Diksi (2025), Vol. 33, No. 2, **225-248**
doi.org/10.21831/diksi.v32i2

Is there any item or test bias in the Business English Test at Universitas Terbuka?

Agus Santoso[1]*, Heri Retnawati[2], Timbul Pardede[1], Dyah Paminta Rahayu[1], Munaya Nikma Rosyada[2], Rugaya Tuanaya[2], Rimajon Sotlikova[3], Begimbetova Guldana Atymtaevna[4]

AFFILIATION
[1]Universitas Terbuka, Indonesia
[2]Universitas Negeri Yogyakarta, Indonesia
[3]Webster University in Tashkent, Uzbekistan
[4]Abai Kazakh National Pedagogical University, Kazakhstan
*Corresponding Author:
aguss@ecampus.ut.ac.id

ARTICLE HISTORY
- Received 19 April 2025
- Accepted 17 September 2025
- Published 30 September 2025

CITATION (APA STYLE)
Santoso, A., Retnawati, H., Pardede, T., Rahayu, D. P., Rosyada, M. N., Tuanaya, R., ... Atymtaevna, B. G. (2025). Is there any item or test bias in the Business English Test at Universitas Terbuka?. *Diksi, 33*(2). https://doi.org/10.21831/diksi.v33i2.84570

Abstract: In an exemplary test implementation, the items used should be fair and free from bias. This biased content threatens the validity that will affect the interpretation of the test results. This study aims to analyze the biased content of items and tests on the test device for the Commercial English Course, with the code ADBI4201, a course in the Business Administration Study Program at the Faculty of Law, Social, and Political Sciences, Open University (UT). It uses the quantitative approach. Data were collected through documentation in the form of grids and responses from final semester test participants in January 2024. Data analysis was carried out by (1) estimating item parameters using classical test theory and item response theory for tests, (2) estimating item parameters based on regional groups and based on gender, which are used to identify DIF content, (3) testing the significance of DIF with a maximum likelihood comparison. The study's findings showed 26 items contained DIF by gender and 25 items contained DIF by region. The results of the test bias detection showed that the test slightly favored the female group and the students from the Java region.

Keywords: *Area and gender, differential item functioning, differential test functioning, English test.*

INTRODUCTION

A test is one of the main tools used in education to assess students' abilities and knowledge (Argianti & Retnawati, 2020). The purpose of conducting tests is to obtain valid and reliable information about the things to be measured (Bilyakovska, 2022; Eliaumra et al., 2022; Fulcher, 2016). The administration must be fair and impartial so that the results can be used effectively and meaningfully (Wallace, 2018). This means that every student must have an equal opportunity to demonstrate their abilities without any discrimination or bias that may arise from background, gender, race, or other factors that are not relevant to the material being tested. A fair test must also be administered under the same conditions for all participants (Makkink & Vincent-Lambert, 2020; Rasooli et al., 2019). This includes a conducive testing environment, clear and consistent instructions, and equal treatment during the testing process. In addition, the construction of questions must be done carefully to ensure that each item in the test does not contain language or content that could be misinterpreted or more difficult for certain groups to understand (Effiom, 2021).

Therefore, a good test should meet several substantial criteria, including validity, reliability, and good item characteristics (Nurrahman et al., 2022; Otaya et al., 2020; Wilsa et al., 2023). Validity measures the extent to which a test measures what it is supposed to measure (Retnawati, 2016). In developing instruments, two types of validity are known, namely, content validity, which aims to measure whether the test covers all aspects of the material that should be tested; and construct validity, which aims to measure whether the test measures the intended concept or construct (Bademci, 2022; Bilyakovska, 2022; Setiawan et al., 2023). Reliability shows the consistency of test results when the test is repeated under the same conditions. A reliable test will provide similar results when administered to the same group at different times (Babu & Kohli, 2023). Some ways to measure reliability are test-retest reliability which measures the consistency of results over time, internal consistency reliability which measures the extent to which items in a test give similar results, and inter-rater reliability which measures the consistency of results when scored by different raters (Babu & Kohli, 2023; Leventhal & Gregg, 2022). Meanwhile, item characteristics refer to the individual quality of each item in the test. Test items must be clear, unambiguous, and appropriate to the desired difficulty level. In addition, items must distinguish well between participants who understand the material and those who do not (H. H. Dewi et al., 2023).

Differential Item Functioning (DIF) or item bias is a condition in which an item functions differently for different groups, even though the measured abilities are the same (Sumin et al., 2022). The presence of DIF in a test can threaten the validity of the test because the test results no longer reflect the factual abilities of the participants (Sumin et al., 2022). DIF can arise for various reasons, including cultural, linguistic, or experiential differences between the tested groups (Wallin et al., 2024). For example, an item in an English test may be easier to understand for native English speakers than for those whose first language is not English, even though both groups have the same level of language ability (Bormanaki & Ajideh, 2022). Therefore, it is necessary to detect item bias in each question. Detecting bias in test items is critical in ensuring the test is fair and valid (Effiom, 2021). If bias is not detected, test results can provide an unfair advantage to some groups of participants while disadvantaging others. This impacts individual fairness and affects educational policies and decision-making based on test results (Canay et al., 2022). The bias detection process involves statistical analysis to identify items that may function differently for different groups.

One of the methods that is often used is DIF analysis, which can help reveal unfair items (Dubbelman et al., 2020). By detecting and correcting bias, test designers can improve the validity and reliability of tests and ensure that all participants have an equal opportunity to demonstrate their abilities (Penfield & Camilli, 2006). Based on this description, this study aims to analyze the biased content of items and tests in the Commercial English

language test used in the undergraduate program at Universitas Terbuka. This analysis is necessary, considering that Universitas Terbuka has diverse student backgrounds that require fair and impartial evaluation. The focus of this study is to identify whether there are items or parts of the Commercial English language test that show DIF and to understand the implications of these findings for the validity and reliability of the test. Universitas Terbuka, as a higher education institution that serves various levels of society with different backgrounds, must ensure that the tests they use reflect the abilities and knowledge of students objectively. Thus, investigating test bias is a necessary step in achieving this goal.

DIF is a substantial concept in educational evaluation and measurement. DIF occurs when items on a test show functional differences between different groups, even though both groups have the same ability level on the construct being measured by the test (Wallin et al., 2023). More technically, DIF can occur when the probability of a correct answer on a particular item differs for two groups of test takers with equal ability. DIF can be classified into two types: uniform DIF and non-uniform DIF. Uniform DIF occurs when the functional differences of an item are consistent across ability levels. In other words, one group always has a higher or lower probability of answering the item correctly than the other group, regardless of their ability level. In contrast, non-uniform DIF occurs when the functional differences of an item vary across ability levels. In this case, the biasing effects of the item can change depending on the ability level of the participants. A thorough understanding of DIF is essential because the presence of DIF can indicate that an item on a test may contain elements that are unfair to one group of participants. Identifying and managing DIF helps ensure that tests remain valid and reliable in measuring learner ability.

Various methods are used to detect DIF in test items, and one of them is using item characteristic curves (ICC) (Lord, 1980; Setiawati et al., 2017). The item characteristic curve is a graphical representation that shows the relationship between participant ability and the probability of answering an item correctly (Baker & Kim, 2017). In DIF analysis using ICC, a comparison is made between the item characteristic curves for two or more groups. If the curves for these groups differ significantly, this may indicate DIF (Andrich & Marais, 2019). For example, if, at one ability level, one group has a higher probability of answering an item correctly than another group, then the item shows DIF. To assess whether the difference in the item characteristic curves indicates significant DIF, the maximum likelihood method can be used. This method involves calculating the item parameters that most likely give the observed data and comparing models with and without DIF to determine the significance of the differences (Ito et al., 2019; Patnala et al., 2024; Szmańda & Witkowski, 2021).

In addition to analyzing DIF at the item level, it is also necessary to understand DIF at the overall test level, known as Differential Test Functioning

(DTF) (Walker & Gocer Sahin, 2023; Yavuz Temel, 2023). DTF occurs when the entire test exhibits bias toward a particular group, even though each item may not exhibit significant DIF. DTF can occur due to the cumulative effects of DIF from multiple items or due to complex interactions between items on the test. For example, if several items on a test are consistently easier for one group than another, this can lead to DTF, resulting in unfair test results for one group. DTF analysis involves measuring the overall differences in performance between the groups being tested and using statistical methods to determine whether those differences are significant and caused by bias in the test. By understanding and managing DTF, test designers can ensure that the test set is fair and valid for all test takers.

