Investigating Gender Differential Item Functioning (DIF) in the eTIMSS 2019 Math Assessment

Yiqi Zheng

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Education

University of Washington

2025

Committee:

Min Li

Chun Wang

Program Authorized to Offer Degree:

Education

University of Washington

Abstract

Investigating Gender Differential Item Functioning (DIF) in

the eTIMSS 2019 Math Assessment

Yiqi Zheng

Chair of the Supervisory Committee:

Min Li

College of Education

The Trends in International Mathematics and Science Study (TIMSS) has long been a crucial tool for evaluating students' mathematical achievement on a global scale. The reliability, validity, and fairness of its score interpretations have been extensively researched. With the introduction of the digital version (eTIMSS) and the addition of innovative Problem-Solving and Inquiry (PSI) tasks in 2019, it becomes essential to explore how these newly adopted changes affect item and assessment quality. Differential Item Functioning (DIF) analysis is an advanced method in educational assessment used to detect potentially biased items across different demographic groups, such as gender groups (boys and girls). However, previous research on math assessments has yielded inconclusive findings on which item features may be the sources of gender-related DIF. By utilizing eTIMSS 2019 United States student response and

demographic data, this study has two main objectives: (1) identifying DIF items and estimating their magnitude using both Non-Item Response Theory (Non-IRT)-based and Item Response Theory (IRT)-based DIF detection methods, and (2) systematically coding item features through consensus coding to explore their associations with DIF patterns using stepwise linear regression. The findings indicated that the gender DIF patterns were related to the number of clauses, the presence of context, the involvement of construction, and the number of metric system units. Additionally, newly added PSI booklets exhibited a higher percentage of DIF items compared to regular digital items. The inconsistency in linking item features to DIF estimates between regular digital math items and PSI items stressed the need for further research on the effects of item features on how students respond to different task types, especially in the era of digital technology.

Keywords: Differential Item Functioning (DIF), Gender DIF, Item Features, Large-Scale Mathematics Assessment, Assessment Fairness, TIMSS

Investigating Differential Item Functioning (DIF) in the eTIMSS 2019 Math Assessment

Detecting Differential Item Functioning (DIF) has been an important issue in educational assessment. Differential Item Functioning (DIF) refers to the phenomenon where items on a test function differently between two groups of examinees who have the same level of ability; in other words, an item is flagged as DIF if the probabilities of examinees from two groups answering this item correctly are significantly different, even though they have equal ability (Holland & Wainer, 1993). There are two types of DIF: uniform DIF and non-uniform DIF. Uniform DIF means differences in item performance of two equal-ability groups are consistent across all ability levels, whereas non-uniform DIF means the differences in item performance between groups vary across different ability levels (Mellenberg, 1982).

DIF analysis is crucial for detecting item bias because simply comparing groups on total scores can lead to incorrect conclusions about fairness. For example, Martinková and colleagues (2017) illustrate that item bias does not exist despite a significant difference between groups on total scores, while item bias can exist even if two groups have the same total score. DIF analysis examines the difference at the item level, specifically, locates the significant performance differences between two groups after controlling the ability level and finds patterns among differences (Zieky, 2003). Items flagged by DIF analysis are suggested for further expert review and other additional analyses to decide whether the significant differences between groups are fair, thereby promoting the validity and fairness of the assessment (Martinková et al., 2017).

DIF Detection Methods

DIF detection methods proposed by researchers can be generally divided into two categories: Non-Item Response Theory (Non-IRT)-based methods and Item Response Theory

(IRT)-based methods. Non-IRT-based methods, including the ManteI-Haenszel Method and Logistic Regression, mainly rely on comparing item performance between groups, which often use observed scores. Meanwhile, IRT-based methods compare item parameter estimates or model fit between groups based on estimates of latent ability, including the Likelihood Ratio Test and Raju's Area Measure. Each method has its properties in detecting DIF (see Table 1).

ManteI-Haenszel Method

The Mantel-Haenszel (MH) method is one of the most common methods used to detect DIF. Initially, the Mantel-Haenszel (MH) procedure was proposed by Nathan Mantel and William Haenszel (1959) to analyze contingency tables for retrospective studies of disease. Later, the MH procedure was introduced and proved powerful in detecting items that function differently in two groups of examinees (Holland & Thayer, 1988).

The Mantel-Haenszel (MH) method involves two parts: conducting the chi-square test for the hypothesis and measuring the DIF effect size by log-odds ratio. To perform the MH DIF analysis, the focal group, and the reference group are matched by total scores. The focal group is the focus of study, and the reference group is the basis for the focal group for the comparison (Dorans & Holland, 1993). For each level of matched total score, the odds ratio of correct response for the reference group and the focal group is calculated for each item. Next, α_{MH}, as the common odd ratio, is formed based on the weighted average of the odds ratio for two groups on a certain item of interest at each level of matching proficiency. α_{MH} yields a measure of each item's effect size for evaluating the DIF magnitude by the formula MH D-DIF= -2.35 In (α_{MH}). If MH D-DIF is equal to 0, then no DIF is detected. Negative MH D-DIF indicates the item favors the reference group and disadvantages the focal group; Positive MH D-DIF indicates the item favors the focal group and is biased against the reference group.

Table 1

Comparison of Differential Item Functioning (DIF) Detection Methods

Method	Description	Type of DIF Detected		Advantages	Limitations
		Uniform	Non-uniform		
Mantel-Haenszel (MH) Method	Compare the odds of correct response for two groups by using the chi-square test while controlling for total score	x		● Used widely in large-scale assessment ● Produce results easy to interpret	● Limited to detect uniform DIF
Logistic Regression (LR)	Use the Likelihood Ratio Test to determine, when modeling the probability of a correct response as a function of total score as latent ability, whether adding group membership or an interaction term significantly improves model fit	x	x	● Exhibit strong power in DIF detection	● Logistic regression model usually deals with dichotomous data
Likelihood Ratio Test (LRT)	Fit two IRT models, DIF model allowing DIF and baseline model not, and use the Chi-square test to assess model fit	x	x	● Fit different IRT models for multiple types of items ● Provide robust statistical testing by selecting anchor items	● Computationally intensive to fit complex IRT models ● Results do not directly indicate the direction or magnitude of DIF, and need further interpretation
Raju's Area	Measure the area between the item characteristic curves (ICCs) for focal and reference groups, reflecting differences in expected scores across groups	x	x	● Capture and quantify subtle differences in item functioning	● Sensitive to sample size and model assumptions ● Challenging to interpret the quantified area difference, especially for non-uniform DIF

Then, the MH method provides chi-square tests for each item with 1 degree of freedom for the null hypothesis that the common odds ratio α_{MH} is equal to 1 (MH D-DIF = 0), while the alternative hypothesis is that the common odds ratio α_{MH} is different from 1 (MH D-DIF ≠ 0), to determine whether the common odds ratio α_{MH} significantly deviates from 1.

$$\alpha_{MH} = \frac{\sum_{k=0}^{K} \frac{A_k D_k}{N_k}}{\sum_{k=0}^{K} \frac{B_k C_k}{N_k}}$$

$$\text{MH CHISQ} = \frac{\left(\left|\sum_k A_k - \sum_k E(A_k)\right| - \frac{1}{2}\right)^2}{\sum_{k=0}^{K} \text{Var}(A_k)}$$

Note. $E(A_k) = \frac{(A_k+B_k)(A_k+C_k)}{N_k}$ and $\text{Var}(A_k) = \frac{(A_k+B_k)(C_k+D_k)(A_k+C_k)(B_k+D_k)}{N_k^2(N_k-1)}$

The ETS system classifies MH D-DIF into three categories: A (negligible or non-significant DIF), B (slight to moderate DIF), or C (moderate to large DIF) (see Table 2; Dorans & Holland, 1993; Zieky, 2003).

Table 2

Mantel-Haenszel (MH) D-DIF Categories and Criteria

Category	Criteria
A - Negligible or non-significant DIF	|MH D-DIF| < 1; or MH CHISQ not significant ($p \geq .05$)
B - Slight to moderate DIF	If items do not meet the criterion for C and if MH CHISQ > 3.84 and |MH D-DIF| ≥ 1
C - Moderate to large DIF	|MH D-DIF| is significantly larger than 1 indicated by MH CHISQ ($p < .05$), and |MH D-DIF| ≥ 1.5

Note. D-DIF = differential item functioning effect size based on the Mantel-Haenszel statistic.

MH method is uniformly the most powerful test for the null hypothesis against the alternative of uniform DIF. Each item is computed for its log odds ratio and chi-square statistic individually, as a result, the decision is made as to DIF for each item. However, the MH method is a test of uniform DIF for dichotomous items, though it is somewhat sensitive to the non-uniform DIF.

Logistic Regression

The Logistic Regression method was introduced by Swaminathan and Rogers (1990). The Logistic Regression method is not an IRT-based method as it does not estimate the item parameters or latent traits; instead, it takes the total score as the proxy of ability. Unlike the MH method, the Logistic Regression method is capable of detecting both uniform DIF and non-uniform DIF.

The Logistic Regression method approaches the 2PL IRT model with a logistic regression model. That is, the probability of a correct response is modeled as a linear function of predictor variables. The baseline model (i.e., Model 1) only has a total score as a predictor, while the two augmented models include the total test score, group membership, and the interaction between group membership and the total test score. By examining whether the log-likelihood values of the model increase significantly after adding additional predictor terms, an item can be determined to exhibit DIF or not. Uniform DIF could be tested by comparing the log-likelihood values for Model 1 and Model 2 (df=1); Non-uniform DIF could be tested by comparing the log-likelihood values for Model 2 and Model 3 (df=1). An overall test of the total DIF effect is tenable by comparing Model 1 and Model 3 (df=2).

$$P(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}$$

$$\text{logit}\big(P(\theta)\big) = \log\left(\frac{P(\theta)}{1-P(\theta)}\right) = a_i(\theta - b_i) \equiv \alpha_i + \beta_{i1}\theta$$

$$\text{Model 1:} \quad \text{logit}\big(P(x_i = 1)\big) = \alpha_i + \beta_{i1} \cdot \theta$$

$$\text{Model 2:} \quad \text{logit}\big(P(x_i = 1)\big) = \alpha_i + \beta_{i1} \cdot \theta + \beta_{i2} \cdot \text{Group}$$

$$\text{Model 3:} \quad \text{logit}\big(P(x_i = 1)\big) = \alpha_i + \beta_{i1} \cdot \theta + \beta_{i2} \cdot \text{Group} + \beta_{i3} \cdot \theta \cdot \text{Group}$$

Note. θ is used as a proxy for the examinee's latent ability and corresponds to the total observed score in logistic regression DIF models.

The Logistic Regression method has the advantage of directly comparing models, while it usually deals with dichotomous data. For each studied item, dichotomous item scores are required to serve as the dependent variables.

Likelihood Ratio Test

The Likelihood Ratio Test for DIF was introduced by Thissen, Steinberg, and Wainer (1988). This method can be used to examine both uniform DIF and non-uniform DIF for both dichotomous and polytomous items. Items are examined for DIF by comparing the log-likelihood values of a series of nested models, for example, a more constrained model (such as assuming no DIF on the target item, which means the item parameters are constrained to be equal between two groups) and a more relaxed model (such as assuming DIF on the target item, which means the item parameters allowed to vary across two groups). The likelihood ratio test (LRT) is then used to compare these two models. If the change in the log-likelihood value between the constrained and unconstrained models is significant, it indicates the target item has

DIF. The significance is determined by comparing the change in the log-likelihood to a chi-square distribution with degrees of freedom equal to the number of parameters of interest.

The free-baseline anchor approach is considered one of the most effective implementations of the Likelihood Ratio Test (LRT) for detecting DIF, particularly when multiple DIF items are present (Stark et al., 2006; Wang & Yeh, 2003). When there is no prior knowledge about which items exhibit DIF or which items can serve as anchors, the constrained-baseline method can be initially employed to select appropriate anchors. This involves fitting an overly constrained baseline model, which assumes all items are DIF-free, and comparing it with a slightly less constrained model that drops one constraint at a time. By computing and comparing the log-likelihood values for both models, items with the least DIF can be identified and used as anchors. Once the anchors are selected using the constrained-baseline method, the free-baseline method is applied to detect DIF items. This method involves fitting a baseline model where the parameters of the anchor items are held equal across groups, ensuring these items are DIF-free, while all other items' parameters vary across the focal group and the reference group. The comparison models are created by constraining one item at a time in addition to anchor items. Comparison models are compared with the baseline model by using the likelihood ratio test. Significant changes in the log-likelihood values between the baseline model and comparison models indicate the presence of DIF. With anchor items, the free-baseline anchor approach provides a more stable and accurate detection process when multiple DIF items exist.

