting student identities in SCS1 responses and identified problematic items that may favor the non-white male group. These studies demonstrated the value of DIF analysis in identifying biases and promoting fairer assessments in computer science education.

Although those studies empirically evaluated CS assessments at a range of grade levels, there are a limited amount of studies in CS DIF research in comparison to other STEM disciplines. Additionally, DIF at the item level only indicates which item may cause inequity issues, without adequate information on how distractors behave, it will be rather difficult for designers to know which part of the item needs further revision. To fill those two gaps, this study conducted DIF at the distractor level to further pinpoint which options showed bias against which groups conditional on their ability, aiming at offering greater details for item revision.

Importance of Assessment Fairness

In educational assessments of any kind, ensuring equitable opportunity for students to demonstrate their knowledge is a key focus for test developers. The notion of equitable opportunity can be linked to the concept of "fairness." However, determining fairness in education assessments is difficult (Gipps & Stobart, 2009). Zwick (2019) discusses the term fairness regarding biased measurements. In educational measurement, bias occurs when an item

is consistently easier (or more difficult) for a particular subgroup. Biased assessment items may produce incorrect measurements of a student's ability, which affects the overall fairness of the assessment and, consequently, adversely impacts any conclusions made using the assessment.

When developing a high-quality assessment, test designers should consider to what degree a test item is affected by bias or impact. Item bias here refers to the misspecification of the latent variable, where multiple latent abilities are measured instead of the intended one, introducing noise or error into the measurement. Essentially, item bias occurs when the differential functioning of the item is unrelated to the test's purpose or interpretation of the measure; in other words, the item assesses a latent construct other than the intended. In contrast, item impact occurs when the differential performance pattern between two groups of examinees reflects the true difference in the underlying ability between the groups. In the latter case, item impact occurs when groups differ in their performance on the skills or understanding that the items are intended to measure. This is an expected outcome of many assessments and should not be seen negatively.

Differential Item Function

Differential Item Functioning (DIF) occurs when the performance of an item, after conditioning for ability, is different for multiple groups. Consider a multiple-choice assessment where we are only concerned with item correctness. An intuitive method for determining if an item functions differently across subgroups is to compare the proportion of individuals within each group that answered the item correctly. By matching members from each group on ability level and comparing item performance, we can determine if that item favors (or is easier) a particular subgroup. The purpose of matching groups on ability level is to control for ability level so that any observable difference in performance must be attributed to something other than

ability. If substantial and consistent differences exist between a reference and focal group across the latent ability continuum, we can conclude that the item functions differently between groups and that DIF exists (Penfield & Lam, 2000). Typically, the two groups are referred to as the reference group and the focal group; the reference group is anticipated to show the expected response pattern, while the focal group represents those who might be influenced by DIF in certain items. As for DIF result interpretation, there are two types of DIF based on how an item favors certain groups: uniform DIF and non-uniform DIF. Uniform DIF assumes an item is consistently easier or harder for all individuals in one group compared to another, while non-uniform DIF suggests that the level of difficulty varies based on an individual's overall performance.

Differential Distractor Functioning

Differential Distractor Functioning (DDF) is concerned with group differences in the option selection as an extension to DIF, which focuses on group differences surrounding incorrect and correct responses. DIF deals with the between-group difference in the probability of correct response conditioning on ability, focusing only on the correct response. By focusing only on the correct answer, the assumption is that if DIF exists, there must be a between-group difference in the probability of selecting one or more distractors. In a DDF analysis, however, we examine whether a group-level difference exists in the probability of selecting a particular distractor after conditioning on ability (Ioannis et al., 2018).

Consider a multiple-choice assessment where each item has one correct answer. The remaining plausible answer choices are incorrect options, serving as distractors. Group-level differences in distractor choice will not affect the overall test scores since all distractors are wrong answers. Therefore, when conducting a DIF analysis, distractor choice is not considered.

However, information about how students make sense of the item responses can be obtained by studying what distractor is preferred by different groups of students. Such an analysis is pivotal in evaluating the degree to which items may function differentially across multiple groups with respect to selecting distractors. In other words, distractors can function differentially across groups when they are more appealing to test takers from a given group, thereby being considered biased and should be examined further (Greene et al., 1989).

Methods for Conducting DDF Analysis

Mapuranga and colleagues (2008) classified these methods into four categories: (a) generalized linear model methods, (b) expected item score methods, (c) item response theory (IRT) methods, and (d) nonpara-metric odds-ratio methods. In the first category, *generalized linear model methods*, involves procedure based on *logistic regression* approach, introduced by Abedi, Leon, and Kao (2008). With this approach, the responses were first categorized into two groups: one for individuals who chose the most common distractor and another combining students who selected one of the two or three less frequent distractors. A standard logistic regression analysis was then performed on the test scores. Kato et al. (2009) expanded previous methods by applying multi-step multinominal logistic regression to assess both DIF and DDF effects at the same time. This approach models the probabilities of all response options, including item key and distractors for each test item as functions of an ability proxy. This provides a comprehensive view of how all response options, both correct and incorrect, behave for each item. The strength of multinominal logistic regression lies in its ability to quantify the amount of DIF/DDF, both overall and for individual response options of every item, while also visualizing item characteristics. Thus, multinominal regression was selected as the first method for analyzing the DDF effect in this study.

The second method selected in this study was Nominal Response Model (NRM) (Bock, 1972). It falls under the third category IRT framework. In this category, DDF methods utilize a latent variable in the definition of the null DIF (Koon & Kamata, 2013). One method in this category uses the mixture item response model, introduced by Bolt, Cohen, and Wollack (2001), to examine DDF for each distractor. Nominal Response Model (NRM) (Bock, 1972) functions similarly to the mixture model but includes an additional parameter to account for DDF effects associated with each distractor. The NRM model estimates the probability of choosing each distractor, without considering the order of the options, and generates item parameters for each distractor by applying regular DIF analysis.

The third method included in this study, GPCM-lasso (Schauberger & Mair, 2020), falls into the IRT framework. This approach aims at the more *generalized partial credit model* (GPCM) proposed by Muraki (1992) on the lasso principle for parameter selection in detecting uniform DIF. This GPCM-lasso approach is applicable to polytomous item response models and allows several covariates of potentially mixed scale levels. Unlike the second method, GPCM-lasso extends DDF analysis by ordering the options from severely incorrect to completely correct. The gap between each option and its adjacent ordered one is defined as a step. The Differential step function (DSF) in GPCM-lasso method then deals with those steps by detecting differences in the probability that individuals from different groups will progress from one option to the adjacent more accurate option.