Several studies show various applications of DIF and DTF analysis in many fields, including education, psychology, health research, and psychometrics. Advanced statistical techniques and innovative methodologies improve assessments' validity, fairness, and reliability across different populations and contexts. One study (Shykhnenko, 2020) explored the optimization of an assessment system in an English for Specific Purposes course using DIF and DTF. This study highlights the practical application of DIF analysis in educational settings to improve assessment outcomes and ensure fairness in evaluation. Psychology research discusses a Bayesian approach to detecting DIF using the Generalized Graded Unfolding Model. This study emphasizes the importance of DIF analysis in psychological assessment, demonstrating the relevance of advanced statistical methods in identifying item bias and ensuring the accuracy of psychological measurements (Joo et al., 2022). Another study contributes to the refinement of DIF analysis techniques, especially in cases where group characteristics are not explicitly specified (Wallin et al., 2024). In addition, another study conducted a cross-country comparison of trends in adolescent psychosomatic symptoms using Rasch analysis and identified DIF on items related to depressive mood across periods (Hagquist et al., 2019). This study demonstrates the application of DIF analysis in health research to understand the variation in symptom reporting among adolescents from different Nordic countries.

Nedungadi et al. (2022) investigated DIF in the Fundamental Concepts for Organic Reaction Mechanisms Inventory to assess whether students from different gender groups and majors scored differently on certain items despite having the same proficiency. This study emphasizes the importance of DIF analysis in validating assessment tools and ensuring unbiased measurement of knowledge and skills. Terluin et al. (2018) explored the equivalence of web-based and paper-based questionnaires using DIF and DTF analysis. By comparing different modes of questionnaire administration, this study demonstrated the usefulness of DIF analysis in evaluating measurement consistency across assessment formats. Li & Becker (2021) introduced the concept of Differential Bundle Functioning to quantify the amount of accumulated DIF within an item group, providing a comprehen-

sive approach to assessing measurement bias within a given item group. This study contributes to the advancement of DIF analysis techniques by focusing on the collective impact of item bias on assessment outcomes. Moradi et al. (2022) focused on the fairness of reading comprehension tests across gender and learning modes, emphasizing the importance of investigating DIF and unidimensionality to ensure the validity of test results. By examining potential biases in reading comprehension assessments, this study highlights the role of DIF analysis in promoting fair and accurate evaluations. In a study by Hagquist & Andrich (2017), recent advances in DIF analysis in health research using the Rasch model are discussed. This study emphasizes the increasing use of Rasch analysis with a focus on DIF to evaluate the psychometric characteristics of health outcome measures, emphasizing the significance of DIF analysis in health care settings.

METHOD

This study uses a quantitative approach, and the data are the scores of the Commercial English Course test, with the code ADBI4201, a course in the Department of Business Administration, Faculty of Law and Social and Political Sciences, Universitas Terbuka (UT). The data used are in the form of documentation of the grid and responses of the final semester test participants in January 2024. There were 6619 test participants, providing student answers to the test in the course. The variables used in this study include region and gender. The region variable identifies the location of the test participants' schools in the research sample, namely the Java and Outside Java regions. The gender variable identifies the gender of the test participants in this research sample, namely male and female. The distribution of test participants by both gender and region of origin can be seen in Table 1.

Table 1. Test takers data

Origin	Gender		Total
	1 (Male)	2 (Female)	
1 (Java)	1,189	2,664	3,853
2 (out of Java)	979	1,787	2,766
Total	2,168	4,451	6,619

The data analysis process was carried out in several steps; first, the most appropriate model for the available data was selected using model fit analysis. After the appropriate model was constructed, the next step was to conduct an assumption test for each question item. Then, the item characteristics of the test device used were estimated. Furthermore, the question items were evaluated based on the gender group and student region. DIF was identified based on the Item Characteristic Curve for gender and region.

After that, the significance of the DIF load was tested using the maximum likelihood ratio. Finally, DIF interpretation was conducted by looking at each test item in detail. The data analysis was carried out using the open-source software R Studio.

RESULTS AND DISCUSSION

Result

Model Fitness

Table 2. Model fit test results

Set 85	
Model	Fit test result
Rasch	2
1-PL	2
2-PL	17
3-PL	22
4-PL	23

Table 2 shows the results of the model fit test conducted on the test items of Set 85. In Set 85, the Rasch model showed a fit test result of 2, which indicates that there are only two items that fit the model. Model 1-PL also showed a fit test result of 2. In contrast, model 2-PL showed a fit test result of 17, which means that only 17 items fit this model. Model 3-PL showed a fit test result of 22, indicating a higher fit compared to model 2-PL. Finally, model 4-PL showed a fit test result of 23, the highest fit among all models tested in Set 83. From the results of the model fit test in Table 2, it can be concluded that 4-PL showed the best fit with the data in the items of Set 85.

Unidimensional assumption test

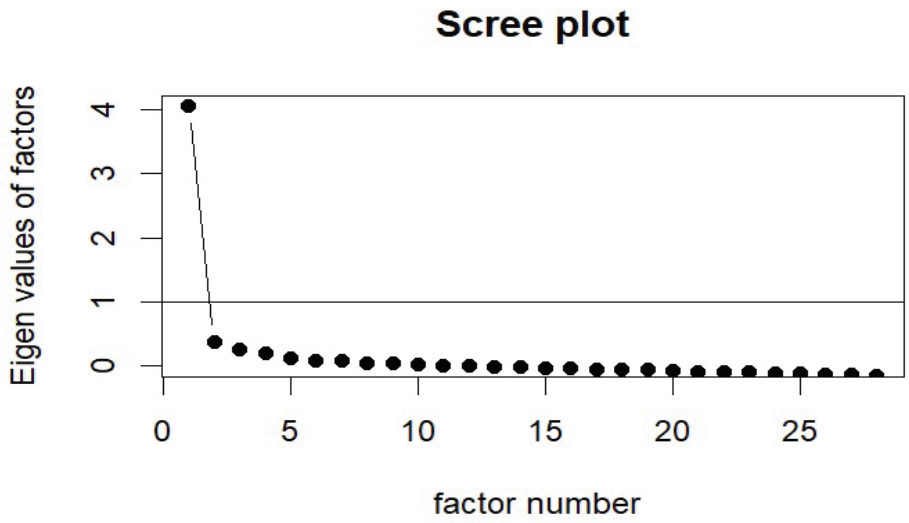

Figure 1. Scree Plot of Unidimensional Test

Figure 1 shows a scree plot of the eigenvalues of various factors. This scatter plot shows the number of significant factors in the data. The scree plot shows that the largest eigenvalue is in the first factor, with a value approaching 4. This eigenvalue decreases sharply in the second and subsequent factors and begins to level off after the second or third factor. This indicates that most of the variance in the data is explained by the first factor. Therefore, it can be concluded that the data have a strong unidimensional structure, with one dominant factor explaining most of the variance in the data. Figure 1 shows a strong unidimensional structure; the unidimensional assumption is met.

Parent data parameters

Table 3 shows that most of the questions have a difficulty index in the "Medium" category, which means that the items have a balanced level of difficulty, not too difficult or too easy for test takers. In other words, most items in this test can be answered by respondents with a reasonable level of difficulty. Some items were categorized as "Easy," such as B6, B8, B13, B15, B20, and B23. These items were easier for respondents to answer, which may indicate that the content or questions were simpler or more familiar to the respondents. However, items B11 and B17 were considered "Difficult," meaning that these items were more challenging and required a higher level of knowledge or skill for test takers to answer correctly.

Table 3. Test Item Characteristics

Item	Difficulty Index	Category	Discrimination Power	Category
B1	0.46	Medium	0.3118	Good
B2	0.67	Medium	0.4187	Excellent
B4	0.70	Medium	0.1535	Fair
B5	0.59	Medium	0.3781	Good
B6	0.54	Medium	0.3853	Good
B7	0.76	Easy	0.4124	Excellent
B8	0.51	Medium	0.4297	Excellent
B9	0.75	Easy	0.3068	Good
B11	0.56	Medium	0.0363	Poor
B12	0.40	Medium	0.0278	Poor
B13	0.25	Difficult	0.1085	Fair
B14	0.69	Medium	0.3576	Good
B15	0.72	Easy	0.3351	Good
B16	0.60	Medium	0.0852	Poor
B17	0.95	Easy	0.2989	Fair
B18	0.55	Medium	0.3992	Good
B19	0.21	Difficult	0.0095	Poor
B20	0.69	Medium	0.4777	Excellent
B21	0.39	Medium	0.3611	Good
B22	0.72	Easy	0.4415	Excellent
B23	0.58	Medium	0.3480	Good
B24	0.50	Medium	0.3432	Good
B25	0.79	Easy	0.4475	Excellent
B26	0.63	Medium	0.2444	Fair
B27	0.41	Medium	0.3958	Good
B28	0.58	Medium	0.3856	Good
B29	0.69	Medium	0.3197	Good
B30	0.60	Medium	0.4288	Excellent