Raju's Area

Raju's area method was originally proposed by Raju (1988) in the form of formulas computing the exact areas between item characteristic curves (ICCs) of different groups for one-,

two-, and three-parameter IRT models. By fitting IRT models to data of the focal group and the reference group, item parameters can be estimated on a common metric and used to generate the ICC of each item for both groups. If the area difference is significantly different from 0, it indicates that the item functions significantly differently across those groups, indicating DIF.

The area between ICCs of the two groups can be calculated as the signed or unsigned area (Raju, 1988). The signed area, either negative or positive, takes the direction of the DIF into account, while the latter measures the absolute value of the area difference without indicating the direction of the DIF.

Raju's area method can be adapted widely with various IRT models. However, Raju (1988) illustrated some concerns regarding the area method. The areas calculated based on item parameters are subjected to sampling fluctuations. In other words, the area varies from one sample to the other due to the sampling error, especially when the sample size is small. This is because the area calculated is based on estimated item parameters, which suffer from sampling fluctuations as well. Moreover, how to calculate the exact area differences rather than the estimated area differences has been discussed. Especially, the calculation of area differences becomes less accurate for 3PL ICCs when the guessing parameters are not equal. Though area estimates can be obtained by integrating areas between two finite points on the theta scale, the estimates are always smaller than the exact area, and the appropriate ability interval is ambiguous.

$$\text{Signed Area (SA)} = \int_{-\infty}^{\infty} (F_1 - F_2)\ \mathrm{d}\theta$$

$$\text{Unsigned Area (UA)} = \int_{-\infty}^{\infty} |F_1 - F_2|\ \mathrm{d}\theta$$

Gender Differential Item Functioning (DIF) in Mathematics Assessment

Mathematics assessments are critical globally as they play an essential role in evaluating students' mathematical capacity within the educational field. Scholars have extensively researched various types of Differential Item Functioning (DIF) and their sources in mathematics assessments, particularly DIF of different demographic group memberships. For example, items tend to function differently for English Learners (ELs) and non-English Learners (non-ELs) in math word problems because of certain linguistic features increasing the linguistic complexity of items, such as infrequent/unfamiliar vocabulary (Martiniello, 2008), passive voice, conditional clause (Banks et al., 2016), and abstract presentations (Buono & Jang, 2021). The increased linguistic complexity level causes heavier cognitive loading, thus leading to items favoring against EL students.

Gender DIF, which occurs when items function differently between gender groups, is another significant issue that has been extensively researched. Various potential sources of gender DIF in Mathematics assessment have been studied, such as item type, content domain, cognitive demand, and item difficulty.

In the aspect of item type, dichotomously scored items tend to favor males and polytomous items tend to favor females, and an item-by-format interaction may exist where females perform better on constructed response items (Henderson, 2001). Taylor and Lee (2012) and Shear (2023) drew a similar conclusion that multiple-choice items generally favored males while constructed-response items generally favored females by analyzing the math tests from a state testing program and the PISA respectively. However, there was also research indicating that generally both multiple-choice and open-ended items were equally difficult for both boys and girls (Yan, 2005).

Regarding content domain, geometry, and mathematics reasoning items were more difficult for females while algorithmic, computation-oriented items were relatively easier for females (Doolittle & Cleary, 1987). The content analysis conducted by Taylor and Lee (2012) after their DIF analysis indicated that items that measured geometry, probability, and algebra favored males, and items that measured statistical interpretations, multi-step problem-solving, and mathematical reasoning favored females. Moreover, female students tended to perform relatively better on items that were abstract or included variables such as *X*, *a*, or *b*, while male students performed relatively better on real-life word problems (Harris & Carlton, 1993). Yan (2005) had similar findings with 1132 U.S. 8th grade students who took TIMSS 1999 math test: girls significantly outperformed boys on algebra items, whereas boys significantly outperformed girls on measurement items. Additionally, items related to statistics and probability were also more likely to have real-world contexts and more character counts, tending to favor males (Steedle et al., 2023).

Cognitive demand was researched as a potential source for gender DIF as well. Gierl et al. (2003) adopted a multidimensionality-based DIF analysis to explore content- and cognition-related sources of gender differences in mathematics achievement based on the taxonomy proposed by Gallagher et al. (2000) and their findings indicated that, among Grade 9 students, math items related to spatial skills favored boys, while those associated with memorization skills favored girls. Moreover, sizable gender differences in Space and Shape items (one of four main mathematics domains assessed by PISA which related to the understanding of spatial and geometric phenomena and relationships (OECD, 2000)) were reported by Liu and Wilson (2009), which resonated with Gallagher et al. (2000) that boys showed better spatial ability. Shanmugam (2020) analyzed the performance of non-English-speaking students in Malaysia on

20 selected Grade 8 computation math items from TIMSS 1999 and 2003, concluding that items assessing lower-order thinking skills (LOTs, solving familiar "textbook" problems using well-known algorithms) tend to favor girls, while items assessing higher-order thinking skills (HOTSs, such solving non-routine problems using unrehearsed algorithms) tend to favor boys.

Currently, DIF research still suffers from a scarcity of well-thought-out studies that draw on learning science and cognitive theories to articulate potential sources of DIF. According to Gierl et al. (2003), while statistical DIF analyses are typically followed by a judgmental review to understand why items exhibit DIF, this approach has made limited progress in explaining the underlying causes. Referencing the Standards for Educational and Psychological Testing (1999), the authors stated that "once items on a test have been statistically identified as functioning differently from one examines group to another, it has been difficult to specify the reasons for the differential performance or to identify a common deficiency among the identified items" (p. 78). Therefore, this study seeks to address this gap by integrating cognition and sociocultural learning theories to investigate a broader range of item features attributed to DIF estimates.

Current Study

This study aims to provide insights into the item and booklet quality of the large-scale math assessment, and further explore the potential sources of gender DIF among U.S. students within the context of eTIMSS 2019, an international large-scale mathematics assessment conducted in 2019. eTIMSS 2019 introduced new item types emphasizing problem-solving in real-world contexts and marked the first time U.S. students took the TIMSS math assessment in a newly designed digital format, offering a meaningful context for this study. Specifically, the following research questions will be studied:

1. What items in the eTIMSS 2019 Grade 8 math assessment exhibit gender DIF (boys vs girls) among U.S. students?
2. How are gender DIF estimates associated with various math item features?

Method

Data Source

The data for this study is drawn from the Trends in International Mathematics and Science Study (TIMSS) 2019. As a well-established international assessment conducted by the International Association for the Evaluation of Educational Achievement (IEA), this international effort monitors the math and science achievement trends over the world every four years since 1995. The TIMSS database covers a wide range of data, including student achievement data as well as student, home, teacher, school, and national context data (Fishbein et al., 2021). It also releases limited restricted-use item data to users who require access, which facilitates the analysis of concrete assessment items. Therefore, TIMSS data is a valuable resource for researchers to study math assessment.

TIMSS 2019, the seventh cycle of the study, was conducted in 64 educational systems for Grade 4 and 46 for Grade 8. This cycle marked the transition to a digital assessment format (eTIMSS), with about half of the participating countries administering the assessment via computer, while the rest used paper and pencil (Martin et al., 2020). The United States was among the countries that implemented eTIMSS. eTIMSS included not only the digital counterpart of regular TIMSS items but also additional innovative Problem-Solving and Inquiry (PSI) tasks. PSI items provided extended coverage of mathematics and science frameworks by simulating real-world contexts and requiring students to integrate and apply both process skills and content knowledge to solve mathematical problems. As a result, PSI items were substantially

more complex than the standard eTIMSS items (Martin et al., 2020, p. 5, Chapter 4). Besides the features of the transition to digital and the inclusion of new types of items, as the last pre-pandemic year, the 2019 achievement data offers valuable insights into students' math abilities and educational outcomes before the disruptions caused by COVID-19.

This study focuses on the student achievement data of the eTIMSS Grade 8 Math assessment of the United States. eTIMSS 2019 Grade 8 Math assessment covers four major content domains: Number, Algebra, Geometry, and Data and Probability, with the target percentages of 30%, 30%, 20%, and 20% within the assessment respectively. This assessment also required participating students to draw on three cognitive domains: knowing, applying, and reasoning (Mullis & Martin, 2017, p. 22). Several sub-domains were described in detail in the assessment framework under each content and cognitive domain to ensure this assessment was developed appropriately.

Consistent with many other large-scale testing programs, the TIMSS 2019 Grade 8 mathematics assessment used matrix sampling for its booklet design. Specifically, items were first grouped into 14 blocks, with about 12 to 18 items per block, by matching the content and cognitive domain distribution across the overall item pool. Then, blocks were arranged within 14 booklets so that each block appeared in two booklets. Each student only needs to complete one booklet. eTIMSS 2019 Grade 8 Math assessment as the counterpart of the regular paper-and-pencil format TIMSS assessment not only included 14 blocks of regular items but also included 2 blocks of PSI items, which means that the first 14 blocks of regular items were arranged into 14 booklets as the regular assessment, while the last two blocks of PSI items were arranged into 2 PSI booklets.

Though the eTIMSS Grade 8 Math assessment includes 16 booklets consisting of 16 blocks, only three blocks, two blocks of regular items (ME02 and ME06), and one block of PSI items (MI01), were available for the concrete items. That is to say, given this study aimed at linking DIF results to assessment item features, this study had to focus on booklets including three accessible blocks, which are Booklet 1, 2, 5, 6, 15, and 16. Among these six booklets, Booklet 1 and Booklet 2 both contained Block ME02; Booklet 5 and Booklet 6 both contained Block ME06; Booklet 15 and Booklet 16 both contained Block MI01. Because of these shared blocks, the response data of all six booklets were grouped into pairs based on their shared blocks to facilitate the analysis, as bi-booklets, such as Booklet 1&2, Booklet 5&6, and Booklet 15&16. This organization allowed for consistent comparison of DIF results between the booklets that shared the same blocks and provided a larger sample size for more robust DIF analysis.

There was a total of 142 unique items across all six booklets, comprising 61 Multiple Choice (MC) items and 81 Constructed Response (CR) items. Each booklet included a combination of multiple-choice and construction response items, with total item counts ranging from 30 to 45. Regular booklets had a balanced mix of MC and CR items with Booklets 1 and 2 containing the most items overall. In contrast, PSI booklets (Booklets 15 and 16) had the fewest MC items and relied heavily on CR items, indicating the variation in item type distribution across booklets. One item (ME62342) from Block 6, which appeared in both Booklet 5 and Booklet 6, had no student response data and was therefore excluded from this study.

Table 3

eTIMSS 2019 Math Assessment Booklet Design

Block ID / Booklet ID	ME01	ME02†	ME03	ME04	ME05	ME06†	ME07	ME08	ME09	ME10	ME11	ME12	ME13	ME14	MI01†	MI02
1*																
2*																
3																
4																
5*																
6*																
7																
8																
9																
10																
11																
12																
13																
14																
15*																
16*																

Note.

(1) Superscript notation: booklet ID with * indicates the booklet is included in the DIF analysis of this study, and block ID with † indicates the block is included in the item feature analysis of this study.

(2) Color coding: this table uses a four-color scheme to represent the positioning of math blocks within each booklet.
Dark green: first Math Block in Part 1; Light green: second Math Block in Part 1;
Dark blue: first Math Block in Part 2; Light blue: second Math Block in Part 2.

(3) Booklet structure: each booklet contains two blocks of math items and two blocks of science items. In half of the booklets, two math blocks appear in Part 1 and two science blocks appear in Part 2; in the other half of the booklets, the order is reversed for math blocks and science blocks.

Table 4

Item Type Distribution by Booklet

Booklet ID	Multiple Choice (MC)	Constructed Response (CR)	Total Items
1	26	18	44
2	23	22	45
5	16	18	34
6	16	16	32
15	2	28	30
16	2	28	30

Note. One item (ME62342) appeared in both Booklet 5 and Booklet 6 but had no student response data, thus it was not included in the item counts.

For some items, students were required to provide more than one answer or a multi-part response (Fishbein et al., 2021, p. 67). These items were referred to as derived items, while their individual parts were called sub-items. There were 10 derived items across all six booklets along with 28 sub-items. For example, ME72198A and ME72198B were two 1-point items corresponding to the x-axis and y-axis coordinates for the point asked by the same question. ME72198 was derived from those two sub-items and was worth 1 score point. Given that IRT application requires adherence to the local independence assumption, and items should be treated as a whole when coding item features later, derived items were retained, and all sub-items were removed. As a result, 114 items were retained for the DIF analysis.

