Sociology & Cultural Research Review (SCRR)
Available Online: https://scrrjournal.com
Print ISSN: 3007-3103 Online ISSN: 3007-3111
Platform & Workflow by: Open Journal Systems
https://doi.org/10.5281/zenodo.18437578

Investigating Differential Item Functioning in Chemistry Test Items of the Secondary School Certificate Examination

Dr. Mobeen Ul Islam
Assistant Professor, Department of Education, University of Gujrat
drmobeen.islam@uog.edu.pk
Hafiz Muhammad Salman Naveed
PhD Scholar, Department of Education, University of Gujrat
salmanvd01@gmail.com
Afeefa Amjad
Department of Education, University of Gujrat

ABSTRACT
The study was aimed at analyzing test items by means of differential item functioning. The test questions are exercises that learners are to undertake during the test. Differentiation between items functioning analysis is highly significant in test development at both national and international levels to confirm the biased and unbiased items in the test. This research was useful in establishing the validity of the test items and the data regarding the degree of difference in item functioning performance regarding chosen groups. It was a quantitative paradigm descriptive study. Data were analyzed using the Mantel-Haenszel. Differential item functioning was useful to assess the strengths and weaknesses of the tests of the target groups with non-similar traits using the test items based on polytomous items in the science subject, that is, Chemistry of grade 10 th, in the annual examination of BISE Sahiwal 2018. The conclusion was reached that there is a necessity to create the test items, which are to be developed by the experts, in order to conduct an appropriate assessment.

Keywords: *DIF, Annual Examination, Mantel-Haenszel, Secondary School Certificate Examination, Chemistry Test Items*

Introduction

The examination systems in Pakistan are critical in determining the course of education in the country, specifically at the secondary level, whereby the Secondary School Certificate (SSC) examination is a key to higher secondary education and future academic prospects. The central location of chemistry among SSC subjects is necessitated by its role in science streams and careers in both medicine, engineering, and applied sciences. Since SSC chemistry examinations are high-stakes tests, the validity and fairness of test scores are an extremely important issue to examination boards, educators, and policymakers.

Different item functioning (DIF) is one of the greatest threats to test fairness, and this happens when examinees in different groups, but with similar underlying ability, have different chances of getting a test item correct (Holland & Wainer, 2012). DIF is item-level behavioral instead of aggregate test scores, and is used extensively to determine possible item bias. In cases when DIF exists, the differences in scores can be based on something besides differences in achievement. In such circumstances, the validity of inferences made based on the results of the examination is compromised (Zumbo, 2019).

The issue of fairness in examinations is aggravated by strong contextual and structural disparities amongst the groups of students in the Pakistani SSC system. The differences can be especially observed in the gender, student enrollment status (private candidates and regular candidates), and the type of school (public school and private school). Pakistan has similar divergent instructional practices, classroom activities, and sociocultural expectations between male and female students with regard to science education. Even though the general gender gaps in science performance have decreased worldwide, item-level research still shows gender-based DIF in science questions, meaning that certain science questions favor one gender over another despite the ability levels being equalized (Liu & Wilson, 2023).

The point that is equally significant in the Pakistani context is the difference between regular students and the candidates of private institutions, whom they prepare by their own or under informal education arrangements. The private candidates usually do not have uniformity in laboratory facilities, systematic teaching, and curriculum encouragement, and such aspects could affect the way they perceive and react to chemistry examination questions. In a similar fashion, the disparities between the public and the private schools, such as the divergence in the teacher education, teaching equipment, and laboratory apparatus, as well as the preparation of the item, can be a source of systematic disparities in the functionality of the item (OECD, 2023). Loss of these contextual differences is likely to obscure item bias that is inherent in the examination system.

The DIF is especially sensitive to chemistry tests as they take both conceptual knowledge and symbolic reasoning, as well as mathematical manipulation and exposure to laboratory settings. Test items can also unintentionally demonstrate assumptions of instructional exposure or familiarity on a contextual basis that are disproportionately represented between the sexes and school sectors (Taber, 2021). Such differences are of interest to psychometric inquiry in the SSC chemistry examinations in Pakistan, which use one standardized paper to test a very heterogeneous population.