Table 4. Description and Internal Consistency of Test Scale

N of Item	28.000000
N of Person	6308.000000
Alpha	0.791054
Scale Mean	16.504756
Scale SD	5.056473

For the discrimination index, several items showed "Excellent" discrimination, such as B2, B6, B7, B18, B20, B23, and B28. This means that these items are very effective in distinguishing between respondents with high and low abilities. These items can well separate respondents who have a strong understanding or skills from those who do not. Many items are categorized as "Good", such as B1, B4, B5, B8, B12, B13, B16, B19, B21, B22, B24, B25, B26, and B27. These items are also effective in distinguishing between respondents with different abilities, although not as strong as those in the "Excellent" category. Several items fall into the "Fair" category, such as B3, B11, B15, and B24. These items have quite good discrimination power but not as strong as the "Good" or "Excellent" category. Items with "Poor" discrimination power include B9, B10, B14, and B17. These items are less effective in differentiating between high and low-ability respondents. These items may not be very useful in determining different levels of understanding or skills among respondents. In addition, Table 4 shows the results of the Alpha Coefficient of 0.791054, which means that the data have a good level of internal consistency, indicating that the test items consistently measure the same construct. The analysis results in Table 3 indicate that the test items generally have appropriate levels of difficulty and discrimination power, with most items performing well in both aspects. However, some items may need to be reviewed to ensure that the level of difficulty and discrimination are appropriate to the desired test characteristics.

DIF by Gender

The data presented in Table 5 are DIF data based on gender for all items in the test. The data include three main parameters: discrimination parameter (a1), difficulty parameter (d), and guessing parameter (g), which are analyzed for both male and female groups. A high a1 value indicates the ability of the item to differentiate participants with different abilities. For example, Items b12 and b13 show very high discrimination values for males (a1 = 3.028 and 3.016), while Item b15 shows high discrimination values for females (a1 = 1.789). Furthermore, a higher d-value indicates a more difficult item. Some items show significant differences in difficulty values between males and females. For example, Item b17 has a higher difficulty value for females (d = 4.033) compared to males (d = 3.254). Items b2 and b29 also show significant differences in difficulty values between the two gender groups. For the parameter g, it is highly expected to show a low value because it shows that participants are more likely to answer correctly based on their abilities rather than guessing. In general, the g-value varies, but some items, such as Items b12 and b13, show higher g-values, indicating a greater possibility of students guessing the answer. Table 5 shows a variation in the three parameters between males and females.

Table 5. Item parameter for participants' gender

Item	Males			Females		
	a1	d	g	a1	d	g
b1	1.037	-0.653	0.036	0.739	-0.107	0.019
b2	1.341	0.549	0.047	1.261	1.019	0.013
b4	0.464	0.567	0.064	0.314	0.777	0.117
b5	1.864	-0.340	0.204	1.035	0.358	0.083
b6	1.155	-0.123	0.005	1.020	0.288	0.009
b7	1.885	0.913	0.195	1.611	1.218	0.275
b8	1.169	-0.415	0.024	1.403	0.171	0.046
b9	1.049	0.625	0.207	0.902	1.166	0.170
b11	0.093	0.136	0.024	0.053	0.248	0.023
b12	3.028	-5.670	0.416	2.271	-5.499	0.361
b13	3.016	-5.162	0.197	3.216	-6.577	0.220
b14	1.960	0.051	0.343	1.319	0.403	0.268
b15	1.479	0.013	0.357	1.789	-0.004	0.483
b16	0.113	0.205	0.032	0.200	0.453	0.019
b17	2.537	3.254	0.453	1.985	4.033	0.433
b18	1.873	-0.405	0.153	1.072	0.267	0.024
b19	-0.026	-2.081	0.071	-0.032	-1.572	0.065
b20	2.965	-0.321	0.273	2.346	1.085	0.244
b21	2.172	-1.923	0.160	2.019	-1.497	0.167
b22	1.723	0.868	0.083	1.400	1.415	0.009
b23	1.812	-0.612	0.263	1.488	-0.373	0.268
b24	1.300	-0.727	0.142	1.464	-0.752	0.242
b25	1.833	1.421	0.035	1.545	2.161	0.009
b26	0.763	0.267	0.017	0.532	0.683	0.009
b27	2.001	-1.787	0.129	1.164	-0.343	0.021
b28	1.873	-0.655	0.216	1.109	0.333	0.092
b29	2.301	-0.965	0.473	1.685	-0.076	0.422
b30	2.200	-0.549	0.194	1.267	0.571	0.056

Figure 2 shows item probability functions for 28 test items based on two gender categories (cat1: Female and cat1: Male). The horizontal axis (θ) represents the ability level of the participants, while the vertical axis (P(θ)) shows the probability of answering each item correctly. The blue (cat1: Female) and yellow (cat1: Male) curves in each subplot provide important

information about the performance differences between the two gender groups. The overlapping curves indicate that the probability of answering correctly for both groups is almost the same across ability levels, which means that the items function fairly for both genders. However, some items show a striking difference between the two curves, indicating Differential Item Functioning (DIF). For example, in Item b17, a significant difference is seen at the low ability level, indicating that the male group has an advantage in answering the item compared to the female group. Similarly, Items b28, b26, b22, and b7 show differences between the groups, which could lead to item bias toward one group. In contrast, Items b11, b13, b14, and b25 show almost complete overlapping curves between the two groups. This indicates that the probability of answering correctly is almost the same for both categories at different ability levels, so these items do not show any significant bias.

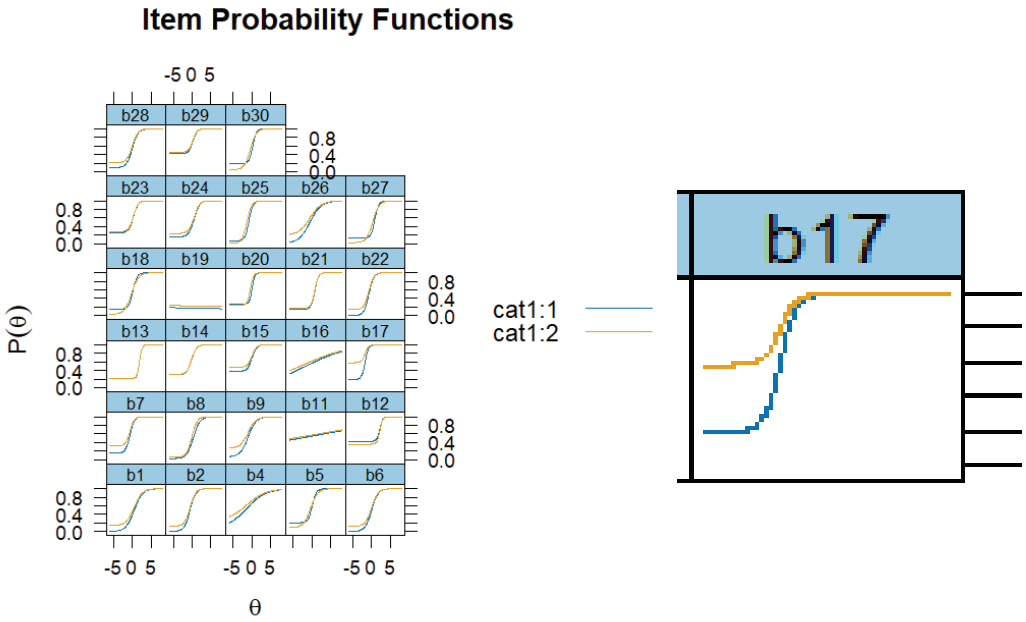

Figure 2. ICC DIF based on test takers' gender (The 17th item in the right side)

It can be seen in Table 6 that item b1 shows an AIC value of -53.868, SABIC of -43.153, HQ of -46.854, BIC of -33.619, chi-square value (X2) of 59.868 with 3 degrees of freedom, and a p-value of 0, indicating that this item is significant and has a DIF load. Item b11, on the other hand, shows a p-value of 0.198, which means it is not significant in the context of DIF analysis. Table 5 also shows that out of a total of 28 items, most items show significant DIF indications except for Items b11 and b23, which do not show significance based on a p-value greater than 0.05. Items that are indicated to contain DIF mean that there are differences in the way the item is respond-

ed to by different groups, even though they have the same level of ability. Lower AIC, SABIC, HQ, and BIC values indicate a better model. Table 5 shows that most items have negative values, indicating that the model used is quite good at explaining the data.