The raw data of student responses to items of their booklets were scored by using the SPSS score program provided by the TIMSS 2019 International Database (Fishbein et al., 2021, p.100). All single-part items were worth one scoring point, while derived items could be worth either one or two score points according to their pre-determined scoring guide by TIMSS 2019 (Fishbein et al., 2021, p. 67). For each bi-booklet, the count of total items and the total scores are shown in Table 5. PSI booklets (Booklets 15&16) contained the lowest score points and the least

counts of both derived and single-part items, compared with regular bi-booklets. This distribution highlighted the variation of item composition and scoring across bi-booklets.

Table 5

Summary of Item Count and Scores for Bi-Booklets

Booklets	Count of Derived Items	Count of Sub-Items	Count of Binary Items	Count of Graded Items	Total Items	Total Score
B1 & B2	5	15	44	2	46	48
B5 & B6	3	8	41	2	43	45
B15 & B16	2	5	22	3	25	28

To facilitate the use of Logistic Regression for DIF analysis, all student response data, except those for binary items, was further recoded into an alternative binary form. For the graded items that initially had scores of 0, 1, or 2, scores of 0 and 1 were recoded to 0, indicating that the student did not fully answer the item correctly. A score of 2 was recoded to 1, indicating that the student provided the fully correct answer.

This recoding implies that only students who received full points (a score of 2) for an item were considered to have answered it correctly (coded as 1). Partial credit (a score of 1) was treated similarly to an incorrect answer (a score of 0), both recoded to 0. By this approach, all items were transformed into a binary format, where 0 represents an incorrect or partially correct response and 1 represents a fully correct response.

As this study intended to detect gender DIF among all the items, gender information was required. In the student context data, gender was coded as a nominal variable with the categories of boy and girl. By linking student score data to student context data by unique student ID numbers, each student's scored responses were matched with their gender.

Participants

9,944 U.S. students participated in the eTIMSS 2019 Grade 8 Mathematics assessments in total. 285 students didn't indicate their gender information in their context data. Considering that gender DIF detection was part of the study objectives, students with gender information were filtered and their data was retained.

Since the entire math assessment costs 90 minutes for Grade 8 students, students could be unmotivated to provide quality responses thus contaminating the data. By referring to a derived variable from the eTIMSS codebook known as responder classification, motivated student responses could be filtered. Responder classification categorizes students into three groups: reached all items, indicating that students attempted to finish all items and they did finish; ran out of time, which means students who attempted but could not complete all items due to time constraints; stopped responding, which means students who had remaining time but chose not to complete the assessment. As this behavior suggests disengagement from the assessment, students who were classified as stop responding were deleted from the sample.

Table 6

Participant Distribution by Gender and Booklet

Booklet ID	N of Boys	N of Girls	N of Total
1	319	263	582
2	311	265	576
5	277	303	580
6	290	320	610
15	285	273	558
16	274	268	542
Total	1756	1692	3448

Note. N = number of participants.

Finally, the sample size for each studied booklet is shown in Table 6. The total sample consisted of 3448 students with 1756 boys and 1692 girls distributed across six target booklets. The number of students per booklet ranges from 542 to 610, relatively balanced between both gender groups.

DIF Analysis Procedures

Four DIF detection methods were selected for this study: The Mantel-Haenszel method, logistic regression, likelihood ratio test, and Raju's area method. By applying these four different methods, the outcomes can be cross-validated in order to identify and quantify potential DIF across gender groups. These methods provided diverse insights into how items may behave differentially between boys and girls.

Mantel-Haenszel Method

The difMH function from the difR package was used to operate the Mantel-Haenszel method. The boy group was set as the reference group and the girl group was set as the focal group. Benjamini-Hochberg (BH) correction was used to control the False Discovery Rate (FDR) when performing the MH method because when conducting multiple comparisons or statistical tests, the risk of incorrectly rejecting a true null hypothesis increases (Benjamini & Hochberg, 1995). αMH, which is referred to as the common odd ratio, as a measure of each item's effect size for evaluating the DIF magnitude from the MH method output was used to get the MH D-DIF estimates by the formula MH D-DIF estimates = -2.35 In (α_{MH}). If MH D-DIF is equal to 0, then no DIF is detected. Negative MH D-DIF indicates the item favors the reference group and disadvantages the focal group; Positive MH D-DIF indicates the item favors the focal group and disadvantages the reference group.

Logistic Regression

The difLogistic function from the difR package was used to perform the Logistic Regression method. This method was applied to the alternative binary student score data. Again, the referenced group was set as boys and the focal group was set as girls. DIF was assessed in both directions to determine favoring which group. The Benjamini-Hochberg (BH) correction was applied to adjust p-values and control for the False Discovery Rate (FDR). The effect size (ΔR^2) from the output was used as the DIF estimates for the Logistic Regression method. This measure reflects the impact of the group variable (in this study is gender) on the item response, indicating the magnitude of DIF for each item. By referring to the Item Characteristic Curve (ICC) plot of two gender groups for each item, the DIF type can be identified.

Likelihood Ratio Test

The multipleGroup function from the mirt package was utilized to fit multiple-group IRT models for items in each bi-booklet. The IRT model selection for different types of items was informed by the TIMSS. According to TIMSS 2019 Technical Report, TIMSS used the three-parameter logistic (3PL) IRT model for multiple-choice items, the 2PL IRT model for constructed response items worth 1 score point, and the generalized partial credit model for constructed response items worth up to 2 score points (Martin et al., 2020, p. 4, Chapter 11). Consequently, multiple groups of items were fitted with their corresponding IRT models.

The constrained-baseline method was applied to select anchors for each bi-booklet at first. The initial multiple-group IRT model allowed for group differences in means and variances to vary freely, while constraining all item parameters across groups. The DIF function from the mirt package was used to evaluate DIF for each item based on likelihood ratio tests. For each item, the chi-square statistic and degrees of freedom were computed by comparing a constrained

baseline model in which item parameters (i.e., a_1, d_1, d_2, and d) were allowed to vary across groups. The ratio of the chi-square statistic to degrees of freedom was calculated to identify items with minimal DIF as anchors. The proportion of anchor items selected per bi-booklet was approximately 15–20%, which was the minimum required to ensure model convergence.

Next, a second model was fitted using the free-baseline method. The "multipleGroup" function in the mirt package was used to fit a baseline model where the parameters of the anchor items were held constant across groups, while the parameters of all other items were allowed to vary between the groups. The comparison models were created by sequentially adding constraints for the key item parameters (i.e., a_1, d_1, d_2, and d) of one non-anchor item at a time and compared with the baseline model by using the likelihood ratio test using the DIF function in the difR package. Significant changes in the log-likelihood values between the baseline model and comparison models indicate the presence of DIF.

Raju's Area Method

Raju's area is calculated by measuring the area between the item characteristic curves (ICCs) of two gender groups. There are two approaches to calculating this area: signed and unsigned. In this study, the unsigned area was used considering its applicability to both uniform and non-uniform DIFs, representing the total magnitude of the differences between the two groups' ICCs, regardless of which group was favored at different ability levels. Item parameters of two gender groups were extracted from the output of the Likelihood Ratio Test. For 2PL models, ICCs were generated by using the item discrimination (a) and difficulty (b) parameters; For 3PL models, an additional guessing parameter (c) was included; in GPCM models, ICCs were computed by taking into account the category-specific difficulty parameters (i.e., b_1 and b_2). The areas were calculated by numerically integrating the absolute differences in probabilities

across ability levels ranging from -6 to +6 so as to better approximate the entire distribution of abilities in the sample.

In general, all DIF detection methods provided measures to quantify DIF magnitude. The Mantel-Haenszel (MH) Method produced log-odds estimates (MH D-DIF), and Raju's Area Method provided numerical values representing the absolute area differences between item characteristic curves, both of which can be directly used as outcome variables in subsequent linear regression analyses. In contrast, delta R^2 (the change in pseudo R^2) from the Logistic Regression output and the chi-square statistic from the Likelihood Ratio Test (LRT) are measures of model fit rather than direct DIF effect size estimates. While they indicate the presence and significance of DIF, these statistics are less straightforward for use as outcome variables in linear models without additional transformation or interpretation.

Coding of Item Features

Exploring how gender DIF estimates are associated with various math item features is a crucial aspect of this study following the DIF detection process. To this end, item features were coded by drawing upon cognition and sociocultural learning theories that may influence DIF estimates of math assessment items.

The first step in item feature coding was the creation of a comprehensive codebook. Various item features related to the item stem, options, entire items, and visual components of math items were identified and defined within this codebook based on relevant literature (Buono & Jang, 2021; Mahoney, 2003; Martiniello, 2008, 2009; O'Neill & McPeek, 1993). Features of the stem included aspects such as stem length, count of conditional clauses, and the presence of context. Features of the options included the option length and the number of equations, formulas, functions, or expressions. Features of the whole item included the involvement of

Table 7

Codebook for Sample Item Features

Item Feature by Category	Definition	Value Scheme Detailed
Stem		
Stem length	The total number of words contained in the stem of an item, where a "word" is defined as any sequence of characters separated by spaces, excluding punctuation that is directly attached to words	Numeric
Count of conditional clauses	The count of condition statements within the stem that have conditional markers such as "if" or "then"	Numeric
Presence of context	Information is presented in narrative structures such as a scenario or a story	0: No; 1: Yes
Options		
Option length	The total number of words for the total item options (including any instructions for how to operate the player to respond)	Numeric
Count of equation/ formula/function/ expression in option	The count of equation/formula/function/expression in all options of each item	Numeric
Item		
Involvement of construction	The question requires the transformation, construction, or creation of a mathematical object	0: No; 1: Yes
Count of metric system unit	The total number of measurement units used in the item that is part of the metric system (e.g., metric units such as meters, liters, or grams)	Numeric
Visual		
Presence of visuals	The presence of visual components used in the item	0: No; 1: Yes
Interactive visual representation	the existence of visual components that can be drawn, erased, reset, etc.	0: No; 1: Yes

construction and the count of metric system units, among others. Visual element features included the presence of visuals, interactive visual components, etc. The codebook provided detailed definitions of all item features and their corresponding value scheme (see Appendix A Codebook Codes for details and Table 7 for selected features). Following the revised codebook, item features were coded through consensus coding and test-retest coding, ensuring both inter-rater and intra-rater reliability. Two trained coders participated in the coding process, both with three years of K-12 teaching experience. One coder brought expertise in educational assessment,

Table 8

One Mock-up Item and Selected Sample Codes of Item Features

On Saturday, the lowest temperature in City A was 14 °C and the lowest temperature in City B was -10 °C. What was the difference between the lowest temperatures in the cities?

Answer: [ ] °C

Item Feature	Code	Reasoning
Content domain	Number	Coded by TIMSS
Topic area	Integers	Coded by TIMSS
Cognitive domain	Knowing	Coded by TIMSS
Format	Construction response	Coded by TIMSS
Stem length	32	32 words contained in the stem
Count of equation/formula/function/ expression/variable within the stem	0	No equation/formula/function/expression/variable in the stem
Count of clauses in the stem	0	No passive clauses, conditional clauses, or relative clauses in the stem
Presence of visual components	No	No visual components are provided
Involvement of the three-dimensional concept	0	No three-dimensional mathematical object contained or three-dimensional reasoning is needed to solve the problem
Involvement of construction	0	No transformation, construction, or creation of a mathematical object is required
Presence of the context	Yes	Information is presented in narrative structures with a concrete location and date, such as “On Saturday” and “the lowest temperature in City A”
Count of metric system units	2	°C is a metric system unit. It occurs twice in the stem

while the other specialized in curriculum and instruction, with a focus on language and literacy. A mock-up item is presented in Table 8 to illustrate the reasons for coding decisions. Some of the item features, such as content domain, topic area, cognitive domain, and format had been coded by eTIMSS 2019. Codes of other item features of this mock-up item and their reasoning

are reported in Table 8. Although there were 114 items across all six booklets in the DIF detection phase, only three blocks containing 41 restricted-use items were fully released by the IEA for consensus coding to analyze item features.

Analysis of DIF Estimates and Item Features

Following the DIF analysis and item feature coding, the next step involved analyzing how DIF estimates produced each selected method were associated with item features within the context of this study. Specifically, different DIF detection methods had different types of outputs: Mantel-Haenszel method provided the D-DIF effect size, logistic regression had odds ratio as a continuous measure of DIF magnitude, and Raju's area quantified DIF through the unsigned area between item characteristic curves (ICCs), while likelihood ratio test had chi-square statistics and associated *p*-values, indicating the statistical significance of DIF without providing direct effect size. DIF estimates and item features were linked for each restricted-use item. Stepwise linear regression was applied to examine the relationship between DIF estimates of each DIF detection method and the selected item features, providing insights into how specific characteristics of the items may contribute to differential item functioning.