Applications of DDF Analysis

In this section, we provide two examples of DDF studies that illustrate how DDF can be used to approach test fairness and bias. In the first study, Middelton and Laitusis (2007) used DDF to determine which distractors on an assessment needed to be revised. In the second study,

Jamalzadeh et al. (2021) utilized DDF to evaluate the validity claims of a General English Achievement Test.

The Middelton and Laitusis (2007) study aimed to understand whether distractor choices functioned differently for students without learning disabilities as compared to those with learning disabilities who received accommodations. Three subgroups of accommodations were also considered: no accommodation, read-aloud accommodation, and other forms of accommodation. Approximately 45,000 students were included in the study. The state assessment consisted of 75 items, 42 measuring reading ability and 33 measuring writing ability. Utilizing standardized distractor analysis (SDA), Middelton and Laitusis concluded that seven items measuring reading ability were flagged for DDF when comparing students without learning disabilities and those with disabilities. In comparison, only three items measuring writing performance were flagged for DDF.

Jamalzadeh et al. (2021) aimed to investigate how the combination of DIF and DDF could be used to evaluate the validity of a General English Achievement Test (GEAT) administered at a large university. Specifically, their study sought to understand whether integrating DIF and DDF analysis can mitigate test bias and improve fairness. Participants were between 19 and 27 years of age, and English was not their first language. 63% of the participants were female. The assessment consisted of 60 multiple-choice items and measured four sub-sections: vocabulary, grammar, cloze test, and reading comprehension. The results of this assessment were used to determine if a student needed further instruction.

By combining DIF and DDF, Jamalzadeh et al. identified two items flagged for DIF in favor of female students and three items in favor of male students. Additionally, one of the items in favor of male students also exhibited DDF. Jamalzadeh et al. concluded that by combining

DIF and DDF, test fairness and bias could be improved by determining which distractors need to be revised to address DIF.

Focuses of Current Study

The current study applies DDF analysis with student responses to an introductory CS assessment, called the second CS1 assessment (SCS1). The assessment was validated by Parker and her colleagues (2016), and further scrutinized with internationality DIF (Parker et al., 2024). As the DIF analysis in the recent validation research is limited to the item level, examining only the item stem and the correct option while neglecting the distractors, the current study extends DDF to analyze student responses at the option level, filling a critical gap in creating equitable assessments. Specifically, my research question is: Do distractors on SCS1 assessments exhibit differential distractor functioning (DDF) for students of different genders?

This DDF analysis aims to ensure that all options function equitably for establishing an assessment's overall validity and fairness. The results not only highlight potential distractors that might be biased against specific groups but also offer more actionable and efficient insights for revising items. Furthermore, uncovering option selection patterns across groups after DDF evaluation can be pedagogically helpful in designing assessments and learning materials to address disparities in learning opportunities and access.

Methods

Participants

The response comes from multiple independent administrations of SCS1 in different contexts. At the end of the SCS1 assessment, demographic questions were listed but were not required. Due to this, the sample size with demographic information is only 259 undergraduate

students: 159 males and 100 females, with 61.4% and 38.6% in total, respectively; 102 whites and 157 non-whites, with 39.3% and 60.7% in total, respectively.

Instrument

The SCS1 assessment is a pseudocode-based test, making it suitable for use across a range of CS introductory courses, regardless of the programming languages being taught (Parker et al., 2016). It includes 9 concept topics (variables and assignments, logical operators, conditionals. for loops, while loops, arrays, function parameters, function return values, and recursion), each with 3 question types (definitional, code tracing, and code completion) for a total of 27 multiple-choice questions. Each item has five options: 1 correct option and 4 distractors.

Analytic Procedures

DDF methods were used to detect the invariance of all responses rather than just the invariance between correct and incorrect responses.

Multinominal Logistic Regression (MLR)

This method, proposed by Kato (2009), analyzes every response option in an item by producing a response characteristics curve (RCC) using a multi-step multinomial logistic regression approach. The likelihood of selecting the related option as a function of the ability proxy is represented by an RCC. Assume that there are K response options in an item. The RCC for response option k is the multinomial logistic function.

$$P_k(z) = \frac{exp(a_k + b_k Z)}{\sum_{l=1}^{k} exp(a_l + b_l Z)}$$

Where Z denotes the ability proxy (the standardized scale total score), and a_k and b_k are regression coefficients that represent the intercept and slope of the RCC of a given item. The

calculation process for MLR is based on a comparison of two models. The first model restricts item RCCs to be the same across the groups, indicating that there are no group differences, while the second model frees the restriction, which allows item RCCs to vary across the groups. The pseudo$-R^2$, which estimates the variances explained by the ability proxy from two models, is compared to detect items exhibiting DDF. A significant difference in R^2 indicates that the corresponding options show significant DDF.

Nominal Response Model (NRM)

In contrast to the classical test theory method in distractor analysis based on the total score, IRT distractor analysis not only assesses whether a distractor is functioning well but also allows the analyst to use distractors for estimating students' latent abilities. There are two methods based on IRT models for distractor analysis: the nominal-response model (Bock, 1972) and the graded-response model (Samejima, 1972).

The nominal-response model was developed by Bock (1972) to examine distractors in multiple-choice items. The nominal-response model estimates the likelihood of selecting each multiple-choice option without assuming any ordering among the options, in contrast to traditional IRT models that estimate focus only on the probability of selecting the correct option. The nominal-response model can be denoted as

$$P_j(k|\theta) = \frac{exp(a_{kj}\,\theta + d_{k.j})}{\sum_{i=1}^{n} exp(a_{i.j}\,\theta + d_{i.j})}$$

where $P_j(k|\theta)$ is the probability of selecting option k in item j condition on the student's ability θ, $a_{kj}$is the item discrimination for distractor k, $d_{k.j}$ is the difficulty of the distractor k, and m_j is the total number of options in item j. Furthermore, the set of $\{a_{kj}, d_{k.j}\}$ parameters must include

anchoring conditions to be able to uniquely determine the parameter values (e.g., $a_{0j} = d_{0.j} = 0$, $\forall\, j$) (Bock, 1972); a single a_{kj} or $d_{k.j}$ values are only comparable within the specific item.