Table 6. Results of the DIF artificial significance test

Item	AIC	SABIC	HQ	BIC	X2	df	p	DIF Content
b1	-53.868	-43.153	-46.854	-33.619	59.868	3	0	Significant
b2	-24.422	-13.706	-17.408	-4.173	30.422	3	0	Significant
b4	-24.566	-13.850	-17.552	-4.317	30.566	3	0	Significant
b5	-24.759	-14.043	-17.745	-4.510	30.759	3	0	Significant
b6	-36.048	-25.333	-29.034	-15.800	42.048	3	0	Significant
b7	-34.480	-23.764	-27.466	-14.231	40.48	3	0	Significant
b8	-78.366	-67.650	-71.352	-58.117	84.366	3	0	Significant
b9	-34.557	-23.841	-27.543	-14.308	40.557	3	0	Significant
b11	1.339	12.055	8.353	21.588	4.661	3	0.198	Insignificant
b12	-24.678	-13.962	-17.664	-4.429	30.678	3	0	Significant
b13	-3.850	6.866	3.164	16.399	9.85	3	0.02	Significant
b14	-1.832	8.884	5.182	18.417	7.832	3	0.05	Significant
b15	-20.771	-10.056	-13.757	-0.522	26.771	3	0	Significant
b16	-12.080	-1.365	-5.066	8.168	18.08	3	0	Significant
b17	-43.930	-33.214	-36.915	-23.681	49.93	3	0	Significant
b18	-20.859	-10.143	-13.845	-0.610	26.859	3	0	Significant
b19	-16.581	-5.865	-9.567	3.668	22.581	3	0	Significant
b20	-96.311	-85.596	-89.297	-76.062	102.311	3	0	Significant
b21	-4.026	6.689	2.988	16.223	10.026	3	0.018	Significant
b22	-31.929	-21.214	-24.915	-11.680	37.929	3	0	Significant
b23	-0.309	10.406	6.705	19.939	6.309	3	0.097	Insignificant
b24	-18.610	-7.894	-11.596	1.639	24.61	3	0	Significant
b25	-81.503	-70.787	-74.489	-61.254	87.503	3	0	Significant
b26	-73.928	-63.213	-66.914	-53.679	79.928	3	0	Significant
b27	-63.451	-52.736	-56.437	-43.202	69.451	3	0	Significant
b28	-39.560	-28.845	-32.546	-19.311	45.56	3	0	Significant
b29	-7.051	3.665	-0.037	13.198	13.051	3	0.005	Significant
b30	-58.858	-48.143	-51.844	-38.609	64.858	3	0	Significant

DIF by place of origin

Table 7 shows that Items b13 and b12 in the Java group have negative d-parameter values (–4.016 and –4.443), meaning that these items are very difficult for respondents from Java. On the other hand, Items b7, b18, and b30 have positive d-parameter values (1.529, 1.346, and 0.753), indicating that these items are relatively easier for respondents from Java. High discrimination parameters (a1), such as in Items b13 (2.031) and b12 (1.884), indicate that these items are effective in differentiating respondents based on their abilities. The results of the analysis based on the non-Java group show that Item b13 also has a negative d-value (-8.857), indicating significant difficulty for respondents from non-Java. On the other hand, Item b11 has a high positive d-value (0.293), indicating that this item is easier for respondents from non-Java. High discrimination parameter (a1) is seen in Item b13 (4.649), indicating effectiveness in differentiating respondents based on ability.

A comparison between Java and Outside Java shows a striking difference in Items b13 and b12, where the d-value is very negative for both groups but more extreme in Outside Java. Item b7 has a positive and high d-value in both groups, although the d-value is higher in Java (1.529) than Outside Java (0.725). This means that this item is easier for respondents from Java. Items such as b11 and b16 show similarities in the d-parameter between the two groups, indicating that the difficulty of this item is relatively consistent regardless of the respondent's origin.

Table 7. Item parameter for participants' place of origin

Item	Java			Outside Java		
	a1	d	g	a1	d	g
b1	0.924	-0.222	0.087	0.760	-0.517	0.003
b2	1.191	1.127	0.016	1.471	0.297	0.106
b4	0.400	0.770	0.090	0.361	0.741	0.041
b5	1.048	0.689	0.011	1.753	-0.936	0.262
b6	0.984	0.377	0.002	1.468	-0.599	0.128
b7	1.668	1.529	0.210	1.564	0.725	0.219
b8	1.215	0.314	0.008	1.572	-0.706	0.097
b9	0.890	0.938	0.317	1.043	0.842	0.100
b11	0.041	0.180	0.020	0.163	0.293	0.009
b12	1.884	-4.443	0.371	0.032	-0.405	0.011
b13	2.031	-4.016	0.193	4.649	-8.857	0.225
b14	1.294	0.561	0.325	1.781	-0.413	0.320
b15	1.565	0.448	0.395	1.603	-0.362	0.443
b16	0.103	0.448	0.020	0.262	0.282	0.016
b17	1.961	4.749	0.115	2.039	3.239	0.336
b18	1.134	0.346	0.062	1.549	-0.526	0.130
b19	0.004	-1.631	0.029	0.093	-1.256	0.022
b20	2.517	1.248	0.211	2.439	-0.057	0.279
b21	1.846	-1.172	0.159	2.417	-2.467	0.172
b22	1.423	1.581	0.007	1.617	0.649	0.111
b23	1.494	-0.159	0.295	1.676	-1.034	0.257
b24	1.340	-0.368	0.180	1.298	-1.053	0.203
b25	1.653	2.292	0.005	1.635	1.459	0.019
b26	0.558	0.661	0.012	0.708	0.399	0.005
b27	1.193	-0.279	0.010	1.672	-1.477	0.121
b28	1.090	0.483	0.063	1.598	-0.569	0.197
b29	1.665	0.120	0.443	1.959	-1.054	0.440
b30	1.282	0.753	0.029	1.794	-0.489	0.175

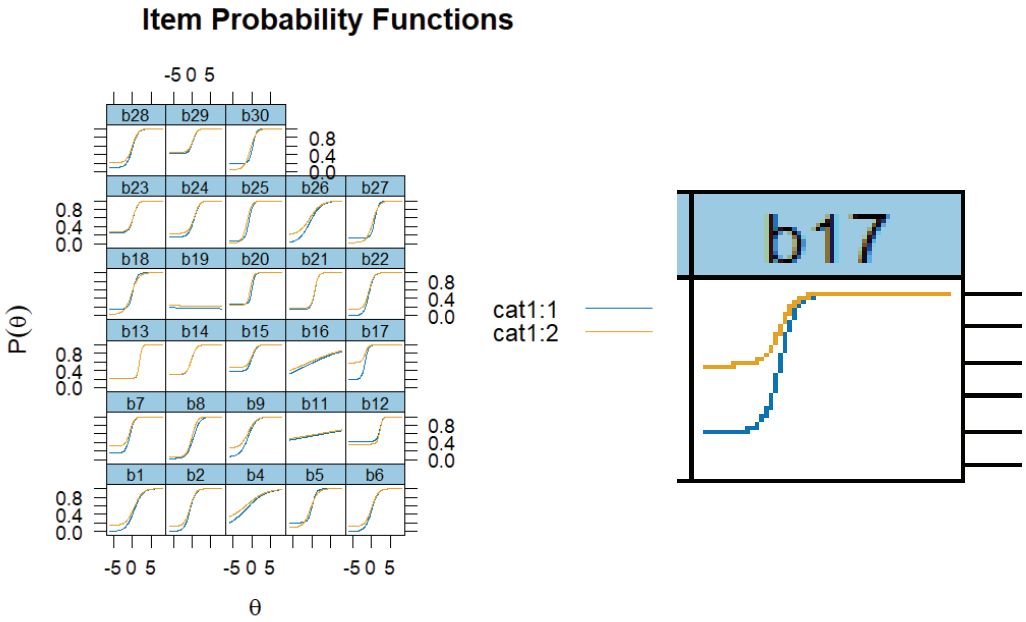

Figure 3. ICC DIF based on test-takers' place of origin

Figure 3 shows the item probability function for 28 test items based on two regional categories (cat1: Java and cat1: Outside Java). The horizontal axis (θ) represents the ability level of the participants, while the vertical axis (P(θ)) shows the probability of answering correctly for each item. The blue (cat1: Java) and yellow (cat1: Outside Java) curves provide important information about the differences in performance between the two regional groups. The difficulty level of an item can be interpreted from the position of the curve on the horizontal axis. Items that are more to the left (e.g., Items b1 to b6) are easier because the probability of a correct response increases at lower θ values. Conversely, items that are more to the right (such as Items b13 to b16) indicate more difficult items, requiring higher θ values to achieve the same response probability.