Based on previous research and the consideration of eTIMSS's features as a global and digital-format assessment, eight features were ultimately selected or composited for the linear regression analysis. Among these item features, some were individual variables that were directly coded, while some were composite variables. Composite variables were created by summing the individual variables within the same category together since some coded item features in this study shared similarities. For example, the count of clauses was calculated by adding three kinds of clauses (independent clauses, dependent clauses, and relative clauses), to obtain a total count of clauses for each item. They were relatively independent of each other,

thereby minimizing the risk of multicollinearity. This approach allowed for a clearer interpretation of each broad type of item feature's contribution to the model without overlapping effects. These eight item feature variables are listed and described in Table 9 below.

Table 9

Selected Item Features for Linear Regression

Item Feature	Definition
Stem length	The total number of words contained in the stem of an item, where a "word" is defined as any sequence of characters separated by spaces, excluding punctuation that is directly attached to words
Count of equation/formula/function/ expression/variable within the stem	The composite variable represents the total number of mathematical elements, including equations, formulas, functions, or expressions, mentioned in the stem, combined with the number of distinct variables appearing in the stem.
Count of clauses in the stem	Composite variable by adding the count of passive clauses, conditional clauses, and relative clauses within the stem together.
Presence of visual components	Whether any visual element was included or not in the item
Involvement of the three-dimensional concept	Whether the item contained a three-dimensional mathematical object or needed three-dimensional reasoning to solve the problem
Involvement of construction	Whether the item required the transformation, construction, or creation of a mathematical object
Presence of the context	Whether information within the item was presented in narrative structures, such as a scenario or a story
Count of metric system units	The total number of measurement units used in the item that is part of the metric system (e.g., metric units such as meters, liters, or grams)

Note that although more item features were coded in the initial step, due to the similarity to other item features, insufficient variation, or the presence of these features in only a few items, they were not considered for the subsequent linear analysis. For instance, no items had the feature of unit conversion, where the question required converting between units of measure to solve. Additionally, the presence of interactive visual components overlapped with the presence of visual components, so this feature was also discarded.

Results

Descriptive Statistics by Booklet and Gender

Descriptive Booklet-Level Statistics

Considering this study analyzed the data by bi-booklet, the descriptive statistics were also examined at the booklet level based on these bi-booklet groupings. This means that the data for Booklets 1&2, Booklets 5&6, and Booklets 15&16 were combined and analyzed together to provide insights into the performance of students across these groupings. Descriptive statistics, such as mean, median, standard deviation, and score ranges, were calculated separately for boys and girls within each bi-booklet group to capture assessment performance differences to ensure a more comprehensive understanding of the bi-booklet data while keeping the booklet pairings consistent for analysis. The descriptive statistics are shown in the following table (Table 10).

In Booklets 1&2, two girls missed all the items within the booklet they were assigned. Boys had a mean score of 13.9, which was slightly lower than the girls' mean score of 14.4. The score distribution between boys and girls was similar, with boys showing slightly more variability, indicated by a standard deviation (SD) of 8.05 compared to the girls' SD of 7.95. The range of scores for both gender groups span from a minimum of 0 (for boys) or 1 (for girls) to a maximum of 32.

Table 10

Descriptive Summary of Scores by Gender of Each Bi-Booklet

Booklets	Gender	Count	Mean	SD	Median	Min	Max
B1 & B2	Boy	628	13.9	8.05	13	0	32
	Girl	528	14.4	7.95	13	1	32
B5 & B6	Boy	567	13.4	7.17	12	1	30
	Girl	622	13.4	7.17	12	0	30
B15 & B16	Boy	559	8.46	6.76	6	0	27
	Girl	541	8.06	6.53	6	0	28

In Booklets 5&6, one girl missed all the items in the booklet she was assigned. Both gender groups had the same mean score of 13.4 and a median score of 12. The standard deviation (SD) was also identical at 7.17 for both groups, indicating a similar distribution of scores. The score ranges were similar, with girls having a slightly wider range, from 0 to 30, while boys' scores ranged from 1 to 30.

In Booklets 15&16, which are two PSI booklets, no students missed all items in their assigned booklets. Boys had a slightly higher mean score of 8.46 compared to girls' mean score of 8.06. Both gender groups had the same median score of 6, but boys exhibited greater variability with a standard deviation (SD) of 6.76 compared to girls' SD of 6.53. The minimum score for both groups was 0, while the maximum score was 27 for boys and 28 for girls.

Item-Level Statistics

Since the likelihood ratio test was used to detect DIF across multiple groups, it estimated item-level parameters. For multiple-choice items, a three-parameter logistic (3PL) IRT model was employed, which estimated the discrimination (*a*), difficulty (*b*), and guessing (*g*) parameters. For constructed response items scored with 1 point, a two-parameter logistic (2PL)

IRT model was applied to estimate the discrimination (*a*) and difficulty (*b*) parameters. For constructed response items scored up to 2 points, the generalized partial credit model (GPCM) was used to estimate the discrimination (*a*) parameter along with two difficulty parameters (b_1 and b_2). Two difficulty parameters represent the difficulty thresholds for transitioning between score categories. b_1 is the difficulty parameter for a student getting a score of 0 to 1, indicating the ability at which the student has a 50% probability of scoring at least 1 score point. Similarly, b_2 is the difficulty parameter to get a score of 1 to 2 indicating the ability at which a student has a 50% probability of scoring 2 points.

Generally speaking, all items had reasonable item parameters. Multiple-choice items demonstrated a greater ability to discriminate between different ability levels compared to 1-point and 2-point constructed response items, with the 2-point constructed response items having the lowest mean discrimination parameters. Additionally, most 2-point constructed response items had b_1 parameters higher than b_2, indicating that moving from 0 to 1 point required higher ability levels than moving from 1 to 2 points.

Each bi-booklet exhibited distinct patterns in *a* and *b* parameters across gender groups (Figure 1). In Booklets 1&2, difficulty parameters showed greater variability among boys, while discrimination parameters were relatively consistent across genders. In contrast, Booklets 5&6 demonstrated higher average difficulty parameters than Booklets 1&2, with greater variability in difficulty among boys and more variability in discrimination among girls. For both Booklets 1&2 and Booklets 5&6, a few items had negative difficulty parameters (*b*), indicating that these items were relatively easy, allowing students with lower ability levels to have a higher probability of answering them correctly. For 2-point constructed response items, b_1 parameters were mostly smaller than b_2, showing that moving from a score of 0 to 1 was easier than moving from a score

of 1 to 2. Booklets 15&16, which included fewer items, showed the highest average values for both difficulty and discrimination parameters compared to the other bi-booklets. Discrimination was more variable among boys, while difficulty parameters exhibited the least variability across both gender groups.

There were consistent patterns in mean scores by item type within and across all bi-booklets (Figure 2). In both Booklets 1&2 and 5&6, multiple-choice items had higher mean scores than 1-point constructed response items. Additionally, among all item types, 2-point constructed response items showed the least variability in mean scores in both bi-booklets. However, students performed better on 2-point items in Booklets 1&2 (mean = 0.65) than in Booklets 5&6 (mean = 0.50). Booklets 15&16 included mostly 1-point constructed response items, along with only one multiple-choice item and three 2-point constructed response items. All item types in this booklet had lower mean scores compared to the other bi-booklets, with averages of 0.18 for the multiple-choice item, 0.33 for the 1-point constructed response items, and 0.36 for the 2-point constructed response items. Importantly, the 2-point constructed response items in Booklets 15&16 exhibited the greatest variability in mean scores across all bi-booklets.

Figure 1

Item Parameters by Gender of Each Bi-Booklet

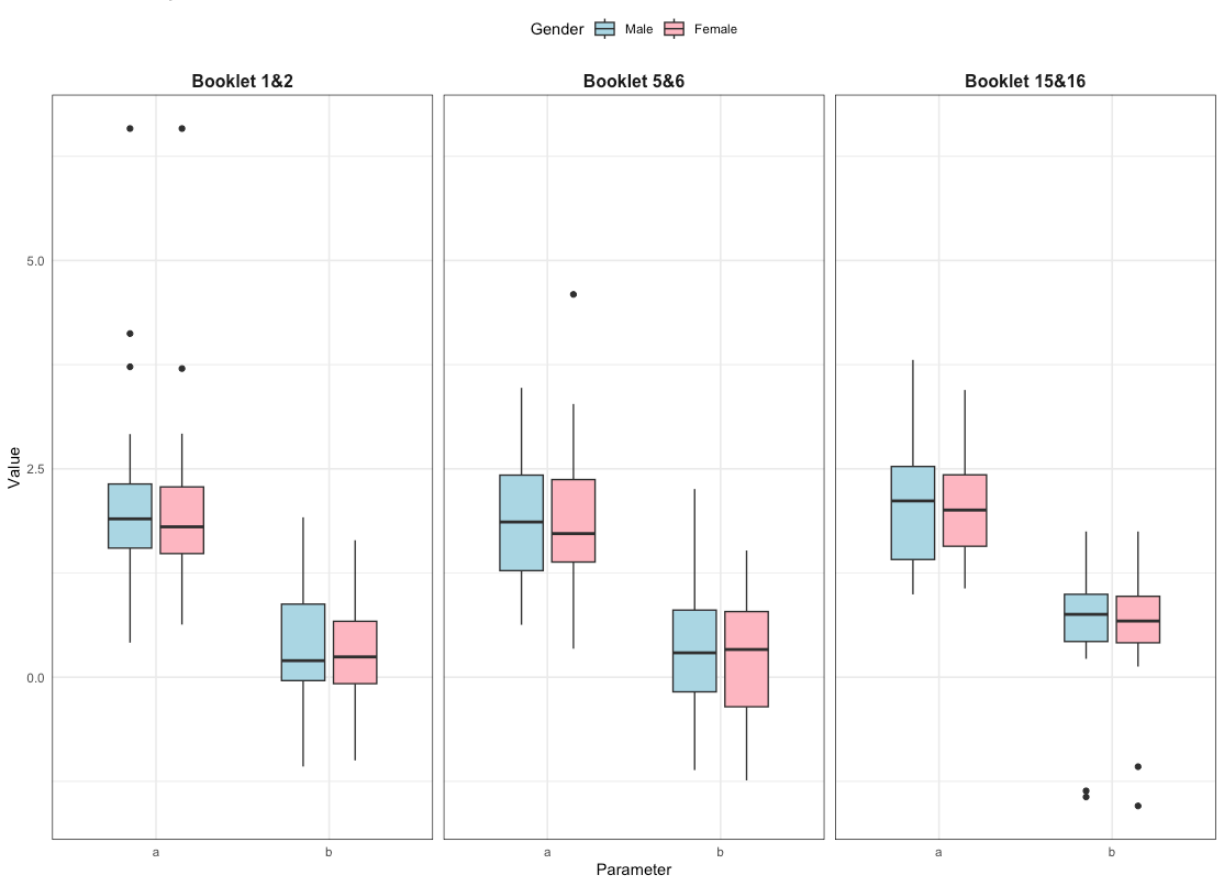
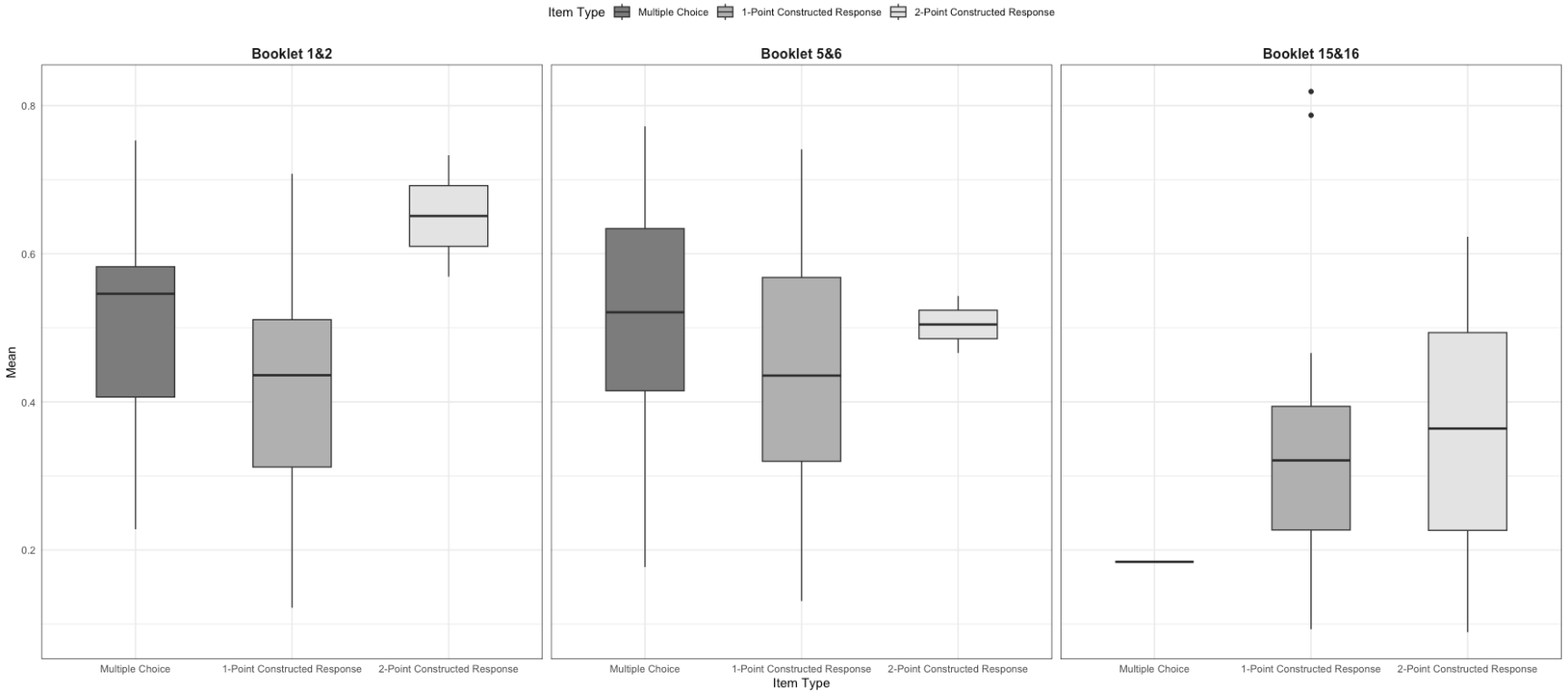