Although DIF analysis is crucial in ensuring equity in assessment, there has not been much empirical research done on DIF in SSC-level chemistry examinations in Pakistan. Previous studies with an SSC focus on DIF have mainly considered the fairness of tests in general or focused on mathematics and general science courses, which have not been examined more thoroughly in comparison with chemistry. Further, few studies have concurrently addressed gender, private students and regular students, as well as public schools and private schools within one analytical framework. To fill this gap, the current study examines the phenomenon of Differential Item Functioning in the Chemistry test items of the Secondary School Certificate examination in Pakistan in terms of gender, private or regular student, and public or private school affiliation. Using well-established DIF detection methods, the research will detect the possible biased items, help in the psychometric purification of SSC chemistry tests, and assist in the evidence-based decisions to employ in more fair evaluations in the Pakistani secondary education system.

Research Objectives

The research objectives of the study were:

1. To investigate whether chemistry test items used by 10th-grade students in secondary school examinations exhibit differential item performance for different traits of students (gender, student type, and school type).

Research Questions

The following research questions were developed to achieve the research objectives of the study:

1. Is there any Chemistry test item used in *the Board of Intermediate and Secondary Education that shows differential item functioning across gender groups*?
2. Is there any Chemistry test item used in *the Board of Intermediate and Secondary Education that shows* differential item functioning among regular and private students?
3. Is there any Chemistry test item used in *the Board of Intermediate and Secondary Education that shows* differential item functioning for students' institution type?

Literature Review

Differential Item Functioning (DIF) is a popular psychometric method that is utilized to determine the fairness and validity of educational measurements. DIF analysis examines the performance of test items in use by different groups of examinees with equal levels of underlying ability (Zumbo, 2019). A student shows DIF when there is a difference in the item-level performance by students who have equal overall achievement. However, this difference is caused by the group membership rather than individual proficiency (Holland & Wainer, 2012). The main point of the DIF analysis is to reveal the item bias so that the test developers can revise or eliminate problematic items, thus making sure that the results of the assessment can be interpreted and that the results are fair to all subgroups.

Educational measurement literature includes a significant number of works that use DIF in large-scale and curriculum-based testing. In a state-of-the-art review, it is stressed that DIF has recently become an essential part of test fairness assessment, with Item Response Theory (IRT) and logistic regression being frequently used to measure uniform and non-uniform DIF (Zumbo & Chan, 2022). High-stakes public exams, classroom tests, and international comparative tests have been approached with the DIF methodology frameworks, and differences in item functioning are found according to gender, language, socioeconomic status, and educational conditions (Chen et al., 2025).

Much attention has been paid to the study of gender variations in achievement in science and mathematics. Indicatively, the literature on IRT-based DIF analyses has reported that some science items, particularly those that need spatial learning or familiarity with the surrounding environment, will perform differently between male and female students, which suggests the possibility of item bias (Liu & Wilson, 2023). This trend brings up the importance of subject-specific DIF studies in chemistry tests, where problem representations and cognitive levels can interact with student background factors. Additionally, the global testing, including PISA and TIMSS, has regularly provided assessments of DIF to guarantee that the measurement is the same between the countries involved and among the gender groups (OECD, 2023). These articles prove that the DIF analysis plays a vital role in ensuring sound international comparisons and fair assessment systems.

Along with gender, it is possible to note that the variables of the educational context have a significant effect on the performance of items, as shown in the literature. The school sector differences, including between the public and the privatized institutions, represent the differences in the quality of the instruction, access to resources, and laboratory experience, and each of these might lead to the differences in the results of the science-related assessment (OECD, 2023). Equally, the difference between the private candidates and ordinary students in the public exam systems, such as that of Pakistan, brings another aspect of heterogeneity in the context of the situation. Those candidates who are privately educated take voluntary study, and may not attend school regularly, may have varied curricular priorities and experience of examination formats, which is likely to influence their interpretation and response to particular items.

The study by Islam, Naveed, and Amjad (2025), where the authors treated mathematics test items in the 2018 SSC examination with DIF analysis, is also an important contribution to the research on DIF

in Pakistani SSC. The authors examined the degree to which individual items operated differently in subgroups in mathematics, a topic that has close relations with chemistry in cognitive demand and test design. Their studies found that the DIF framework was useful in determining the strengths and weaknesses of items against the performance of the student groups, and this indicated the usefulness of DIF in item quality improvement in high-stakes testing (Islam et al., 2025). Even though it was a study of mathematics and not chemistry, it created a methodological precedent for the application of DIF techniques in SSC examinations. It highlights the importance of similar analyses in areas of science where item bias can be detrimental in the sense of unfairness.