Differential item functioning in the nominal-response model from Bock (1972) is performed with distractors by the following steps. First, conduct DIF analysis in the nominal-response model. Second, flagged DIF items from DIF parameters are converted to IRT parameters, freeing the constraints of DIF parameters that fix the first and last options to be constant, which $a_{k0} = 0$ and $a_{k-1} = k\text{-}1$. Since there is no item flagged DIF, we assume no difference between groups on the correct answer. Third, parameters a and d for the focal group should be adjusted to match the reference group based on the item key. Fourth, compute the z-score difference between the two groups based on the following formula:

$$Z_{ak} = \frac{a_{Fk} * (\frac{a_{Mc}}{a_{Fc}}) - a_{Mk}}{\sqrt{(SE_{aFk} * (\frac{a_{Mc}}{a_{Fc}}))^2 + (SE_{aMk})^2}},$$

$$Z_{dk} = \frac{d_{Fk} - (d_{Fc} - d_{Mc}) - d_{Mk}}{\sqrt{(SE_{dFk})^2 + (SE_{dMk})^2}}$$

Z_{ak} is identified a Z score difference among the two groups for option discrimination (parameter a). Where a_{Fk} is parameter a of the focal group (females) for option k and $a_{Mk}$is parameter a of the reference group (males) for option k. a_{Mc} and a_{Fc} represent the correct option as a parameter for male and female groups, respectively, for each item. SE_{aFk} and SE_{aMk} are the standard error parameters for the two groups. Z_{dk} is identified as the Z score difference between the two groups for option difficulty (parameter d). d_{Fk} and d_{Mk} are difficulty parameters for two groups, respectively, on option k. d_{Fc} and d_{Mc} are difficulty parameters for two groups on the correct option for the item. SE_{dFk} and SE_{dMk} are the standard errors of parameter d for the two groups.

Uniform DIF option is flagged when Z_{dk} exceeds criteria $\alpha = 0.05$, indicating that the option exhibits different difficulties across two groups after conditioning on their ability. Non-

uniform DIF option is flagged when Z_{ak} exceeds criteria $\alpha = 0.05$, representing that the option exhibits different discriminations among the two groups given their ability.

Generalized Partial Credit Model Lasso (GPCM Lasso)

Since Bock's NRM does not assume ordered response categories, it can be used to empirically rank distractors where ordering is related to research interest. Ranking all distractors of multiple-choice items can have many benefits for representing students' comprehensiveness of the concept being tested, specifically students who chose different distractors. A well-established ranking of distractors can inform interpretations of test results by indicating students' abilities based on their chosen distractors' varying levels of correctness (Smith & Bendjilali, 2022)

The Generalized Partial Credit Model with Lasso regularization (GPCMlasso) is a unified model with penalized likelihood estimation for parameter selection for Differential Item Functioning (DIF) detection proposed by Schauberger and Mair (2020). This model offers two benefits: (1) considering multiple variables simultaneously, and (2) applicable for both continuous and categorical variables for DIF detection. In GPCMlasso model, uniform DIF is determined by calculating the G variable from the equation below:

$$log(\frac{P(Y_{pi} = r)}{P(Y_{pi} = r - 1)}) = \beta_i = \left[\theta_p + \chi_p^T \alpha - \delta_{ir} - (\gamma_i \times G)\right]$$

Where γ_i represents DIF parameters, are the effects of group variable (G) on item *i,* respectively. If these parameters are non-zero after applying lasso penalization, they are considered uniform DIF. For parameter estimation, the lasso penalized log-likelihood function in the GPCMlasso model follows the equation below:

$$l_p(\theta, \alpha, \delta, \beta, \gamma) = l(\theta, \alpha, \delta, \beta, \gamma) - \lambda \sum_{i=1}^{l} \sum_{j=1}^{m} \omega_{ij} |\gamma_{ij}|$$

Where $l(\theta, \alpha, \delta, \beta, \gamma)$ is the regular version of the log-likelihood function, $\lambda \sum_{i=1}^{l} \sum_{j=1}^{m} \omega_{ij} |\gamma_{ij}|$ is the lasso penalty term, and $\lambda \geq 0$ is the tuning parameter that controls the degree of penalization applied to the vector of regression coefficients γ_{ij}. l_p represents an m-dimensional vector of covariates for individual p, and θ stands for the latent trait of the focus. In addition to the DIF parameter, γ_{ij}, the item step parameter δ_{ij}, the main effect α, and the item discrimination β_i parameters are estimated on item i.

In addition to DIF detection, the more general GPCMlasso model is capable of detecting Differential Step Functioning (DSF) (Schauberger & Mair, 2019b). Here, DSF indicates different covariate values are associated with different levels of a polytomous item as steps. The penalty term can be denoted as follows:

$$J(\theta, \alpha, \delta, \beta, \gamma) = \sum_{i=1}^{l} \sum_{j=1}^{m} \sum_{r=1}^{k_i} \omega_{ij(r)} |\gamma_{ij(r)}| + \sum_{i=1}^{l} \sum_{j=1}^{m} \sum_{r<s}^{k_i} \omega_{ij(r,s)} |\gamma_{ij(r)} - \gamma_{ij(s)}|$$

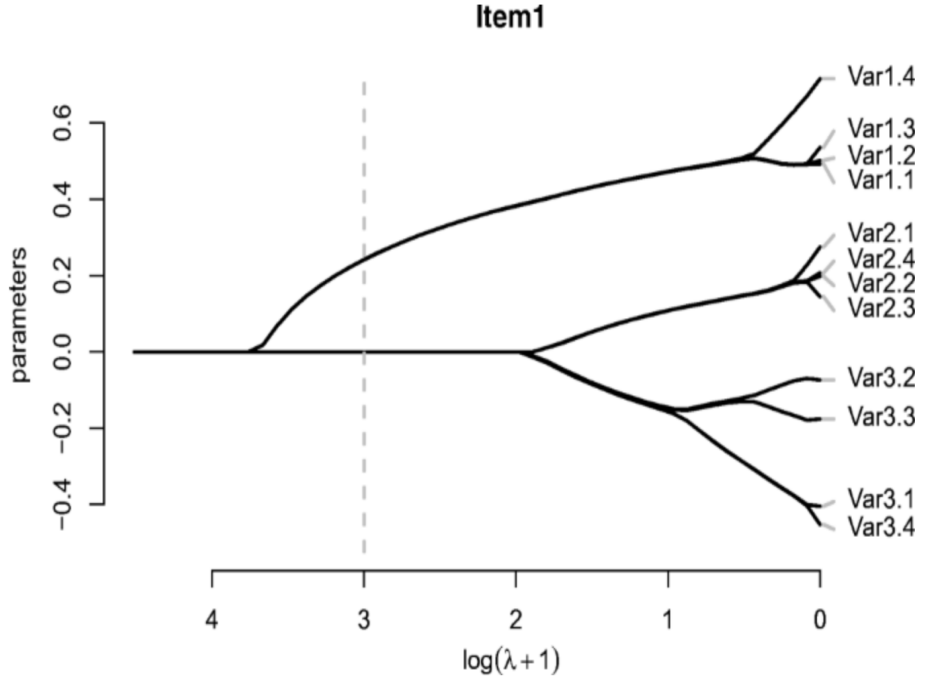
The two parts of the formula serve different purposes. The first part represents the penalty term, penalizes all DIF parameters' absolute values, and possibly sets them all to 0. The next part is to fuse the groups of all k-1 parameters corresponding to the same pair between item i and covariate j. In total, there are $l \times m$ groups. Penalizing all differences between the respective parameters is a way of fusing. Once λ is large enough, this penalty produces parameters that are exactly equal. In this case, it indicates the existence of DIF and the influences of the corresponding covariate on the item but no DSF is observed.