Figure 2

Mean Score by Item Type of Each Bi-Booklet



Gender DIF Results

All four DIF methods quantified the magnitude of DIF and contributed different insights regarding DIF detection. Among all four methods, the Mantel-Haenszel Method, Logistic Regression, and Likelihood Ratio Test successfully identified DIF items, which helped to address the first research question of this study: What items in the eTIMSS 2019 Grade 8 math assessment exhibit gender DIF (boys vs. girls) among US students. Logistic Regression demonstrated the greatest power to detect them, while the Mantel-Haenszel method and the Likelihood Ratio Test complemented the Logistic Regression results, providing additional confirmation.

Regarding the flagged DIF items, a total of 10 items out of 114 items were identified across all three detection methods, including both uniform DIF and non-uniform DIF. The two PSI booklets had the most identified DIF items, while no DIF items were detected in Booklets 1&2. The DIF items exhibited a mixed pattern with some items favoring girls and others favoring boys, with no consistent favor toward either group overall. MQ12D07A consistently exhibited significant differential functioning between boys and girls across all detection methods. ME62150 and MQ12D01 were flagged by two out of the three methods, indicating potential bias.

Mantel-Haenszel (MH) DIF Results

The Mantel-Haenszel method identified two DIF items, ME62150 from Booklet 5&6 and MQ12D07A from the PSI Booklets. ME62150 had a negative DIF estimate, indicating it favored the reference group, while MQ12D07A had a positive DIF estimate, meaning it favored the focal group. The MH method is limited to identifying uniform DIF and it matches groups based on total score to assess group performance differences.

Logistic Regression DIF Results

Logistic regression identified the most DIF items, with four items (ME52034, ME62150, ME52068, and ME62002) from Booklet 5&6 and five items (MQ12R02B, MQ12D01, MQ12D02A, MQ12D05, and MQ12D07A) from the PSI Booklets. Both uniform and non-uniform DIF were detected (Figure 3).

Likelihood Ratio Test DIF Results

The Likelihood Ratio Test flagged three DIF items, all from PSI booklets: MQ12D01, MQ12D07A, and MQ12B03 (Figure 4). The Logistic Regression also flagged MQ12D01 and MQ12D07A as non-uniform and uniform DIF items, respectively. Additionally, the Likelihood Ratio Test detected MQ12B03 as a new DIF item that had not been identified by the Mantel-Haenszel method or Logistic Regression.

Table 11

Summary of DIF Results

Booklets	Item	Mantel-Haenszel Method	Logistic Regression	Likelihood Ratio Test
B5 & B6	ME52034		Non-uniform	
B5 & B6	ME62150	DIF	Non-uniform	
B5 & B6	ME52068		Uniform	
B5 & B6	ME62002		Non-uniform	
B15 & B16	MQ12R02B		Uniform	
B15 & B16	MQ12D01		Non-uniform	Non-uniform
B15 & B16	MQ12D02A		Non-uniform	
B15 & B16	MQ12D05		Non-uniform	
B15 & B16	MQ12D07A	DIF	Uniform	Uniform
B15 & B16	MQ12B03			Non-uniform

Raju's Area Method DIF Results

Raju's area measures the absolute difference between the Item Characteristic Curves (ICCs) of two gender groups. However, it does not offer a threshold or criteria for classifying

Figure 3

Item Characteristic Curves (ICCs) for Logistic Regression DIF Analysis

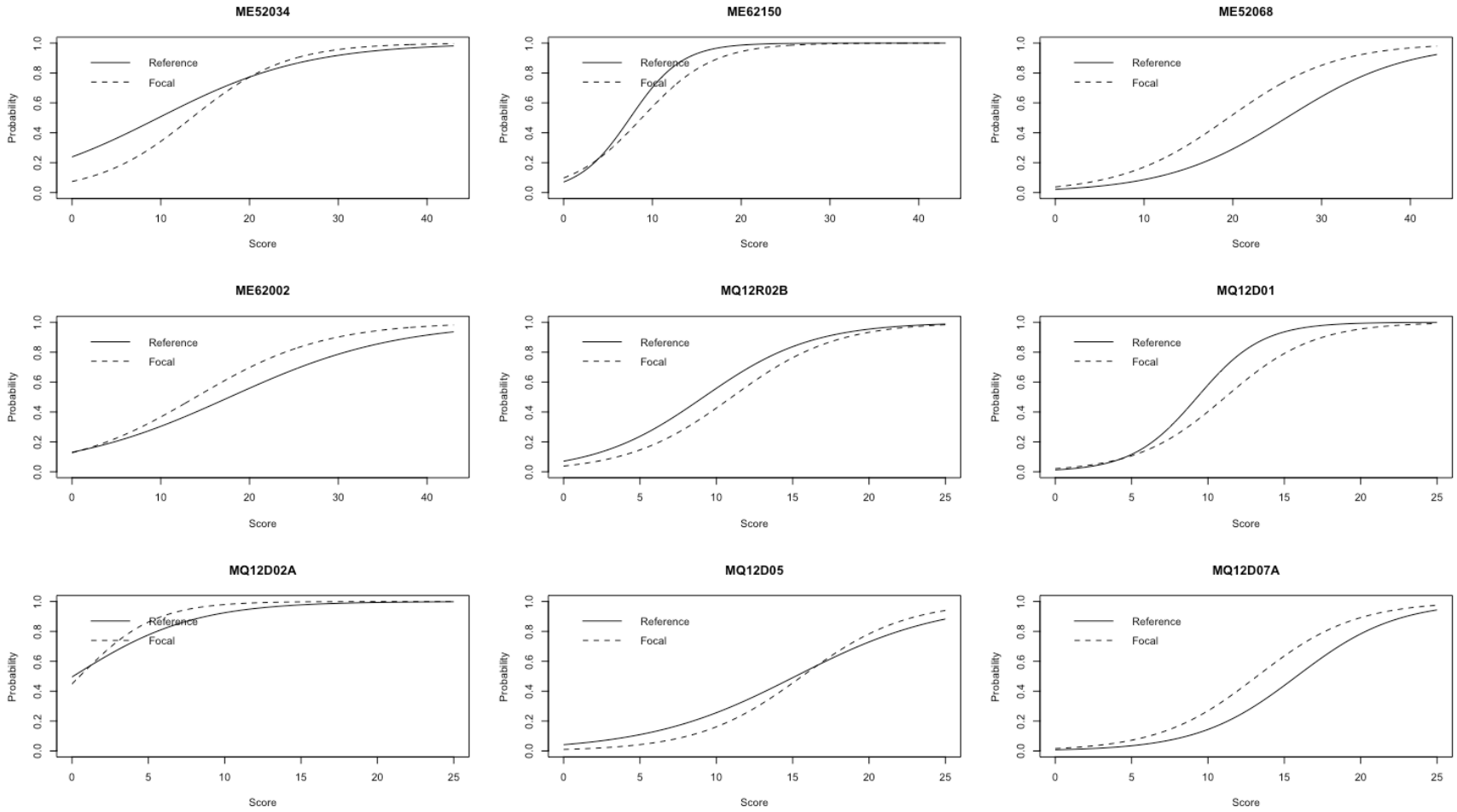
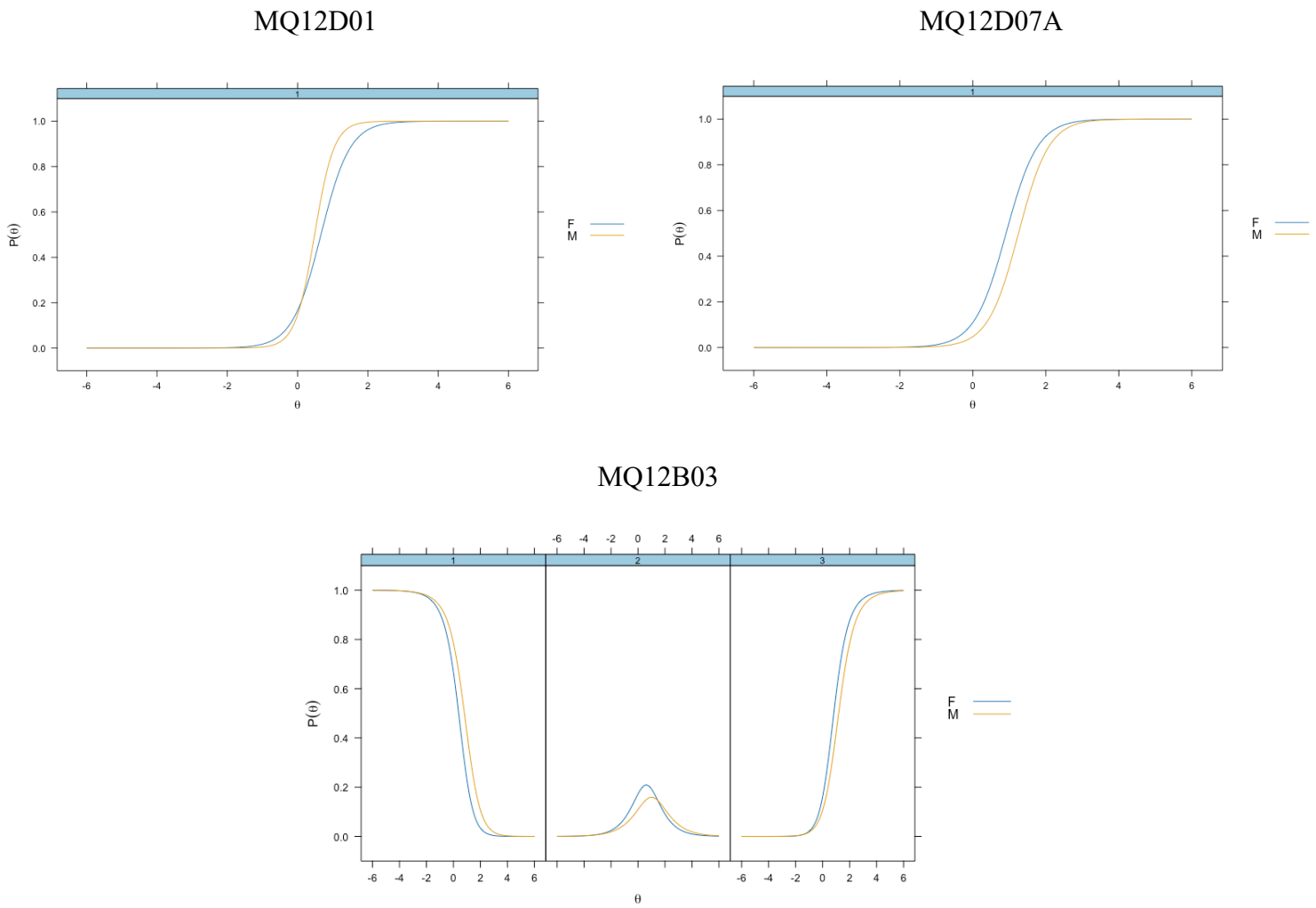


Figure 4

Item Characteristic Curves (ICCs) for Likelihood Ratio Test DIF Analysis



items as exhibiting DIF but instead provides estimates of the magnitude of DIF. For all anchor items, the Raju's areas were zero, while for non-anchor items, the Raju's area ranged from 0.09 to 2.10, with Booklets 5&6 having the largest mean area difference.