The slope of each item's curve provides information about the item's discriminatory power. A curve with a steeper slope indicates an item with high discriminatory power, meaning that the item is more effective in distinguishing respondents with different ability levels. For example, some items, such as b13 to b22, show variations in the slope of the curve, indicating differences in their discriminatory power. The difference between the two curves on each item indicates the difference in response probability between the two regions of Java and Outside Java. If the blue and yellow curves are close together or overlap, such as in Items b1 to b6, this indicates that the difference between the response categories is not significant at various ability levels. However, other items, such as Item b17, show a clearer distance between the blue and yellow curves, indicating a greater difference between the Java region (blue curve) and the Outside Java region (yellow curve) at the same ability level. In addition, several items, such as Items b2,

b5, b6, b28, b30 indicate that the DIF formed is a non-uniform DIF; other items tend to form a uniform DIF.

Table 8. The Result of Goodness of Fit

Item	AIC	SABIC	HQ	BIC	X2	df	p	DIF Content
b1	-72.927	-62.212	-65.913	-52.679	78.927	3	0	Significant
b2	-81.393	-70.678	-74.379	-61.144	87.393	3	0	Significant
b4	2.571	13.287	9.585	22.820	3.429	3	0.33	Significant
b5	-100.514	-89.799	-93.500	-80.266	106.514	3	0	Significant
b6	-71.051	-60.335	-64.037	-50.802	77.051	3	0	Significant
b7	-62.594	-51.879	-55.580	-42.345	68.594	3	0	Significant
b8	-92.746	-82.031	-85.732	-72.498	98.746	3	0	Significant
b9	-66.930	-56.214	-59.915	-46.681	72.93	3	0	Significant
b11	23.471	34.187	30.485	43.720	-17.471	3	NaN	Undetermined
b12	26.875	37.591	33.889	47.124	-20.875	3	NaN	Undetermined
b13	-1.750	8.966	5.264	18.499	7.75	3	0.051	Significant
b14	-93.911	-83.196	-86.897	-73.662	99.911	3	0	Significant
b15	-26.491	-15.775	-19.477	-6.242	32.491	3	0	Significant
b16	12.907	23.622	19.921	33.155	-6.907	3	NaN	Undetermined
b17	-51.239	-40.523	-44.224	-30.990	57.239	3	0	Significant
b18	-70.547	-59.832	-63.533	-50.298	76.547	3	0	Significant
b19	-18.829	-8.113	-11.815	1.420	24.829	3	0	Significant
b20	-81.999	-71.284	-74.985	-61.750	87.999	3	0	Significant
b21	-51.451	-40.736	-44.437	-31.203	57.451	3	0	Significant
b22	-87.776	-77.060	-80.762	-67.527	93.776	3	0	Significant
b23	-91.493	-80.777	-84.479	-71.244	97.493	3	0	Significant
b24	-43.901	-33.185	-36.887	-23.652	49.901	3	0	Significant
b25	-78.153	-67.437	-71.138	-57.904	84.153	3	0	Significant
b26	-25.276	-14.560	-18.262	-5.027	31.276	3	0	Significant
b27	-61.677	-50.962	-54.663	-41.428	67.677	3	0	Significant
b28	-62.852	-52.136	-55.838	-42.603	68.852	3	0	Significant
b29	-74.659	-63.943	-67.645	-54.410	80.659	3	0	Significant
b30	-102.749	-92.034	-95.735	-82.501	108.749	3	0	Significant

Table 8 shows the results of the analysis to assess the model fit using the AIC (Akaike Information Criterion), SABIC (Sample-size Adjusted Bayesian Information Criterion), HQ (Hannan-Quinn Criterion), and BIC (Bayesian Information Criterion) values. Most items have negative values, indicating that the model fits the existing data. In addition, Table 7 shows that all items have a df of 3, indicating consistency in the model used. The Chi-square test results show that all items have a very low p-value (0), indicating that the results are statistically significant. This shows that the model is significantly different from the null model, which confirms that the data fit the model. Items that are stated as "Significant" indicate significant DIF. However, some items, such as Items b11, b12, and b16, have "Undetermined" results for DIF.

DFT by gender

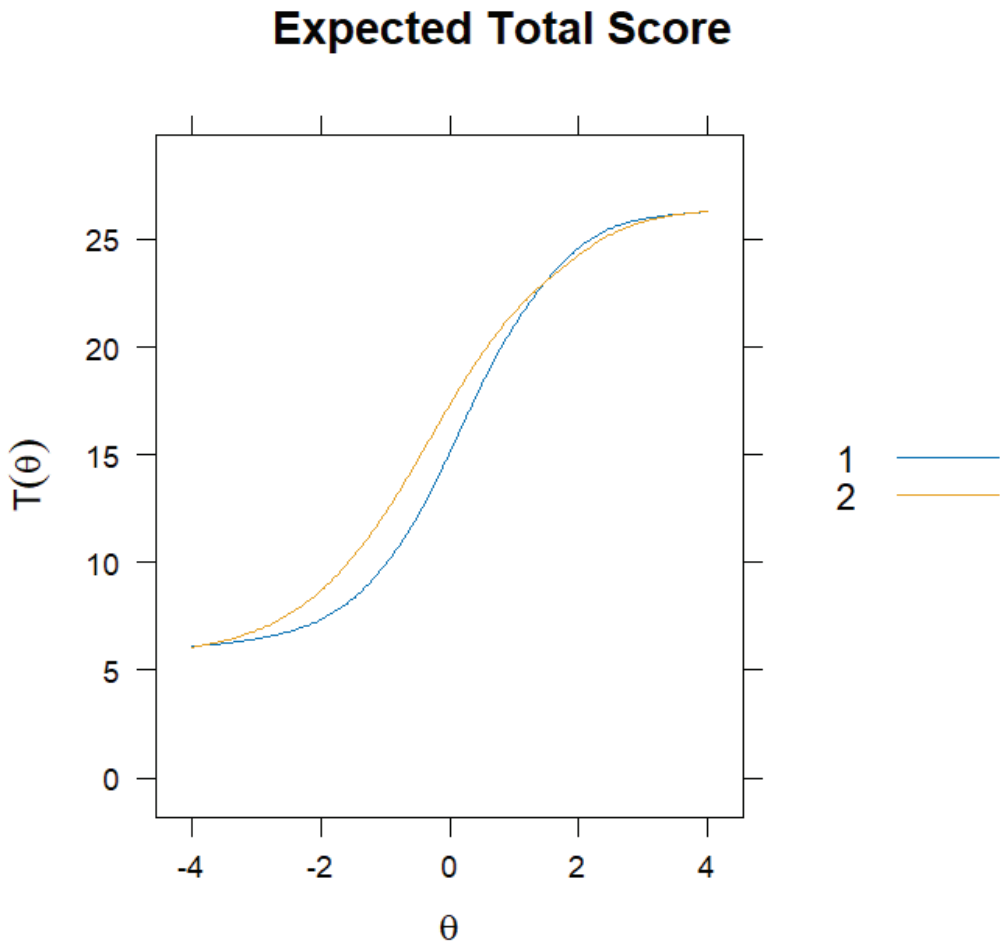

Figure 4. TCC DTF based on test-takers' sex

Figure 4 shows the Differential Total Function (DTF) analysis by gender, where the blue curve represents the male group, while the yellow curve represents the female group. Overall, both curves have similar shapes but are non-uniform DTFs. Figure 4 shows that the relationship between ability and expected total score is consistent across the two gender groups. However, there are a few differences across the ability range. For example, at low ability (around -4 to 0), the curve for the female group (yellow) is slightly higher than the curve for the male group (blue). This suggests that at a low ability level, females tend to obtain higher total scores than males for the same ability level. At medium ability levels (around 0 to 2), the two curves begin to approach each other, suggesting that the difference in scores between males and females becomes less significant. At high ability levels (above 2), the two curves tend to converge, indicating that at a high ability level, the expected total scores for males and females are the same.

The difference between the two curves indicates a slight Differential Total Function (DTF) based on gender in this test. This DTF shows that the items in this test function differently for males and females at a given ability

level. Although this difference is trivial, it is important to note because it can affect the interpretation of test results and indicate potential bias that needs to be corrected to ensure the fairness and validity of the measurement.

DTF by origin

Figure 5 is the DTF by region, with the blue curve representing the Java region and the yellow curve representing the Outside Java region. At low theta (θ) values (from around -4 to 0), the blue (Java) and yellow (Outside Java) curves are very close, indicating that there is no significant difference in expected total scores between test takers from the two regions at low ability levels. However, at higher theta (θ) values (from 0 to 4), the blue (Java) curve is slightly higher than the yellow (Outside Java) curve. This suggests that at medium to high ability levels, test takers from the Java region tend to have slightly higher total scores compared to individuals from the Outside Java region. The differences seen in the DTF graphs in Figure xxx suggest that there are slight differences in how the test performs for individuals from Java and non-Java, particularly at the intermediate to high ability levels. These differences need further investigation to ensure the test is fair and does not favor one group based on region of origin.