Figure 1 below illustrates how the penalty works. The parameter pathways for each DIF parameter that corresponds to this item are shown in the plot. In this example, three potential DIF variables are used, and the response is measured on a five-point scale. In this specific item $m \times (k - 1)$ parameters are used to parameterize all possible DIF and DSF effects. While λ decreases, the model becomes more complex as the penalty term's constraints loosen. In

contrast, with larger λ all parameters according to the same variable are equal. The gray dashed line refers to log(λ +1) = 3. In this model, it would have detected DIF exhibited in variable 1. For the reason that all parameters corresponding to variable 1 are not zero, there is no difference among them, and there is no DSF, but DIF exists.

Figure 1

Parameter Path of DIF/DSF Parameters for an Exemplary Item Illustrating in DSF Analysis



Results

Descriptive Statistics of All Items

Item statistics for each item on SCS1 based on the classical test theory item analysis are presented in Table 1, including item difficulty, item discrimination, and α if deleted. The lower difficulty score indicates that the lower the passing rate, the more difficult the item is. Low discrimination scores represent poor ability to differentiate the overall performance of the assessment. The reliability of SCS1 has Cronbach's alpha 0.79, within the acceptable range.

However, the average total score is 8.94 out of 27, indicating that overall this test is relatively challenging for undergraduate students.

Table 1

Item Statistics

Item	Item Difficulty	Item Discrimination	α if Dropped	Item Difficulty for Female	Item Difficulty for Male
Q1	0.44	0.30	0.79	0.47	0.42
Q2	0.51	0.26	0.79	0.47	0.54
Q3	0.66	0.41	0.78	0.62	0.68
Q4	0.22	0.25	0.79	0.22	0.22
Q5	0.19	0.41	0.78	0.22	0.19
Q6	0.49	0.21	0.79	0.40	0.54
Q7	0.27	0.37	0.78	0.23	0.28
Q8	0.38	0.30	0.79	0.27	0.45
Q9	0.38	0.40	0.78	0.39	0.38
Q10	0.39	0.40	0.78	0.38	0.40
Q11	0.34	0.37	0.78	0.38	0.32
Q12	0.53	0.38	0.78	0.49	0.55
Q13	0.25	0.17	0.79	0.19	0.29
Q14	0.43	0.43	0.78	0.49	0.40
Q15	0.21	0.27	0.79	0.19	0.22
Q16	0.33	0.35	0.78	0.35	0.31
Q17	0.30	0.35	0.78	0.36	0.28
Q18	0.28	0.17	0.79	0.30	0.26
Q19	0.51	0.49	0.78	0.50	0.52
Q20	0.20	0.03	0.80	0.13	0.25
Q21	0.33	0.28	0.79	0.31	0.33
Q22	0.41	0.26	0.79	0.48	0.36
Q23	0.65	0.35	0.78	0.59	0.68
Q24	0.26	0.33	0.79	0.32	0.23
Q25	0.53	0.29	0.79	0.58	0.49
Q26	0.30	0.41	0.78	0.32	0.30
Q27	0.16	0.15	0.79	0.12	0.18

Results for Multinomial Log-linear Regression

The likelihood ratio statistics based on the MLR method were conducted on all items. IRCC's plot includes ICCs for both the correct option and distractors. The IRCCs present the distribution of distractor response across the entire range of observed total z scores to detect the DDF effects. Only Items 6, 8, 11, 18, and 25 were detected with DDF present in distractors. Table 2 includes distractors' difficulty, discrimination, difficulty DIF, and discrimination DIF for those items. The correct option of each item serves as the reference for comparison in distractor analysis, so it is excluded from the result table. Females were the focal group, while males were the reference group in the analysis. Non-zero difficulty DIF and discrimination DIF parameters indicate that DDF is present. A negative value of difficulty DIF means that this distractor is more difficult for males, based on the observed total z-score. In other words, it is more appealing to the males, as they have more difficulty recognizing that this is an incorrect option; in conclusion, it is biased against males. In addition, a negative value of discrimination DIF indicates that this distractor is better at differentiating male performance conditioned on the observed total z-score, which showed non-uniform DIF in this distractor.

All distractors in these five items exhibited non-uniform DIF, whose difficulty DIF and discrimination DIF parameters were all non-zero. Figures 2 show the IRCCs of these 5 items, disparities between the two groups' lines present in every option. However, only for Item 8, all distractors exhibit DDF that were biased against females consistently. The other 4 items, the distractors in the same question, attracted different genders inconsistently. Among these 4 items, two (Items 6 and 25) having 3 out of 4 distractors appeared more appealing to the female group, indicated that those options were biased against female individuals. In contrast, the other 2 items

(Items 11 and 18) with 3 out of 4 distractors were more appealing to the male group, showing that they were biased against the male group.

Table 2

MLR Result

Item Option	Difficulty	Discrimination	Difficulty DIF	Discrimination DIF
Item6-B	-1.67	-1.17	0.05	0.78
Item6-C	-6.43	-0.26	5.09	-2.11
Item6-D	-1.96	-0.82	1.18	-0.68
Item6-E	-1.54	-0.84	-0.18	0.38
Item8-A	-1.32	-1.70	0.72	-1.10
Item8-B	-1.64	-0.51	1.59	-0.26
Item8-C	-1.03	-0.58	1.17	-1.24
Item8-D	-1.53	-1.49	1.31	0.05
Item11-A	-0.75	-1.29	-0.51	-0.66
Item11-B	-0.10	-1.15	-0.15	0.01
Item11-C	-0.49	-1.04	-0.67	-0.92
Item11-E	-0.66	-0.84	0.30	-0.18
Item18-A	0.30	-0.73	0.18	1.05
Item18-C	-0.17	-0.67	-0.91	-0.49
Item18-D	-0.24	-1.61	-0.79	1.09
Item18-E	-0.93	-0.89	-1.22	0.34
Item25-A	-1.09	-1.30	-66.96	1.24
Item25-C	-3.87	-0.47	2.80	-2.10
Item25-D	-1.37	-0.55	0.26	-0.78
Item25-E	-1.73	-1.32	0.24	-0.44