Linking DIF Estimates to Item Features

Following gender DIF analysis, DIF estimates can be obtained and linked to coded item features to address the second research question: How are gender DIF estimates associated with various math item features? Eight item features were used in the linear regression analysis, with DIF estimates modeled as a function of these features. Descriptive statistical analysis was conducted on the item features to gain a deeper understanding of the design of restricted-use eTIMSS 2019 math items. A total of 41 restricted-use items were drawn from three blocks: ME02 (15 items), ME06 (13 items), and MI01 (13 items). ME02 and ME06 are regular blocks, while MI01 is a PSI (Problem-Solving Item) block. Some item features were consistent across all three blocks, while each block also displayed unique characteristics (Table 12).

Descriptives of Item Features by Blocks

(1) The Length of the Stem

Stem length varied across all restricted-use items, ranging from a maximum of 100 words to a minimum of six words. Block ME02 and Block ME06 had moderate stem lengths, averaging 25 and 33 words, respectively. In contrast, Block MI01 from the PSI booklets had an average stem length of 57 words, which was more than double the average length of the multiple-choice items. Based on the stem length, the PSI block required a higher level of reading effort for students to process and thus respond to the items.

(2) The Count of Equation/Formula/Function/Expression/Variable within the Stem

Similar to stem length, PSI items included a higher average number of equations, formulas, functions, expressions, and variables. In contrast, Block ME06 had the fewest, as most items within this block did not contain these mathematical elements.

(3) The Count of Clauses in the Stem

The count of clauses is another aspect of linguistic complexity that may contribute to performance differences between genders. Aligned with stem length, Block MI1 had more clauses than the other two blocks, resulting in longer stem lengths for PSI items as the additional clauses introduced more words into the item stems. In contrast, most items in the regular booklets had succinct and easy-to-read stems, with few or no clauses included.

(4) The Presence of Visual Components

Visual components were heavily emphasized in all PSI items, with each item including at least one visual element. In contrast, in the regular blocks, about half of the items did not contain any visuals. The presence of these visual components added one more layer of cognitive load for students, requiring them to process both the visuals and the text.

(5) The Involvement of the Three-Dimensional Concept

Three-dimensional concepts are often associated with visual components, which help students understand and reason about these concepts more effectively. As a result, it was reasonable for more than half of the items in Block MI01 to involve three-dimensional concepts, whereas only a few items in the regular blocks, ME02 and ME06, included such concepts.

(6) The Involvement of Construction

Construction tasks require students to create or transform a mathematical object. This is more feasible with the digital version of TIMSS, as it allows students to perform object construction directly on a digital device. Incorporating construction tasks into the item design

helps to increase student engagement in mathematical thinking. Interestingly, there was a higher proportion of the regular items in Booklets 1, 2, 5, and 6 involving construction tasks compared to PSI items. I speculate this may be due to the consideration to balance the amount of cognitive demands associated with visual construction and task comprehension. Construction tasks typically require higher cognitive demand, therefore placing the construction tasks that typically require high cognitive demand in regular items that have fewer word counts in the item stem and thus reduced linguistic demand helps ensure a reasonable overall cognitive demand.

(7) The Presence of the Context

The presence of context supports students' understanding when solving abstract mathematical problems. All PSI items included a context, aligning with the design objectives by simulating real-world situations so that students can integrate and apply both process skills and content knowledge to solve mathematical problems. In comparison, about half of the regular items also included a context, highlighting the important role that context plays in math assessment item design to enhance student understanding.

(8) The Count of Metric System Units

Considering that eTIMSS is a global assessment, the units used are not commonly encountered by U.S. students in their daily lives, as the U.S. primarily uses the U.S. customary system of measurement, such as inches, feet, ounces, pounds, and degrees Fahrenheit. The presence of non-U.S. customary units, such as those from the metric system, could pose a challenge for students when solving math problems, as they may be less familiar with these measurements. Block ME06 had the highest average number of metric system units per item, with an average of 2, while Block ME02 had the fewest.

Table 12

Descriptive Statistics of Item Features Across Blocks

Block	Item Count	Stem Length	Count of Equations in Stem	Count Of Clauses in Stem	Presence of Visual Components*	Involvement of 3D Concept*	Involvement of Construction*	Presence of Context*	Count of Metric System Units
ME02	15	25.10 (15.60)	1.20 (1.90)	0.60 (0.83)	0.60 (0.51)	0.07 (0.26)	0.13 (0.35)	0.47 (0.52)	0.27 (0.46)
ME06	13	32.60 (19.20)	0.39 (1.12)	0.69 (1.32)	0.46 (0.52)	0.08 (0.28)	0.15 (0.38)	0.54 (0.52)	1.77 (2.31)
MI01	13	57.10 (25.40)	1.85 (2.08)	1.54 (1.33)	1.00 (0.00)	0.54 (0.52)	0.08 (0.28)	1.00 (0.00)	1.00 (1.53)

Note.

(1) Values are presented as Mean (Standard Deviation).

(2) Item features donated by an asterisk (*) are binary variables, coded as 0 for absence and 1 for presence.

Associations Between DIF Estimates and Item Features

As the Mantel-Haenszel Method and Raju's Area Method were employed to quantify the magnitude of Differential Item Functioning (DIF), the resulting DIF estimates from these two methods were then used to examine the association between DIF estimates and item features.

Stepwise linear regression was conducted to investigate the relationships between item features and DIF estimates, aiming to identify significant predictors. The lm() function and step() function from the stats package were used to fit the initial linear regression model and perform stepwise selection in both forward and backward directions. The stepwise linear regression results are presented in Tables 13 and 14.

In the final model for the Mantel-Haenszel method and item features, three variables were retained: the count of equations/formulas/functions/expressions/variables, the count of clauses, and the presence of context. The count of clauses in the stem showed a significant positive relationship with MH DIF estimates ($p < .05$), indicating that more clauses were associated with higher MH DIF magnitude. The presence of context in the stem exhibited a marginally negative association with MH DIF estimates ($p < .10$), suggesting a tendency for items with context to have lower DIF magnitude. Although the count of equations/formulas/functions/expressions/variables did not show a statistically significant relationship with MH DIF estimates, it was positively associated with MH DIF estimates.

In the stepwise linear regression of Raju's area and item features, three variables were retained in the final model: the involvement of construction, the presence of context, and the count of metric system units. The presence of context demonstrated a significant negative association with Raju's unsigned area ($p < .05$), indicating that items with contextual information tended to have smaller DIF magnitudes. Items involving the construction or creation of a

mathematical object showed a significantly larger Raju's unsigned area between gender groups ($p < .001$). The count of metric system units in the item had a weak positive association with Raju's unsigned area.

Discussions

Key Findings

This study not only detected DIF items in a large-scale mathematics assessment but also offered valuable insights into the potential sources of DIF, specifically how DIF estimates are associated with various item features. Extensive item features were examined, and significant item features were identified, aligning with previous research while also providing new perspectives on large-scale assessments in digital format.

Gender DIF Results

Four DIF methods were adopted for this study, each with unique strengths and limitations. The Mantel-Haenszel and Raju's Area methods effectively quantified DIF magnitude in a more interpretable way. Logistic regression was a more powerful method than the Mantel-Haenszel procedure in detecting DIF items, though with recoded data, which is consistent with findings from many other DIF studies (Swaminathan & Rogers, 1990). The Likelihood Ratio Test, as an IRT-based approach, provided a more precise evaluation of item parameters, surpassing other CTT-based DIF methods. The DIF items flagged by the four methods showed some overlap and could complement each other in detecting DIF, providing a more comprehensive understanding of item bias across two gender groups, which is consistent with previous research that there was close agreement among DIF detection procedures: Lord's Chi-square, Raju's Area Measures, and Likelihood Ratio Test (Kim & Cohen, 1995).

Table 13

Stepwise Linear Regression Results : Predicting the Mantel-Haenszel DIF Estimates by Item Features

Item Features	Model 1				Model 2				Model 3			
	Est (SE)	*t*	*p*	sr^2	*Est (SE)*	*t*	*p*	sr^2	*Est (SE)*	*t*	*p*	sr^2
Intercept	0.11 (0.24)	0.46	.65		0.13 (0.23)	0.55	.59		0.11 (0.23)	0.51	.61	
Stem length	0.00 (0.01)	0.25	.80	0.00	-	-	-		-	-	-	
Count of metric system units	-0.03 (0.06)	-0.44	.66	0.00	-0.02 (0.06)	-0.38	.70	0.00	-	-	-	
Construction	0.41 (0.32)	1.30	.20	0.04	0.41 (0.31)	1.32	.20	0.04	0.38 (0.30)	1.28	.21	0.04
Three-dimension	0.34 (0.29)	1.18	.25	0.03	0.36 (0.26)	1.38	.18	0.04	0.36 (0.26)	1.40	.17	0.04
Presence of visuals	-0.29 (0.22)	-1.30	.20	0.04	-0.28 (0.22)	-1.29	.21	0.04	-0.28 (0.22)	-1.31	.20	0.04
Number of equation and variable	0.07 (0.05)	1.34	.19	0.04	0.07 (0.05)	1.33	.19	0.04	0.07 (0.05)	1.39	.17	0.04
Number of clauses	0.09 (0.11)	0.80	.43	0.01	0.10 (0.08)	1.23	.23	0.03	0.11 (0.08)	1.31	.20	0.04
Presence of context	-0.36 (0.27)	-1.33	.19	0.04	-0.32 (0.21)	-1.51	.14	0.05	-0.34 (0.20)	-1.67	.10	0.06
Item Features	Model 4				Model 5				Model 6			
	Est (SE)	*t*	*p*	sr^2	*Est (SE)*	*t*	*p*	sr^2	*Est (SE)*	*t*	*p*	sr^2
Intercept	0.15 (0.23)	0.67	.51		0.07 (0.22)	0.31	.76		-0.01 (0.16)	-0.07	.94	
Stem length	-	-	-		-	-	-		-	-	-	
Count of metric system units	-	-	-		-	-	-		-	-	-	
Construction	-	-	-		-	-	-		-	-	-	
Three-dimension	0.33 (0.26)	1.27	.21	0.04	-	-	-		-	-	-	
Presence of visuals	-0.22 (0.21)	-1.03	.31	0.02	-0.11 (0.19)	-0.56	.58	0.01	-	-	-	
Number of equation and variable	0.06 (0.05)	1.18	.25	0.03	0.08 (0.05)	1.50	.14	0.05	0.08 (0.05)	1.57	.13	0.00
Number of clauses	0.13 (0.08)	1.64	.11	0.06	0.17 (0.08)	2.12	.04*	0.10	0.16 (0.08)	2.09	.04*	0.00
Presence of context	-0.40 (0.20)	-1.99	.06+	0.09	-0.34 (0.20)	-1.73	.09+	0.07	-0.33 (0.19)	-1.71	.10+	0.00

Note. SE = standard error; sr^2 = semi-partial R^2. + $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 14

Stepwise Linear Regression Results: Predicting the Raju's Area Estimates by Item Features