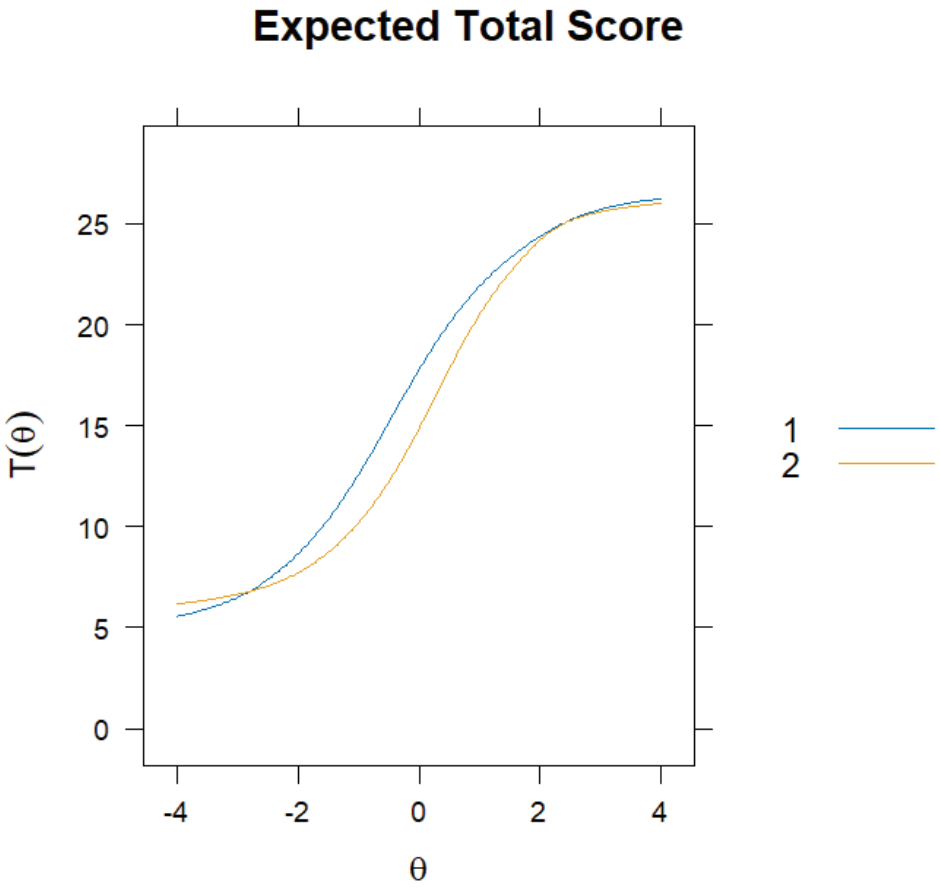

Figure 5. TCC DTF berdasarkan wilayah asal peserta tes

Discussion

The results of the DIF detection showed that 26 test items showed gender bias. This means test items provide unfair advantages or disadvantages to certain gender groups. We found that the bias was in favor of the male gender group. This means the items were easier for or more relevant to men than to women. This finding further emphasizes that gender is one of the factors that causes DIF (Büyükkidik, 2023; Cai & Albano, 2018; Ra & Rhee, 2018). This occurs because of differences in experience, perception, and interpretation between males and females towards test items. These factors can include social and cultural differences, where males and females may have different educational experiences and social contexts, thus affecting the way they answer the questions in the test. For example, items related to sports or household activities may be more familiar to one gender than the other, leading to differences in the level of difficulty of the items (Chubbuck et al., 2016). In addition, there are differences in interest and engagement between males and females in some subjects; men tend to be more interested in topics related to science and technology, while females may be more interested in the humanities and arts (Kans & Claesson, 2022).

These differences may affect the level of motivation and confidence in answering the questions related to a particular topic. Males and females may also have different learning and problem-solving styles. Males are more likely to use an analytical approach. Females may be more likely to use a holistic approach, which affects how they understand and answer the questions in the test, especially those requiring specific problem-solving strategies (Kheder & Rouabhia, 2023; Waschl & Burns, 2020). Additionally, females tend to be better at tasks involving language and verbal comprehension, while males may excel at tasks involving spatial and mechanical comprehension (Granocchio et al., 2023; Hirnstein et al., 2023; Kruchinina et al., 2020; Rinaldi et al., 2023). Items with more complex language or technical contexts may favor one gender. An item that clearly shows a bias favoring the male gender is Item 17.

Figure 6 shows that Item 17 is more likely to be answered correctly by males because the topic of the question is related to monetary policy, inflation, and the role of central banks - material that is more often accessed or of interest to males than females (Bodea & Kerner, 2021; Diouf & Pépin, 2017). This is due to the tendency of males to be more interested in economic and financial topics (Förster & Happ, 2019; Kruger, 2008). In addition to gender, we also found that 25 items showed bias based on the region of origin of the test taker. The test items tended to favor test takers from Java compared to test-takers from outside Java. This finding further emphasizes that the region of origin of test takers, especially between Java and outside Java, can be a factor that causes DIF (Wulandari et al., 2023; Yüksel et al., 2019). Some reasons for this are differences in access to education, culture, and language that affect how test takers understand and answer test items.

Test-takers from Java have better access to educational resources, such as more experienced teachers, more complete learning materials, and better facilities (Dewi et al., 2022; Otok et al., 2021). This gives them an advantage in answering test items that may be designed based on a more general curriculum or educational standards in the Java region.

Culture also plays a significant role in causing DIF. Non-Javanese participants have different cultural backgrounds, which affect how they interpret and respond to test items (Magdolen et al., 2020; Sauer et al., 2018; Wu et al., 2013). For example, items that contain cultural references or everyday practices specific to Java may be irrelevant or even foreign to non-Javanese participants. These differences can lead to subjective testing because non-Javanese participants may encounter difficult items. Language is also a critical factor in the occurrence of DIF (Gibbons et al., 2011; Mach, 2023). The language of instruction in Java may be more standardized and closer to the language used in the test items. Non-Javanese participants may be more familiar with different dialects or regional languages. Language difficulty in understanding the language of the test may put non-Javanese participants at a disadvantage in answering test items, even though they have the same ability in the material being tested.

Item 17 (see Figure 6 and Figure 3) contains the issue of the monetary crisis that occurred in 1997-1998. During the 1997-1998 financial crisis, more people in Java felt it compared to people outside Java. This could be one of the factors that influenced the test results on questions about inflation targets. As the center of Indonesia's economy and finance, Java experienced a more severe impact of the crisis due to the high concentration of economic activities. The decline in the exchange rate and high inflation in Java caused test takers from this region to be more aware and have better knowledge of monetary policy and inflation targets. On the other hand, participants from outside Java may not have felt the impact of the crisis as intensely, so their level of understanding of topics such as inflation targets could be less in-depth.

Our findings support a study by Ahmadi and Jalili (2014) who investigated the sources of DIF in an English as a Foreign Language reading comprehension test among Iranian test-takers and identified DIF related to location and educational level. Prieto and Nieto (2014) also found that the presence of uniform and non-uniform DIF in test items suggests some questions behave differently for Italians and Asians based on their native language. Additionally, Balluerka et al. (2014) used multilevel logistic regression to explore the causes of DIF in a short test, focusing on attitudes toward science in Spanish and English students. Roever's (2007) study investigating DIF in an English as a Second Language test also found similar results, identifying items that function differently for test takers from Asian and European backgrounds.

CONCLUSION

The results of the DIF detection show that 26 test items are gender bias, which tend to benefit male test-takers. This finding confirms that gender is a factor causing DIF, influenced by differences in experience, perception, and interests between males and females. In addition, 25 test items are also biased based on the region of origin of the test takers, with test takers from Java being more advantaged than those from outside Java. The causal factors include different access to education, culture, and language. To overcome DIF caused by gender and region of origin, several steps can be taken. First, identify and revise test items that show gender or regional bias. Second, ask a panel of experts to assess gender equality and understand the context of different regions. Third, analyze with Item Response Theory to ensure that test items work objectively for all participants, regardless of gender or region of origin. In addition, providing training to test developers on the importance of gender equality and encouraging the development of items that are gender-neutral and relevant to the experiences of both genders and all regions is essential. By understanding the factors that cause DIF and implementing strategies to reduce them, we can improve fairness in assessment and provide all individuals with equal opportunities.