Figure 2a

MLR IRCCs by Gender for Item 6

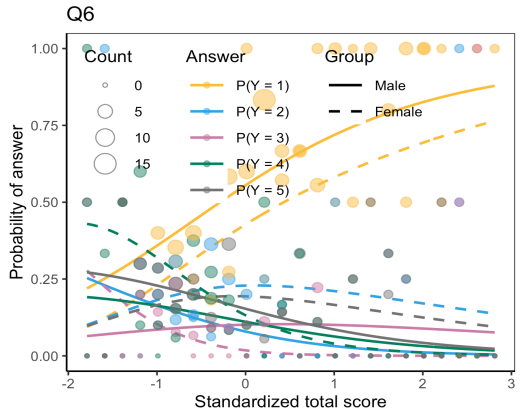


Figure 2b

MLR IRCCs by Gender for Item 8

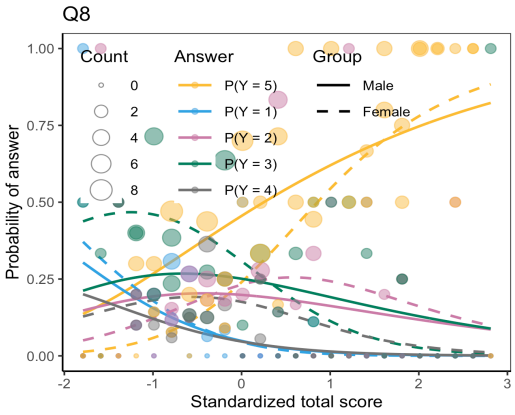


Figure 2c

MLR IRCCs by Gender for Item 11

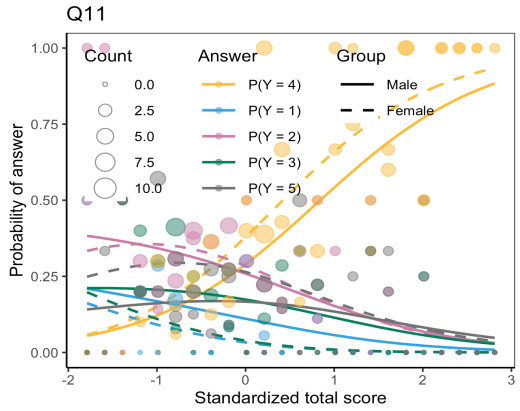


Figure 2d

MLR IRCCs by Gender for Item 18

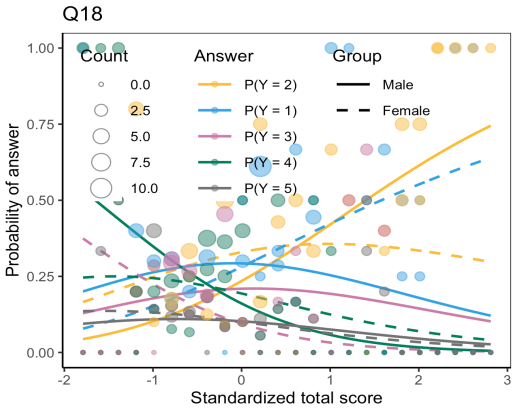
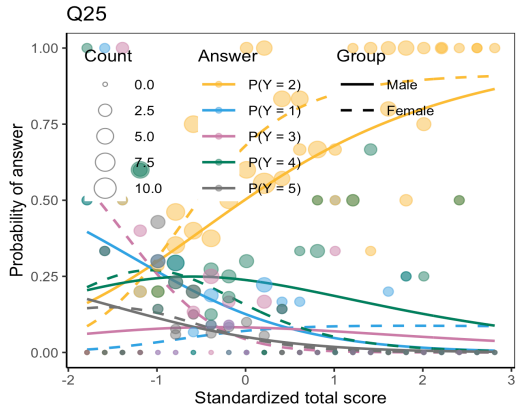


Figure 2e

MLR IRCCs by Gender for Item 25



Results for Nominal response model

No item was detected in the DIF in the 2PL model, so we set 50% of items to be anchor items based on the ratio test. Free baseline designated anchor DIF was applied to the nominal response model. A baseline model is freeing all items except a variance across groups, while subsequent models constrain item parameters to be equivalent across groups. Then the difference in G^2 for baseline and subsequent models was calculated. The advantage of free baseline designated anchor DIF is that it yields higher power to detect "small" DIF with samples.

The result showed that Items 4 and 8 exhibited DIF with criteria adjusted $p < 0.05$. Between-group z scores, differences were calculated by adjusted IRT parameters, which set the correct option to the same between groups. Adjusted distractors' discrimination parameters and standard error of difficulty parameters were calculated by multiplying the ratio of the correct option's discrimination parameter of the focal group over the reference group. The distractor's difficulty parameters were adjusted by subtracting the difference of the correct option's difficulty parameter between the reference group and focal group, while the distractors' standard error remained unchanged. Z score differences of discrimination and difficulty between groups were presented in Table 3 with $p < 0.05$ ($z = \pm 1.96$) as a criterion flagged for DIF distractors. Positive Z score difference of discrimination represents the distractor performed better at differentiating females' ability and vice versa. Positive z score difference in difficulty indicates that the female group felt more difficult in distinguishing the incorrect option compared to the male group, given the same level of ability. As shown in Table 3, item 4 distractors A and B were non-uniformed. From the numeric result, Item 4 option A, with difficulty z score difference -2.57, indicates it is biased against males, while for discrimination -2.52, showing that this option is better at discriminating male performance. As for Item 4, option B with a positive discrimination z score

of 3.12, represents that this option performs better at differentiating female ability. For item 8, options B, C and D exhibit uniform DIF, with positive difficulty z-score differences exceeding the criterion of 1.96. This indicates these options are all against females, as female students are more likely to select the incorrect answers. In other words, it is more challenging for females to distinguish the incorrect options even if they possess the same level of latent trait as their male counterparts. Figure 3a includes the result of NRM for Item 4 distractors' IRCCs while Figure 3b shows the result for Item 8 distractors.