Item Features	Model 1				Model 2				Model 3			
	Est (SE)	*t*	*p*	sr^2	*Est (SE)*	*t*	*p*	sr^2	*Est (SE)*	*t*	*p*	sr^2
Intercept	0.43 (0.12)	3.61	.00***		0.43 (0.11)	3.97	.00***		0.43 (0.11)	4.03	.00***	
Number of equation and variable	0.00 (0.03)	0.02	.99	0.00	-	-	-		-	-	-	
Number of clauses	0.01 (0.05)	0.13	.90	0.00	0.01 (0.05)	0.14	.89	0.00	-	-	-	
Presence of visuals	-0.05 (0.11)	-0.46	.65	0.00	-0.05 (0.11)	-0.47	.64	0.00	-0.05 (0.10)	-0.51	.62	0.00
Three-dimension	0.07 (0.14)	0.53	.60	0.00	0.07 (0.13)	0.56	.58	0.00	0.07 (0.13)	0.57	.57	0.00
Stem length	0.00 (0.00)	-0.69	.50	0.01	0.00 (0.00)	-0.71	.48	0.01	0.00 (0.00)	-0.80	.43	0.01
Count of metric system units	0.06 (0.03)	1.90	.07+	0.06	0.06 (0.03)	1.94	.06+	0.06	0.06 (0.03)	2.02	.05+	0.06
Construction	0.52 (0.15)	3.40	.00**	0.18	0.52 (0.15)	3.48	.00***	0.18	0.53 (0.15)	3.61	.00***	0.19
Presence of context	-0.21 (0.13)	-1.59	.12	0.04	-0.21 (0.13)	-1.61	.12	0.04	-0.22 (0.13)	-1.72	.10+	0.04
Item Features	Model 4				Model 5				Model 6			
	Est (SE)	*t*	*p*	sr^2	*Est (SE)*	*t*	*p*	sr^2	*Est (SE)*	*t*	*p*	sr^2
Intercept	0.39 (0.09)	4.64	.00***		0.39 (0.08)	4.67	.00***		0.37 (0.08)	4.78	.00***	
Number of equation and variable	-	-	-		-	-	-		-	-	-	
Number of clauses	-	-	-		-	-	-		-	-	-	
Presence of visuals	-	-	-		-	-	-		-	-	-	
Three-dimension	0.05 (0.12)	0.43	.67	0.00	-	-	-		-	-	-	
Stem length	0.00 (0.00)	-0.84	.41	0.01	0.00 (0.00)	-0.72	.48	0.01	-	-	-	
Count of metric system units	0.06 (0.03)	2.03	.05*	0.06	0.05 (0.03)	2.01	.05+	0.06	0.05 (0.03)	1.94	.06+	0.05
Construction	0.51 (0.14)	3.63	.00***	0.19	0.51 (0.14)	3.65	.00***	0.19	0.49 (0.14)	3.60	.00***	0.18
Presence of context	-0.21 (0.12)	-1.69	.10	0.04	-0.21 (0.12)	-1.78	.08+	0.04	-0.27 (0.09)	-2.95	.01**	0.12

Note. SE = standard error; sr^2 = semi-partial R^2. + $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

A relatively lower percentage of DIF items were flagged by all methods in this study, suggesting that eTIMSS is a well-designed instrument for measuring Grade 8 math ability across gender groups, especially since many regular items were carried over from the previous cycles. However, with the introduction of PSI items to the eTIMSS assessment for the first time, more DIF items were detected within the PSI block. Overall neither boys nor girls had an overall advantage in the eTIMSS 2019 math assessment, regardless of regular booklets or PSI booklets, which aligned with previous research about gender DIF on large-scale math assessment (Yan, 2005).

Item Features

The restricted-use items released by IEA offer insights into the design of eTIMSS items. The 2019 eTIMSS, which introduced the digital format for the first time, incorporated a variety of item features. This new trial expanded the scope of features that could be included when designing math items. For example, the digital platform of the assessment enabled students to interact with visual components and perform constructions on graphs, which is not possible with traditional paper-pencil format assessments.

The analysis of the restricted-use items' features provided valuable insights into the design of eTIMSS 2019 math assessment items. PSI blocks had longer stems with more clauses, consistent with their emphasis on real-world contexts, which increased the cognitive load required for students to read and understand the items. Compared to regular items, PSI blocks also featured more visual components and three-dimensional concepts to engage students in the context, and they even included videos in the items to help illustrate three-dimensional concepts. Item feature differences were also observed between the two regular blocks, ME02 and ME06.

For example, ME06 involved fewer equations, formulas, functions, expressions, or variables, while featuring more construction tasks and metric system units than ME02.

Relationship between Gender DIF Estimates and Item Features

When examining the relations between DIF estimates and item features, significant predictors were identified through stepwise linear regression. The presence of context was shown as a significant predictor of DIF magnitude for both the Mantel-Haenszel method and Raju's Area method, suggesting that context played an important role in reducing the magnitude of item bias across gender groups. The involvement of construction, a feature mainly introduced by the digital format, significantly increased the DIF magnitude, indicating that items requiring respondents to create or transform mathematical objects may contribute to greater item bias between boys and girls. The metric system units may not have initially attracted much attention when speaking of item features, but in the context of eTIMSS 2019 as a global assessment, adapting to units different from those used in daily life posed a challenge for students and thus led to item bias across gender groups.

Limitations

This study has several limitations regarding the methods, sample, and the relationship between DIF estimates and item features.

First, there is no perfect method for DIF detection. Each method has its pros and cons as illustrated in the previous DIF Method section. Although there was somewhat agreement among all methods, the efficacy of DIF detection methods may not be strong enough (Lopez, 2012). Limited DIF items were detected and their magnitudes were quantified in different ways. Moreover, the results from each method should be interpreted with caution, as each has specific limitations. For example, while the Logistic Regression method detected the most DIF items, it

relied on recoding all items into binary format, which may have led to a loss of information for graded items.

Second, due to security reasons, only 43 items were released and available for item feature coding, limiting the scope of item features that could be analyzed. Even though this study drew upon previous research and considered digital format features when coding item features, more features could be explored and identified in future studies. Additionally, the use of consensus coding for item features introduces subjectivity and lacks empirical evidence of interrater reliability in the coding process.

Third, when linking the DIF estimates with item features, this study examined the correlations, instead of the causal relationship. In other words, this study focused on the pattern between DIF estimates and certain item features but did not imply the presence of which items features within the math items directly cause the changes in DIF estimates.

Implications

This empirical gender DIF analysis offers important insights for assessment design, assessment use and interpretation, and further research.

First, the study identified DIF items and revealed gender DIF patterns associated with the count of clauses, the presence of context, the involvement of construction, and the number of metric system units. These findings suggest that DIF items or items including these item features may need to be revised in order to ensure both ability-matched gender groups could have an equal possibility to respond to the items correctly. This study provides valuable guidance for improving item design and promoting more equitable assessments.

Second, for practitioners who may assess students' math ability, findings from this study stress the importance of selecting proper math assessments or items that do not favor either

gender group. By understanding certain item features may be associated with DIF estimates, practitioners can also interpret assessment results more accurately to ensure that the score of math items could reflect students' latent ability rather than the influences of item design.

Third, the consensus item features that may contribute to gender DIF remain unclear, emphasizing the need for ongoing efforts to identify and mitigate gender bias at the item level. Moreover, gender DIF study results may vary because of the intersection of gender with other social factors, such as culture, country, and socioeconomic status. Therefore, future research may look into the intersecting factors to better understand the gender DIF and address item-level bias.

Acknowledgment: This work was supported by the National Science Foundation under the ECR-HER Core research program (NSF-1920512).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (1999). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Banks, K., Jeddeeni, A., & Walker, C. M. (2016). Assessing the effect of language demand in bundles of math word problems. *International Journal of Testing, 16*(4), 269–287. https://doi.org/10.1080/15305058.2015.1113972

Benjamini, Y., & Hochberg, Y. (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), 289–300. http://www.jstor.org/stable/2346101

Bielinski, J., & Davison, M. L. (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. *American Educational Research Journal, 35*(3), 455–476. https://doi.org/10.3102/00028312035003455

Buono, S., & Jang, E. E. (2021). The effect of linguistic factors on assessment of English language learners' mathematical ability: A differential item functioning analysis. *Educational Assessment, 26*(2), 125–144. https://doi.org/10.1080/10627197.2020.1858783

Doolittle, A. E. (1989). Gender differences in performance on mathematics achievement items. *Applied Measurement in Education, 2*(2), 161–177. https://doi.org/10.1207/s15324818ame0202_5

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement, 24*(2), 157–166. https://doi.org/10.1111/j.1745-3984.1987.tb00271.x

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 35–66). Lawrence Erlbaum Associates. https://doi.org/10.1075/z.62.13kok

Engelhard, G. (1990). Gender differences in performance on mathematics items: Evidence from the United States and Thailand. *Contemporary Educational Psychology, 15*(1), 13–26. https://doi.org/10.1016/0361-476X(90)90002-I

Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 user guide for the international database* (2nd ed.). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/international-database/

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morley, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology, 75*(3), 165–190. https://doi.org/10.1006/jecp.1999.2532

Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement, 40*(4), 281–306. https://doi.org/10.1111/j.1745-3984.2003.tb01148.x

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6*(2), 137–151. https://doi.org/10.1207/s15324818ame0602_3

Henderson, D. L. (2001, April). Prevalence of gender DIF in mixed format high school exit examinations. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129–145). Lawrence Erlbaum Associates. https://doi.org/10.1037/14047-004

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum Associates, Inc.

Kim, S.-H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education*, *8*(4), 291–312. https://doi.org/10.1207/s15324818ame0804_2

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education*, *22*(2), 164–184. https://doi.org/10.1080/08957340902754635

Lopez, G. E. (2012). Detection and classification of DIF types using parametric and nonparametric methods: A comparison of the IRT-likelihood ratio test, crossing-SIBTEST, and logistic regression procedures [Doctoral dissertation, University of South Florida]. *ProQuest Dissertations & Theses Global*. https://digitalcommons.usf.edu/etd/4131

Mahoney, K. S. (2003). Linguistic influences on differential item functioning for second language learners on a standardized test [Doctoral dissertation, Arizona State University]. *ProQuest Dissertations & Theses Global*. https://www.proquest.com/dissertations-theses/linguistic-influences-on-differential-item/docview/305338307/se-2?accountid=14784

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *JNCI: Journal of the National Cancer Institute*, *22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and procedures: TIMSS 2019 technical report*. Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78*(2), 333–368. ERIC.

Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education, 16*(2), rm2. https://doi.org/10.1187/cbe.16-10-0307

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*(2), 105–118. https://doi.org/10.2307/1164960

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 assessment frameworks*. Boston College, TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/timss2019/frameworks/

OECD. (2000). *Measuring student knowledge and skills: The PISA 2000 assessment of reading, mathematical and scientific literacy*. OECD Publishing. https://doi.org/10.1787/9789264181564-en

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495–502. https://doi.org/10.1007/BF02294403

Shear, B. R. (2023). Gender bias in test item formats: Evidence from PISA 2009, 2012, and 2015 math and reading tests. *Journal of Educational Measurement, 60*(4), 676–696. https://doi.org/10.1111/jedm.12372

Shanmugam, S. K. S. (2020). Gender-related differential item functioning of mathematics computation items among non-native speakers of English. *The Mathematics Enthusiast, 17*(1), 108–140. https://doi.org/10.54870/1551-3440.1482

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25*(3), 246–280. https://doi.org/10.1080/08957347.2012.687650

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 147–172). Lawrence Erlbaum Associates. https://doi.org/10.1037/14047-004

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*(6), 479–498. https://doi.org/10.1177/0146621603259902

Yan, S. (2005). Gender-related differential item functioning in mathematics assessment on the Third International Mathematics and Science Study -Repeat (TIMSS-R) [Doctoral dissertation, The University of Toledo]. *ProQuest Dissertations & Theses Global*.

https://www.proquest.com/dissertations-theses/gender-related-differential-item-functioning/docview/305389471/se-2

Zieky, M. (2003). *A DIF primer*. Educational Testing Service. https://www.ets.org/s/praxis/pdf/dif_primer.pdf

Appendix A

Codebook Codes

Category	Item Feature	Definition	Detailed Value Scheme
Stem	Stem length	The total number of words contained in the stem of an item, where a "word" is defined as any sequence of characters separated by spaces, excluding punctuation that is directly attached to words	Numeric
Stem	Count of conditional clauses	The count of condition statements within the stem has conditional markers such as "if" or "then"	Numeric
Stem	Presence of context	Information is presented in narrative structures such as a scenario or a story	0: No; 1: Yes
Stem	Count of variables in stem	The number of variables (symbols representing unknown values or quantities) in the stem, excluding formulas or equations	Numeric
Stem	Count of passive sentences	The number of sentences written in passive voice within the stem	0: No passive voice; 1: One passive voice; 2: Double passive voice
Stem	Count of conditional clauses	The number of conditional statements in the stem that have conditional markers such as "if" or "then", or similar indicators	Numeric
Stem	Count of relative clauses	The number of embedded clauses in the stem that use relative pronouns (e.g., "that", "which", "who", "where") to describe a noun or object	Numeric
Stem	Presence of complex question clauses	Whether the item contains a main clause with one or more subordinating clauses. Complex question phrases often appear in word problems when background information is provided. These clauses can often be simplified by using single-clause sentences with a question word, such as "what", "which", or "when", at the beginning	0: No; 1: Yes

Option	Option length	The total number of words for the total item options (including any instructions for how to operate the player to respond)	Numeric
Option	Count of equation/formula/function/expression in option	The count of equation/formula/function/expression in all options of each item	Numeric
Option	Option length	The total word count of all item options, including any instructional text needed to guide the student's response	Numeric
Item	Involvement of construction	The question required the transformation, construction, or creation of a mathematical object	0: No; 1: Yes
Item	Count of metric system unit	The total number of measurement units used in the item that is part of the metric system (e.g., metric units such as meters, liters, or grams)	Numeric
Item	Item type	The format of the item	1: MC; 2: CR
Item	Content domain	The broader subject area being assessed by the item	1: Number; 2: Algebra; 3: Geometry; 4: Data and Probability
Item	Topic area	The subdomain within the content domain that focuses on a specific set of knowledge or skills	Categorical
Item	Topic	The specific subject or content area within the topic area that the item assesses	Categorical
Item	Cognitive domain	The cognitive process targeted by the item (e.g., recall, apply, analyze) as students engage with the content	1: Knowing; 2: Applying; 3: Reasoning
Item	Presence of stem first	Indicates whether the stem appears before any visual elements in the item presentation	0: Stem First; 1: Visual First