REFERENCES

Ahmadi, A., & Jalili, T. (2014). A confirmatory study of differential item functioning on EFL reading comprehension. *Applied Research on English Language, 3*(2), 55–68

Andrich, D., & Marais, I. (2019). Fit of responses to the model III—differential item functioning (pp. 199–208). https://doi.org/10.1007/978-981-13-7496-8_16

Argianti, A., & Retnawati, H. (2020). Characteristics of math national-standardized school exam test items in junior high school: What must be considered*? Jurnal Penelitian dan Evaluasi Pendidikan, 24*(2). https://doi.org/10.21831/pep.v24i2.32547

Babu, N., & Kohli, P. (2023). Commentary: Reliability in research. *Indian Journal of Ophthalmology, 71*(2), 400. https://doi.org/10.4103/ijo.IJO_2016_22

Bademci, V. (2022). Correcting fallacies about validity as the most fundamental concept in educational and psychological measurement. *International E-Journal of Educational Studies, 6*(12), 148–154. https://doi.org/10.31458/iejes.1140672

Baker, F. B., & Kim, S.-H. (2017). Item characteristic curve models (pp. 17–34). https://doi.org/10.1007/978-3-319-54205-8_2

Balluerka, N., Plewis, I., Gorostiaga, A., & Padilla, J.-L. (2014). Examining sources of DIF in psychological and educational assessment using multilevel logistic regression. *Methodology, 10*(2), 71–79. https://doi.org/10.1027/1614-2241/a000076

Bilyakovska, O. (2022). Test as an effective means of assessing the quality of students' knowledge. *Academic Notes Series Pedagogical Science, 1*(204), 16–20. https://doi.org/10.36550/2415-7988-2022-1-204-16-20

Bodea, C., & Kerner, A. (2021). The gender credibility gap: All-male boards and substantive gender representation in central banking. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3780220

Bormanaki, H. B., & Ajideh, P. (2022). Item performance across native language groups on the Iranian National University entrance English exam: A nationwide study. *Language Testing in Asia, 12*(1), 29. https://doi.org/10.1186/s40468-022-00185-2

Büyükkidik, S. (2023). Purification procedures used for the detection of gender DIF: Item bias in a foreign language test. *International Journal of Assessment Tools in Education, 10*(4), 765–780. https://doi.org/10.21449/ijate.1250358

Cai, L. S., & Albano, A. D. (2018). Examining sources of gender DIF in mathematics knowledge of future teachers using cross-classified IRT models. In *Exploring the Mathematical Education of Teachers Using TEDS-M Data* (pp. 543–561). Springer International Publishing. https://doi.org/10.1007/978-3-319-92144-0_19

Canay, I. A., Mogstad, M., & Mountjoy, J. (2022). On the use of outcome tests for detecting bias in decision making. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4156834

Chubbuck, K., Curley, W. E., & King, T. C. (2016). Who's on first? Gender differences in performance on the SAT ® test on critical reading items with sports and science content. *ETS Research Report Series, 2016*(2), 1–116. https://doi.org/10.1002/ets2.12109

Dewi, D. M., Saingan, A. F., & Fahmi, Y. (2022). Kontribusi teknologi informasi dan komunikasi terhadap rata-rata lama sekolah di pulau Jawa. *PAKAR Pendidikan, 20*(1), 24–36. https://doi.org/10.24036/pakar.v20i1.248

Dewi, H. H., Damio, S. M., & Sukarno, S. (2023). Item analysis of reading comprehension questions for English proficiency test using Rasch model. *REID (Research and Evaluation in Education), 9*(1), 24–36. https://doi.org/10.21831/reid.v9i1.53514

Diouf, I., & Pépin, D. (2017). Gender and central banking. *Economic Modelling, 61*, 193–206. https://doi.org/10.1016/j.econmod.2016.12.006

Dubbelman, M. A., Verrijp, M., Facal, D., Sánchez-Benavides, G., Brown, L. J. E., der Flier, W. M., Jokinen, H., Lee, A., Leroi, I., Lojo-Seoane, C., Milošević, V., Molinuevo, J. L., Pereiro Rozas, A. X., Ritchie, C., Salloway, S., Stringer, G., Zygouris, S., Dubois, B., Epelbaum, S., … Sikkes, S. A. M. (2020). The influence of diversity on the measurement of functional impairment: An international validation of the Amsterdam IADL questionnaire in eight countries. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 12*(1). https://doi.org/10.1002/dad2.12021

Effiom, A. P. (2021). Test fairness and assessment of differential item functioning of mathematics achievement test for senior secondary students in Cross River state, Nigeria using item response theory. *Global Journal of Educational Research, 20*(1), 55–62. https://doi.org/10.4314/gjedr.v20i1.6

Eliaumra, E., Samaela, D. P., & Muhdin, N. K. (2022). Developing diagnostic test assessment to measure creative thinking skills of Biology preservice teacher students. *REID (Research and Evaluation in Education), 8*(2), 152–168. https://doi.org/10.21831/reid.v8i2.50885

Förster, M., & Happ, R. (2019). The relationship among gender, interest in economic topics, media use, and the economic knowledge of students at vocational schools. *Citizenship, Social and Economics Education, 18*(3), 143–157. https://doi.org/10.1177/2047173419892209

Fulcher, G. (2016). Context and inference in language testing. In *The Dynamic Interplay between Context and the Language Learner* (pp. 225–241). Palgrave Macmillan UK. https://doi.org/10.1057/9781137457134_12

Gibbons, L., Crane, P. K., Mehta, K. M., Pedraza, O., Tang, Y., Manly, J. J., Narasimhalu, K., Teresi, J., Jones, R. N., & Mungas, D. (2011). Multiple, correlated covariates associated with differential item functioning (DIF): Accounting for language DIF when education levels differ across languages. *Ageing Research, 2*(1), 4. https://doi.org/10.4081/ar.2011.e4

Granocchio, E., De Salvatore, M., Bonanomi, E., & Sarti, D. (2023). Sex-related differences in reading achievement. *Journal of Neuroscience Research, 101*(5), 668–678. https://doi.org/10.1002/jnr.24913

Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health and Quality of Life Outcomes, 15*(1), 181. https://doi.org/10.1186/s12955-017-0755-0

Hagquist, C., Due, P., Torsheim, T., & Välimaa, R. (2019). Cross-country comparisons of trends in adolescent psychosomatic symptoms – a Rasch analysis of HBSC data from four Nordic countries. *Health and Quality of Life Outcomes, 17*(1), 27. https://doi.org/10.1186/s12955-019-1097-x

Hirnstein, M., Stuebs, J., Moè, A., & Hausmann, M. (2023). Sex/gender differences in verbal fluency and verbal-episodic memory: A meta-analysis. *Perspectives on Psychological Science, 18*(1), 67–90. https://doi.org/10.1177/17456916221082116

Ito, S., Nagao, H., Kurokawa, T., Kasuya, T., & Inoue, J. (2019). Bayesian inference of grain growth prediction via multi-phase-field models. *Physical Review Materials, 3*(5), 053404. https://doi.org/10.1103/PhysRevMaterials.3.053404

Joo, S.-H., Lee, P., & Stark, S. (2022). Bayesian approaches for detecting differential item functioning using the generalized graded unfolding model. *Applied Psychological Measurement, 46*(2), 98–115. https://doi.org/10.1177/01466216211066606

Kans, M., & Claesson, L. (2022). Gender-related differences for subject interest and academic emotions for STEM subjects among Swedish Upper secondary school students. *Education Sciences, 12*(8), 553. https://doi.org/10.3390/educsci12080553

Kheder, K., & Rouabhia, R. (2023). Gender differences in learning languages. European Journal of *Applied Linguistics Studies, 6*(2). https://doi.org/10.46827/ejals.v6i2.456

Kruchinina, O. V., Stankova, E. P., & Galperina, E. I. (2020). Development of spatiotemporal EEG organization in males and females aged 8–30 years during comprehension of oral and written texts. *Human Physiology, 46*(3), 244–256. https://doi.org/10.1134/S036211972003010X

Kruger, D. J. (2008). Male financial consumption is associated with higher mating intentions and mating success. *Evolutionary Psychology, 6*(4), 147470490800600. https://doi.org/10.1177/147470490800600407

Leventhal, B., & Gregg, N. (2022). Reliability and measurement error. In *Reliability and Measurement Error*. Routledge. https://doi.org/10.4324/9781138609877-REE28-1

Li, L., & Becker, B. J. (2021). Assessing differential bundle functioning using meta-analysis. *Journal of Educational Measurement, 58*(4), 492–514. https://doi.org/10.1111/jedm.12303

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge. https://doi.org/10.4324/9780203056615