Table 3

NRM Result

Z Score Difference of	Distractor			
	A	B	C	D
Item 4				
… discrimination	-2.52	3.12	0.01	0.09
… difficulty	-2.57	0.07	1.10	0.60
Item 8				
… discrimination	0.18	-0.05	-1.82	0.91
… difficulty	1.46	2.46	3.14	4.00

Figure 3a

NRM Result for Item 4

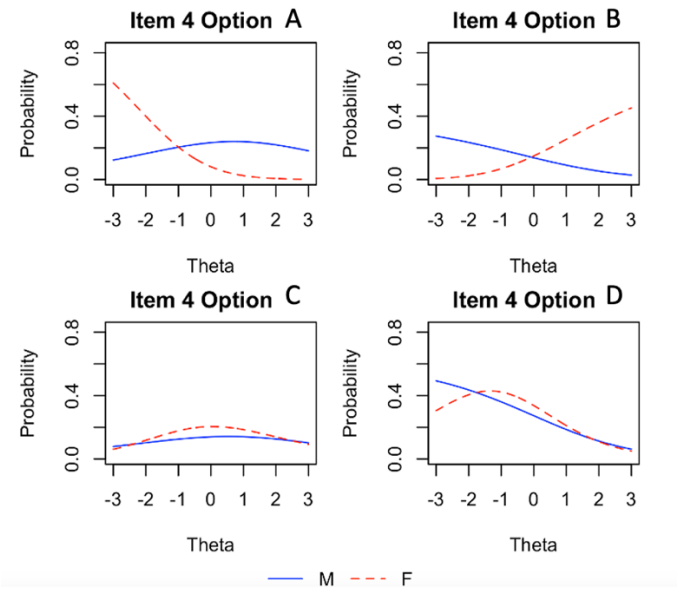
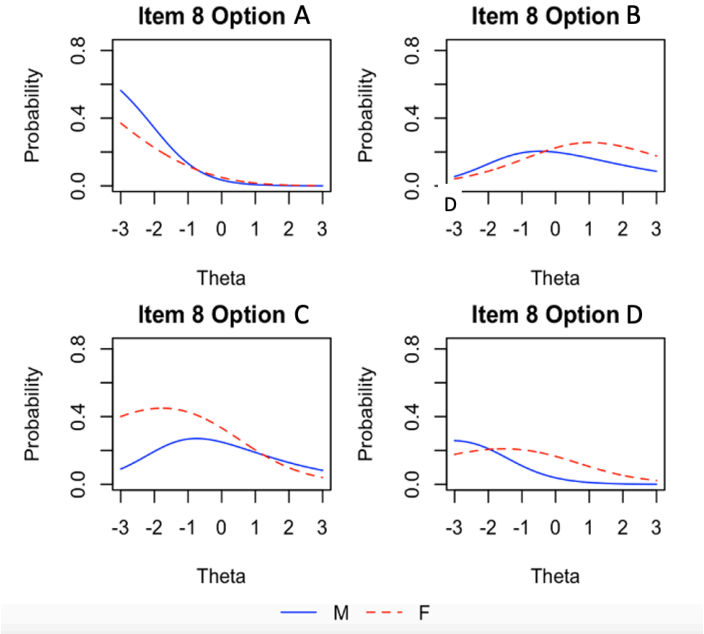


Figure 3b

NRM Result for Item 8



Results for GPCM-lasso

GPCM-lasso was performed with the SCS1 data for ten items which most likely exhibited DIF based on p-value. Again, the female group was the focal group while male was the reference group. In the DSF detection, the step is referred to as the correctness level of the distractors. The correctness level order was determined by an expert panel, which included a professor teaching undergraduate computer science, two doctoral students with experience in computer science education, and a master's student enrolled in a CS1 course previously. Distractors were ordered into 1 to 4 from the least to the most correct, categorized as mostly incorrect, slightly correct, moderately correct, and highly correct.

This section focuses on illustrating the differences I found between the DIF and DSF detections of the SCS1 data. Starting with DIF detection, Figure 4a shows the coefficient paths for the gender variable and all items along the tuning parameter λ. The paths are plotted separately for each item. The dashed lines represent the optimal model according to BIC.

The positive parameter from the result indicates that it favors the focal group (female) and is biased against the reference group (male). While the negative parameter represents that it favors the reference group (male) and is biased against the focal group (female). The result indicates that only Item 8 showed positive gender DIF parameters (0.477), which was biased against the male group at the item level. The other nine items remaining zero indicate no bias between gender groups. However, in DSF detection, the result indicates whether the step between adjacent distractors ordered in levels of correctness exhibits DSF in gender. Only Item 11 step 2 gap between Distractors C and B (slightly correct to moderately correct) showed negative gender DIF parameters (-0.011) that are biased against the focal group (female). The rest of the distractors in these ten items were not flagged for gender DSF.

Item 8, steps 1,3, and 4 (mostly incorrect to slightly correct, moderately correct to mostly correct, and mostly correct to correct options) have different functions favoring females. On the other hand, although item 8 shows DIF in gender at the item level, it does not show DSF, which means that no DIF is detected in distractors. In addition, item 11 step 3 (step between slightly correct to moderately correct) functions differently biased against females and step 4 (mostly incorrect to correct option) favoring females. Only item 11, step 2 (slightly correct to moderately correct), is flagged for gender DSF as biased against females. However, item 11 is not flagged for DIF at the item level. Figure 4a reports the result of DIF at the item level, and Figure 4b shows the result of DSF (i.e., DIF at the distractor level).