Item	Presence of three-dimensional features	Indicates whether the item involves three-dimensional objects or requires 3D spatial reasoning to solve the problem	0: No; 1: Yes
Item	Presence of everyday context	Indicates whether the item uses a context familiar to students from daily life experiences	0: No; 1: Yes
Item	Involvement of unit conversion	Indicates whether the item requires converting between different units of measurement to arrive at a correct solution	0: No; 1: Yes
Item	Presence of cultural context	Indicates whether the item includes contexts or visuals related to a specific cultural background, which may be more familiar to certain student groups and less relevant to others	0: No; 1: Yes
Visual	Presence of visual components	The presence of visual components used in the item	0: No; 1: Yes
Visual	Interactive visual representation	the existence of visual elements components that can be drawn, erased, reset, etc.	0: No; 1: Yes
Visual	Visual functionality	Describes the function of visuals in the stem, such as illustrating concepts, providing context, or supporting problem-solving	1: Display quantitative data or trends (e.g., graphs); 2: Organize and display data in rows and columns (e.g., tables); 3: Illustrate structures, relationships, or processes (e.g., diagrams); 4: Add visual interest without conveying quantitative or structural information (e.g., illustrations); 5: Combine multiple visual aids serving different functions (e.g., graphs and diagrams together)
Visual	Visual relevance	Indicates whether the visuals (e.g., illustrations, diagrams, tables, graphs) effectively support and clarify key information in the item	0: No; 1: Yes

Appendix B

Sample *R* Code

```
### DIF Analysis for PSI Booklets----
#### MH----
PSI_fitMH_gender <- difMH(PSI_data_ordered[, !names(PSI_data_ordered) %in%
      c("genderMF")], PSI_data_ordered$genderMF, focal.name = "F",
      p.adjust.method = "BH")
PSI_fitMH_gender$DIFitems

PSI_fitMH_gender_result <- data.frame(Item = PSI_fitMH_gender$names, DIF_est
      = -2.35 * log(PSI_fitMH_gender$alphaMH), Adj_p =
      PSI_fitMH_gender$adjusted.p)
print(PSI_fitMH_gender_result)



#### LR----
PSI_fitLR_gender <-
      difLogistic(PSI_data_ordered_recoded[, !names(PSI_data_ordered_recoded)
      %in% c("genderMF")], PSI_data_ordered_recoded$genderMF, focal.name =
      "F", type = "both", p.adjust.method = "BH")
PSI_fitLR_gender$DIFitems


PSI_fitLR_gender$logitPar
PSI_fitLR_gender$logitSe


PSI_fitLR_gender_result <- data.frame(Item = PSI_fitLR_gender$names, DIF_est
      = PSI_fitLR_gender$deltaR2, Adj_p = PSI_fitLR_gender$adjusted.p)
print(PSI_fitLR_gender_result)



#### LRT----
PSI_md.cons0 <- multipleGroup(PSI_data_ordered[, 1:(ncol(PSI_data_ordered)-
      1)], 1, group=PSI_data_ordered$genderMF, itemtype = c(rep('3PL',
      length(PSI_threePL_items)), rep('2PL', length(PSI_twoPL_items)),
      rep('gpcm', length(PSI_GPCM_items))),
                              SE=TRUE,invariance=c('free_means',
      'free_var',colnames(PSI_data_ordered)))
coef(PSI_md.cons0,simplify=T)


PSI_d=DIF(PSI_md.cons0, which.par = c('a1','d1','d2','d'), p.adjust =
      'fdr',scheme = 'drop')
PSI_d
PSI_ratio=PSI_d$X2/PSI_d$df

# k is the number of anchors you want to use
k=5

# select anchors with smallest ratios
PSI_anchors=which(PSI_ratio %in% sort(PSI_d$X2/PSI_d$df)[1:k])
```

```
PSI_anchors

# run LRT with selected anchors
PSI_md.noncons0 <- multipleGroup(PSI_data_ordered[,
      1:(ncol(PSI_data_ordered)-1)], 1, group=PSI_data_ordered$genderMF,
      itemtype = c(rep('3PL', length(PSI_threePL_items)), rep('2PL',
      length(PSI_twoPL_items)), rep('gpcm', length(PSI_GPCM_items))),
                                SE=TRUE,invariance=c('free_means',
      'free_var',colnames(PSI_data_ordered)[PSI_anchors]))
coef(PSI_md.noncons0, IRTpars=T, simplify=T)

# J is the number of items
PSI_J <- ncol(PSI_data_ordered[, c(1:25)])
PSI_J

# free-baseline designated anchors
PSI_dif=DIF(PSI_md.noncons0, which.par = c('a1','d1','d2','d'), p.adjust =
      'fdr',scheme = 'add',items2test=c(1:PSI_J)[-PSI_anchors])
which(PSI_dif$adj_p<0.05)
PSI_dif


#### Raju's area----
PSI_coef_M <- data.frame(coef(PSI_md.noncons0, simplify = TRUE, IRTpars =
      TRUE)$M$item)
gender = rep("M", nrow(PSI_coef_M))
item = colnames(PSI_data_ordered)[c(1:PSI_J)]
PSI_coef_M <- cbind(item, PSI_coef_M, gender)
PSI_coef_M


PSI_coef_F <- data.frame(coef(PSI_md.noncons0, simplify = TRUE, IRTpars =
      TRUE)$F$item)
gender = rep("F", nrow(PSI_coef_F))
item = colnames(PSI_data_ordered)[c(1:PSI_J)]
PSI_coef_F <- cbind(item, PSI_coef_F, gender)
PSI_coef_F

##### 3PL----
#(numerical integration)

icc_3pl <- function(theta, a, b, c) {
  c + (1 - c) / (1 + exp(-a * (theta - b)))
}


calculate_area_numerical <- function(a_female, b_female, c_female, a_male,
      b_male, c_male) {
  area_between_3PL_iccs <- function(theta) {
    abs(icc_3pl(theta, a_female, b_female, c_female) - icc_3pl(theta, a_male,
      b_male, c_male))
  }
  area <- integrate(area_between_3PL_iccs, lower = -6, upper = 6)$value
  return(area)
}
```

```
PSI_3PL_area <- data.frame(item = character(), area = numeric(),
      stringsAsFactors = FALSE)

for (item in PSI_threePL_items) {

  a_male <- PSI_coef_M[PSI_coef_M$item == item, "a"]
  b_male <- PSI_coef_M[PSI_coef_M$item == item, "b"]
  g_male <- PSI_coef_M[PSI_coef_M$item == item, "g"]

  a_female <- PSI_coef_F[PSI_coef_F$item == item, "a"]
  b_female <- PSI_coef_F[PSI_coef_F$item == item, "b"]
  g_female <- PSI_coef_F[PSI_coef_F$item == item, "g"]


  area <- calculate_area_numerical(a_female, b_female, g_female, a_male,
      b_male, g_male)


  PSI_3PL_area <- rbind(PSI_3PL_area, data.frame(item = item, area = area,
      stringsAsFactors = FALSE))
}


print(PSI_3PL_area)


##### 2PL----

icc_2pl <- function(theta, a, b) {
  1 / (1 + exp(-a * (theta - b)))
}


calculate_area_numerical_2pl <- function(a_female, b_female, a_male, b_male)
      {
  area_between_2PL_iccs <- function(theta) {
    abs(icc_2pl(theta, a_female, b_female) - icc_2pl(theta, a_male, b_male))
  }
  area <- integrate(area_between_2PL_iccs, lower = -6, upper = 6)$value
  return(area)
}


PSI_2PL_area <- data.frame(item = character(), area = numeric(),
      stringsAsFactors = FALSE)


for (item in PSI_twoPL_items) {

  a_male <- PSI_coef_M[PSI_coef_M$item == item, "a"]
  b_male <- PSI_coef_M[PSI_coef_M$item == item, "b"]

  a_female <- PSI_coef_F[PSI_coef_F$item == item, "a"]
  b_female <- PSI_coef_F[PSI_coef_F$item == item, "b"]


  area <- calculate_area_numerical_2pl(a_female, b_female, a_male, b_male)
```

```
  PSI_2PL_area <- rbind(PSI_2PL_area, data.frame(item = item, area = area,
      stringsAsFactors = FALSE))
}


print(PSI_2PL_area)



##### GPCM----
icc_gpcm <- function(theta, a, b) {
  # Calculate the exponent part for each category
  P <- c(1, sapply(1:length(b), function(k) exp(sum(a * (theta - b[1:k])))))
  prob <- P / sum(P)
  return(prob)
}


calculate_area_numerical_gpcm <- function(a_female, b_female, a_male, b_male)
      {
  area_between_GPCM_iccs <- function(theta) {
    icc_female <- icc_gpcm(theta, a_female, b_female)
    icc_male <- icc_gpcm(theta, a_male, b_male)
    diff <- sum(abs(icc_female - icc_male))
    return(diff)
  }

  vectorized_area_between_GPCM_iccs <- Vectorize(area_between_GPCM_iccs)
  area <- integrate(vectorized_area_between_GPCM_iccs, lower = -6, upper =
      6)$value
  return(area)
}


PSI_GPCM_area <- data.frame(item = character(), area = numeric(),
      stringsAsFactors = FALSE)


for (item in PSI_GPCM_items) {
  a_male <- PSI_coef_M[PSI_coef_M$item == item, "a"]
  b_male <- as.numeric(PSI_coef_M[PSI_coef_M$item == item, c("b1", "b2")])

  a_female <- PSI_coef_F[PSI_coef_F$item == item, "a"]
  b_female <- as.numeric(PSI_coef_F[PSI_coef_F$item == item, c("b1", "b2")])


  area <- calculate_area_numerical_gpcm(a_female, b_female, a_male, b_male)


  PSI_GPCM_area <- rbind(PSI_GPCM_area, data.frame(item = item, area = area,
      stringsAsFactors = FALSE))
}


print(PSI_GPCM_area)
```

```
print(rbind(PSI_3PL_area, PSI_2PL_area, PSI_GPCM_area))



#### Stepwise Linear Regression (MH~Item Features)----
##### Model1: 8 IV----
model_MH_1 <- lm(MH_DIF_est ~ Stem_Length + Equation_Variable + Clause +
      Presence_of_Visuals + Three_Dimensional + Construction +
      Presence_of_Context + Count_of_Non_Us_Metrics, data =
      DIF_itemfeature_data)
summary(model_MH_1)

stepwise_model_MH_1 <- step(model_MH_1, direction = "both")
summary(stepwise_model_MH_1)

###### backward stepwise (MH)----
# Model 1 (initial model)
model_1 <- lm(MH_DIF_est ~ Stem_Length + Equation_Variable + Clause +
                Presence_of_Visuals + Three_Dimensional + Construction +
                Presence_of_Context + Count_of_Non_Us_Metrics, data =
      DIF_itemfeature_data)
summary(model_1)

# Model 2 (after removing Stem_Length)
model_2 <- lm(MH_DIF_est ~ Equation_Variable + Clause + Presence_of_Visuals +
                Three_Dimensional + Construction + Presence_of_Context +
                Count_of_Non_Us_Metrics, data = DIF_itemfeature_data)
summary(model_2)

# Model 3 (after removing Count_of_Non_Us_Metrics)
model_3 <- lm(MH_DIF_est ~ Equation_Variable + Clause + Presence_of_Visuals +
                Three_Dimensional + Construction + Presence_of_Context, data
      = DIF_itemfeature_data)
summary(model_3)

# Model 4 (after removing Construction)
model_4 <- lm(MH_DIF_est ~ Equation_Variable + Clause + Presence_of_Visuals +
                Three_Dimensional + Presence_of_Context, data =
      DIF_itemfeature_data)
summary(model_4)

# Model 5 (after removing Three_Dimensional)
model_5 <- lm(MH_DIF_est ~ Equation_Variable + Clause + Presence_of_Visuals +
                Presence_of_Context, data = DIF_itemfeature_data)
summary(model_5)

# Model 6 (Final model)
final_model <- lm(MH_DIF_est ~ Equation_Variable + Clause +
      Presence_of_Context, data = DIF_itemfeature_data)
summary(final_model)
```