Mach, T. (2023). Literatur im DaF-Unterricht: Einige kritische Anmerkungen. *AUC PHILOLOGICA, 2022*(3), 119–133. https://doi.org/10.14712/24646830.2023.6

Magdolen, M., Behren, S. von, Hobusch, J., Chlond, B., & Vortisch, P. (2020). Comparison of response bias in an intercultural context – evaluation of psychological items in travel behavior research. *Transportation Research Procedia, 48*, 2891–2905. https://doi.org/10.1016/j.trpro.2020.08.231

Makkink, A. W., & Vincent-Lambert, C. (2020). The development of 'SATLAB': A tool designed to limit assessment bias in simulation-based learning. *South African Journal of Pre-Hospital Emergency Care, 1*(1), 26–34. https://doi.org/10.24213/1-1-3024

Moradi, E., Ghabanchi, Z., & Pishghadam, R. (2022). Reading comprehension test fairness across gender and mode of learning: insights from IRT-based differential item functioning analysis. *Language Testing in Asia, 12*(1), 39. https://doi.org/10.1186/s40468-022-00192-3

Nedungadi, S., Brown, C. E., & Paek, S. H. (2022). Differential item functioning analysis of the fundamental concepts for organic reaction mechanisms inventory. *Journal of Chemical Education, 99*(8), 2834–2842. https://doi.org/10.1021/acs.jchemed.2c00242

Nurrahman, A., Sukirno, S., Pratiwi, D. S., Iskandar, J., Rahim, A., & Rahmaini, I. S. (2022). Developing student social attitude self-assessment instruments: A study in vocational high school. *REID (Research and Evaluation in Education), 8*(1), 1–12. https://doi.org/10.21831/reid.v8i1.45100

Otaya, L. G., Kartowagiran, B., & Retnawati, H. (2020). The construct validity and reliability of the lesson plan assessment instrument in primary schools. *Jurnal Prima Edukasia, 8*(2), 126–134. https://doi.org/10.21831/jpe.v8i2.33275

Otok, B. W., Suharsono, A., Purhadi, Standsyah, R. E., & Azies, H. Al. (2021). A meta confirmatory factor analysis of the underdeveloped areas in the Java Island. 020002. https://doi.org/10.1063/5.0059540

Patnala, V., Salla, G. R., Prabhakar, S., Singh, R. P., & Annapureddy, V. (2024). Analysing the grain size and asymmetry of the particle distribution using auto-correlation technique. *Applied Physics A, 130*(3), 191. https://doi.org/10.1007/s00339-024-07332-x

Penfield, R. D., & Camilli, G. (2006). 5 differential item functioning and item bias (pp. 125–167). https://doi.org/10.1016/S0169-7161(06)26005-X

Prieto, G., & Nieto, E. (2014). Influence of DIF on differences in performance of Italian and Asian individuals on a reading comprehension test of Spanish as a foreign language (negative emotionality) in Hong Kong. *Journal of Applied Measurement, 15*(2), 176–188

Ra, J., & Rhee, K. J. (2018). Detection of Gender related DIF in the Foreign Language Classroom Anxiety Scale. *Educational Sciences: Theory & Practice*. https://doi.org/10.12738/estp.2018.1.0606

Rasooli, A., Zandi, H., & DeLuca, C. (2019). Conceptualising fairness in classroom assessment: Exploring the value of organisational justice theory. *Assessment in Education: Principles, Policy & Practice, 26*(5), 584–611. https://doi.org/10.1080/0969594X.2019.1593105

Retnawati, H. (2016). Proving content validity of self-regulated learning scale (The comparison of Aiken index and expanded Gregory index). *REID (Research and Evaluation in Education), 2*(2), 155–164. https://doi.org/10.21831/reid.v2i2.11029

Rinaldi, P., Pasqualetti, P., Volterra, V., & Caselli, M. C. (2023). Gender differences in early stages of language development. Some evidence and possible explanations. *Journal of Neuroscience Research, 101*(5), 643–653. https://doi.org/10.1002/jnr.24914

Roever, C. (2007). DIF in the Assessment of second language pragmatics. *Language Assessment Quarterly, 4*(2), 165–189. https://doi.org/10.1080/15434300701375733

Sauer, J., Sonderegger, A., & Hoyos Álvarez, M. A. (2018). The influence of cultural background of test participants and test facilitators in online product evaluation. *International Journal of Human-Computer Studies, 111*, 92–100. https://doi.org/10.1016/j.ijhcs.2017.12.001

Setiawan, A., Cendana, W., Ayres, M., Yuldashev, A. A., & Setyawati, S. P. (2023). Development and validation of a self-assessment-based instrument to measure elementary school students' attitudes in online learning. *REID (Research and Evaluation in Education), 9*(2), 184–197. https://doi.org/10.21831/reid.v9i2.52083

Setiawati, F. A., Ayriza, Y., Retnowati, E., & Amelia, R. N. (2017). The response patterns of the career interest instrument based on Holland's theory. *ANIMA Indonesian Psychological Journal, 32*(3), 128–147. https://doi.org/10.24123/aipj.v32i3.628

Shykhnenko, K. I. (2020). Optimising assessment system in the ESP course through the use of the methods of differential item functioning and differential test functioning in final test design. *Zhytomyr Ivan Franko State University Journal. Pedagogical Sciences, 2*(101), 156–165. https://doi.org/10.35433/pedagogy.2(101).2020.156-165

Sumin, S., Sukmawati, F., & Nurdin, N. (2022). Gender differential item functioning on the Kentucky Inventory of Mindfulness Skills instrument using logistic regression. *REID (Research and Evaluation in Education), 8*(1), 55–66. https://doi.org/10.21831/reid.v8i1.50809

Szmańda, J. B., & Witkowski, K. (2021). Morphometric parameters of Krumbein Grain Shape Charts—A critical approach in light of the automatic grain shape image analysis. *Minerals, 11*(9), 937. https://doi.org/10.3390/min11090937

Terluin, B., Brouwers, E. P. M., Marchand, M. A. G., & de Vet, H. C. W. (2018). Assessing the equivalence of Web-based and paper-and-pencil questionnaires using differential item and test functioning (DIF and DTF) analysis: A case of the four-dimensional symptom questionnaire (4DSQ). *Quality of Life Research, 27*(5), 1191–1200. https://doi.org/10.1007/s11136-018-1816-5

Walker, C. M., & Gocer Sahin, S. (2023). Differential functioning. In *International Encyclopedia of Education (Fourth Edition)* (pp. 249–259). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10035-1

Wallace, M. P. (2018). Fairness and justice in L2 classroom assessment: Perceptions from test takers. *The Journal of AsiaTEFL, 15*(4), 1051–1064. https://doi.org/10.18823/asiatefl.2018.15.4.11.1051

Wallin, G., Chen, Y., & Moustaki, I. (2023). DIF analysis with unknown groups and anchor items. *Psychometrika, 89*(1), 267–295. https://doi.org/10.1007/s11336-024-09948-7

Wallin, G., Chen, Y., & Moustaki, I. (2024). DIF analysis with unknown groups and anchor items. *Psychometrika, 89*(1), 267–295. https://doi.org/10.1007/s11336-024-09948-7

Waschl, N., & Burns, N. R. (2020). Sex differences in inductive reasoning: A research synthesis using meta-analytic techniques. *Personality and Individual Differences, 164*, 109959. https://doi.org/10.1016/j.paid.2020.109959

Wilsa, A. W., Rusilowati, A., Susilaningsih, E., Jaja, J., & Nurpadillah, V. (2023). Validity, reliability, and item characteristics of cell material science literacy assessment instruments. *Jurnal Penelitian Dan Evaluasi Pendidikan, 27*(2), 177–188. https://doi.org/10.21831/pep.v27i2.61577

Wu, S., Barr, D. J., Gann, T. M., & Keysar, B. (2013). How culture influences perspective taking: Differences in correction, not integration. *Frontiers in Human Neuroscience, 7*. https://doi.org/10.3389/fnhum.2013.00822

Wulandari, R. D., Laksono, A. D., Rohmah, N., & Ashar, H. (2023). Regional differences in primary healthcare utilization in Java Region—Indonesia. *PLOS ONE, 18*(3), e0283709. https://doi.org/10.1371/journal.pone.0283709

Yavuz Temel, G. (2023). A simulation and empirical study of differential test functioning (DTF). *Psych, 5*(2), 478–496. https://doi.org/10.3390/psych5020032

Yüksel, S., Demir, P., & Alkan, A. (2019). Factors causing occurrence of artificial dif: A simulation study for dichotomous data. *Communications in Statistics - Simulation and Computation, 48*(7), 2004–2011. https://doi.org/10.1080/03610918.2018.1429622