Figure 4a

GPCM Lasso DIF at the Item Level

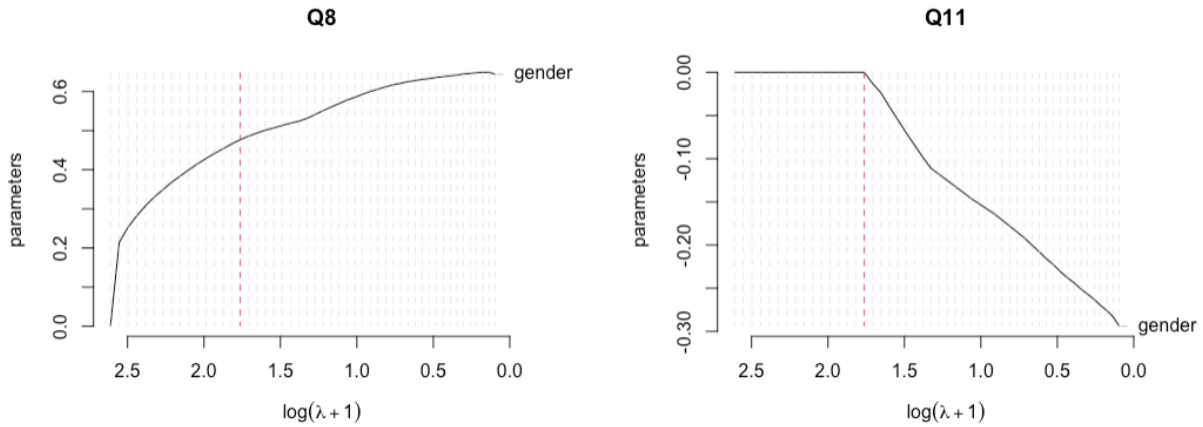
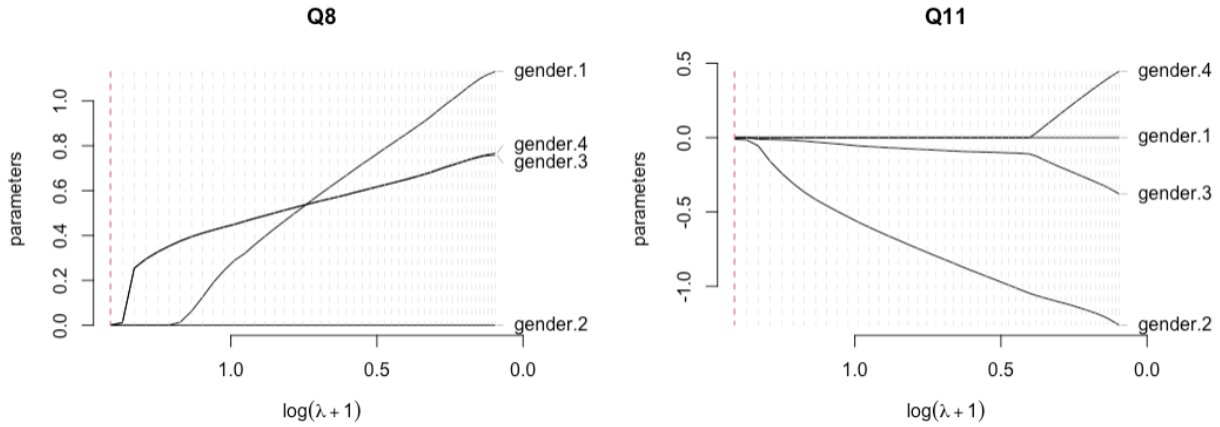


Figure 4b

GPCM Lasso DSF at the Distractor Level



Discussion and Conclusions

This study aims to detect item distractors that exhibit DIF across gender groups when there is no item DIF effect. This ensures that both items and distractors function appropriately, guaranteeing assessment fairness. For this purpose, first, the MLR is categorized as a divided-by-distractor method based on the observed total score, which investigates the differential distractor functioning (DDF) effect independently from differential item functioning (DIF). This method dives deeper into the option level to investigate bias against gender groups. However, the MLR method provides a perspective more toward the classical test theory. Second, DIF detection in the IRT NRM method also focuses on the distractor level DIF detection, but different from the MLR method, this is based on the latent trait, which is the strength of IRT. Third, the GPCM-lasso method, conducting analysis by ordering the distractors by the correctness level, further offers a different perspective that explains the patterns of different levels of misconception across gender groups. These three methods flag different DIF distractors, providing different perspectives on assessment items.

DDF results of the SCS1 show that 5 out of 27 items were flagged as exhibiting DDF by the MLR method. In these 5 items, all the distractors were detected non-uniform gender DIF, each distractor functioned differently across gender groups with different difficulty and performed unequally on differentiating the examinees' ability. The direction of DDF result, only distractors in item 8 showed consistently favoring the same group (males) and biased against female. In the other four items, the distractors showed inconsistent patterns: 3 out of 4 distractors were more appealing to one gender group, while the other distractors favored the other gender.

DDF results from the IRT NRM method tend to pinpoint biased options instead of all distractors in an item, which provides more specific detailed information on which distractors

should be further taken into consideration for revision. Distractors A and B in item 4 were flagged for non-uniform gender DIF options due to different shapes of ICC curves from gender groups, indicating they performed differently across genders in differentiating the examinee's ability. While Distractors B, C, and D in item 8 were detected, uniform DIF, which were easier for males according to lower difficulty parameters, means that males were easier to recognize the distractors were wrong and were less likely to choose these distractors. In other words, these three options in item 8 exhibited bias against females.

DDF results from the GPCM-lasso can detect DIF at both item and distractor levels by treating distractors as different levels of correctness toward the CS concept. The result indicates that item 8 at the item level showed favoring females while no bias was detected at the distractor level. On the other hand, there was no item DIF detected in Item 11 at the item level, but step 2, moving from slightly correct to moderately correct option, was detected in bias against females.

Overall, the different DDF results across these three methods indicate that MLR, IRT-NRM and GPCM-lasso all provide useful and complementary perspectives when evaluating assessment items. In summary, Item 8 has consistently been flagged for DIF at either the item or distractor level for all three methods. Distractors B, C and D were easier for males than females, while at the item level, it was detected favoring females by the GPCM lasso method with the positive parameter.

Item 8 is about the for-loop concept. The item includes an array of characters, print statements and the more advanced concept nested for loops is measured. This item requires the examinee to trace the code and select the correct output. However, the item centered around the word SCIENCE and how to print it, character by character, across new lines. Previous studies (e.g., Parker et al, 2024) speculated that it might be the word "science" that caused the different

performance patterns between genders. However, not all the distractors containing the word "science" were detected in distractor DIF detection. Furthermore, qualitative research such as think-aloud interviews would be needed to understand how different genders perceive the item or replace the word "science" with a more natural word in this item.

Conducting distractor analysis on SCS1 helps give greater details or directions on how to revise a biased item in introductory computing assessment. Even if none of the items are detected with DIF, the results reveal true differences in students' selections of options. These can be valuable assessment results to reveal how students make sense of and resonate with varying partially correct ideas. This assessment information provides instructors with insightful hints in diagnosing learners' performances, especially their struggles and misconceptions.

The findings of this study should be interpreted with caution due to the following three limitations. First, the small sample size of the dataset (n=259) and the slightly unbalanced number of two genders (females: 100 and males: 159) would be a concern, with an even smaller sample size of each distractor while conducting the distractor analysis. The results might be different with a larger and more balanced sample size. Second, the DIF procedure as a quantitative method only flags potentially problematic items or distractors. They are incapable of answering the why question. Protocol interviews and expert panel reviews are essential in uncovering the hidden reasons for the assessment bias. The third limitation of this study lies in the order of the distractors in the GPCM lasso method. There might not be a consensus on the level of correctness of distractors among different learning theories in computing education, therefore the order adopted by this study is not fully justified by computing education literature.

Detecting DIF on distractors is one initial step toward uncovering and disrupting the underlying bias in assessing student learning. In computing education, it is worth investigating

the equity gaps between genders to address the gender imbalance in this field. Assessment items that are free of bias at item and option levels will offer more formative values in support of computing education. With the prosperity of computing education and advancements in psychometric research, educators are now equipped to pedagogically identify students' strengths and pinpoint their conceptual weaknesses through computing assessments, thus ultimately addressing students' learning needs and close the achievement gaps.

Acknowledgment: This work was supported by the National Science Foundation under IUSE-2417207 and 2417208.

References

Astin, A. W., & Astin, H. S. (1992). *Undergraduate science education: The impact of different college environments on the educational pipeline in the sciences* (Final Report). California University.

Abedi, J., Leon, S., & Kao, J. C. (2008). Examining differential item functioning in reading assessments for students with disabilities. *CRESST Report*, 744.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26*(4), 381–409.

Cheek, D. W., & Agruso, S. (1995). Gender and equity issues in computer-based science assessment. *Journal of Science Education and Technology, 4*(1), 75–79.

Cundiff, J. L., Vescio, T. K., Loken, E., & Lo, L. (2013). Do gender–science stereotypes predict science identification and science career aspirations among undergraduate science majors? *Social Psychology of Education, 16*(4), 541–554.

Davidson, M. J., Wortzman, B., Ko, A. J., & Li, M. (2021). Investigating item bias in a CS1 exam with differential item functioning. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (pp. 1142–1148).

De Ayala, R. J. (2009). *The theory and practice of item response theory.* Guilford Press.

Eccles, J. S. (1994). Understanding women's educational and occupational choices: Applying the Eccles et al. model of achievement-related choices. *Psychology of Women Quarterly, 18*(4), 585–609.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*(5220), 41–45.

Hill, C., Corbett, C., & St. Rose, A. (2010). *Why so few? Women in science, technology, engineering, and mathematics.* American Association of University Women.

Jamalzadeh, M., Lotfi, A. R., & Rostami, M. (2021). Assessing the validity of an IAU General English Achievement Test through hybridizing differential item functioning and differential distractor functioning. *Language Testing in Asia, 11*(1), 1–17.

Jagacinski, C. M. (2013). Women engineering students: Competence perceptions and achievement goals in the freshman engineering course. *Sex Roles, 69*(11), 644–657.

Kato, K., Moen, R. E., & Thurlow, M. L. (2009). Differentials of a state reading assessment: Item functioning, distractor functioning, and omission frequency for disability categories. *Educational Measurement: Issues and Practice, 28*(1), 28–40.

Koon, S., & Kamata, A. (2013). A comparison of methods for detecting differential distractor functioning. *International Journal of Quantitative Research in Education, 1*(4), 364–382.

Lewis, C. M., Anderson, R. E., & Yasuhara, K. (2016, August). "I don't code all day": Fitting in computer science when the stereotypes don't fit. In *Proceedings of the 2016 ACM Conference on International Computing Education Research* (pp. 23–32).

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). A review of recent developments in differential item functioning. *ETS Research Report Series, 2008*(2), i-32.

Medel, P., & Pournaghshband, V. (2017, March). Eliminating gender bias in computer science education materials. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education* (pp. 411–416).

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied psychological measurement, 16*(2), 159-176.

Na, C., Clarke-Midura, J., Shumway, J., van Dijk, W., & Lee, V. R. (2024). Validating a performance assessment of computational thinking for early childhood using item response theory. *International Journal of Child-Computer Interaction, 40,* 100650.

National Science Foundation. (2017). *Women, minorities, and persons with disabilities in science and engineering: 2017.* Arlington, VA: National Science Foundation.

Nolan, K., Mooney, A., & Bergin, S. (2019, January). An investigation of gender differences in computer science using physiological, psychological, and behavioral metrics. In *Proceedings of the Twenty-First Australasian Computing Education Conference* (pp. 47–55).

Parker, M. C., Guzdial, M., & Engleman, S. (2016, August). Replication, validation, and use of a language independent CS1 knowledge assessment. In *Proceedings of the 2016 ACM Conference on International Computing Education Research* (pp. 93–101).

Parker, M. C., Ren, H., Li, M., & Wang, C. (2024, March). Intersectional biases within an introductory computing assessment. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (pp. 1021–1027).

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement, 44*(3), 187-210.

Penfield, R. D., Gattamorta, K., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice, 28*(1), 38-49.

Quille, K., Culligan, N., & Bergin, S. (2017, June). Insights on gender differences in CS1: A multi-institutional, multivariate study. In *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education* (pp. 263–268).

Redmond, K., Evans, S., & Sahami, M. (2013, March). A large-scale quantitative study of women in computer science at Stanford University. In *Proceedings of the 44th ACM Technical Symposium on Computer Science Education* (pp. 439–444).

Rommes, E., Overbeek, G., Scholte, R., Engels, R., & De Kemp, R. (2007). "I'm not interested in computers": Gender-based occupational choices of adolescents. *Information, Communication & Society, 10*(3), 299–319.

Schauberger, G., & Mair, P. (2019b). A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behavior Research Methods, 52*(1), 279–294.

Schmidt, F. L. (2011). A theory of sex differences in technical aptitude and some supporting evidence. *Perspectives on Psychological Science, 6*(6), 560–573.

Sibia, N., Bui, G., Wang, B., Tan, Y., Zavaleta Bernuy, A., Bauer, C., ... & Petersen, A. (2024, March). Examining intention to major in Computer Science: Perceived potential and challenges. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (pp. 1237-1243).

Smith, T. I., & Bendjilali, N. (2022). Motivations for using the item response theory nominal response model to rank responses to multiple-choice items. *Physical Review Physics Education Research, 18*(1), 010133.

Tew, A. E., & Guzdial, M. (2010, March). Developing a validated assessment of fundamental CS1 concepts. In *Proceedings of the 41st ACM Technical Symposium on Computer Science Education* (pp. 97–101).

Wang, M.-T., & Degol, J. (2013). Motivational pathways to STEM career choices: Using expectancy–value perspective to understand individual and gender differences in STEM fields. *Developmental Review, 33*(4), 304–340.

Xie, B., Davidson, M. J., Li, M., & Ko, A. J. (2019, February). An item response theory evaluation of a language-independent CS1 knowledge assessment. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (pp. 699–705).

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF).* Ottawa: National Defense Headquarters.