A number of recent DIF research studies also stress the applicability of psychometric fairness in assessment in education. Indicatively, in a study conducted by Eshun (2025) on DIF in core subjects by classical and IRT-based tests, it was found that although many of the items worked equally across subgroups, a non-trivial percentage of items were seen to have DIF in favor of specific groups. The findings of this kind of evidence support the consideration that systematic DIF analysis is a part and parcel of test evaluation procedures in order to maintain fairness and validity in state-sponsored exams. To a certain extent, recent methodological efforts also point out improvements in DIF detection and interpretation, indicating that any combination of detection methods would be stronger in detecting biased items and bolster any evidence of validity (Zumbo & Chan, 2022).

In spite of these developments, there exists little research that specifically addresses DIF in SSC chemistry examinations. The majority of the available literature focuses on larger-scale science tests or mathematics exams, which creates a gap in the comprehension of how the items in chemistry might disproportionately work depending on gender, student status, and school sector groups. Due to the unique cognitive nature of chemistry, in which conceptual mastery, symbolic logic, laboratory work, and solving mathematical problems all come together, there exists a strong demand to conduct research on DIF in the field of chemistry to assist in the fair distribution of assessments in the Pakistani secondary educational system.

Research Design

The research design adopted by the researcher in this study was the descriptive research design. The researcher described, explained, and validated the findings of the research. In the given study, the research question was to detect the presence of differential item functioning (DIF) of the test items under use in the subject of Chemistry of the first group of test takers at BISE Sahiwal in 2018. Differential Item Functioning is also a method of analysis that is applied to determine biased items in an assessment. In 2003, Zieky identified DIF as an effective method of determining the possible unfairness and evaluating the causes of achievement differences, instead of comparing the overall score. In essence, it is administered between two groupings, the focal group and the reference group of test takers, and serves to reduce the bias of the test.

Population and Sample

The sample size of the present research included all the test items in the subjects of Chemistry, as applied to 80000 students who were the test takers in the 2018 examination at BISE Sahiwal.

Research Tool

The research tool used in the study was the 2018 Chemistry exams MCQs test items of the Board of Intermediate and Secondary Education, Sahiwal. The tool that was used in this research was a set of multiple-choice objective test questions in the year 2018 in Chemistry. There were 51 items in the chemistry test.

Data Collection

The information was collected at the Board of Intermediate Secondary Education, Sahiwal. The objective of the research was explained to the chairman of the Sahiwal board, and he was convinced that the data collected during the test would not be used for any other purpose except research.

Data Analysis

An experiment was done to determine the DIF in test items. First, the data were input into an Excel sheet, and it was categorized into various groups. Then, it was analyzed using the Item Response Theory version 4.4 to identify DIF in items and SPSS software to obtain descriptive statistics.

In this part, an analysis of the data was conducted in order to determine the differential item functioning of the items that were used in the annual 2018 exams of Sahiwal in a secondary school certificate. This part was made up of data interpretation and analysis. Mantel-Haenszel statistic and descriptive statistics approach were used by the researcher to analyze the data.

Question A: Is there any Chemistry test item used in *the Board of Intermediate and Secondary Education that shows differential item functioning across gender groups*?

Table 1: DIF Analysis among Gender Groups

Item No.	Mean		S.D.		Rpbis	DeltaMH	P-value
	Female	Male	Female	Male			
MCQ1	.42	.45	.494	.498	0.603	-0.0922 A	0.9101
MCQ2	.49	.39	.501	.490	0.362	1.3898B	0.0051
MCQ3	.19	.22	.396	.416	0.391	-0.1510 A	0.8506
MCQ4	.41	.37	.492	.485	0.487	0.4254 A	0.3640
MCQ5	.55	.52	.498	.501	0.559	0.5673 A	0.2226
MCQ6	.35	.45	.479	.498	0.676	-0.8370 A	0.0656
MCQ7	.32	.33	.466	.469	0.625	-0.0500 A	0.9897
MCQ8	.41	.41	.493	.493	0.393	0.2683 A	0.6145
MCQ9	.19	.24	.396	.426	0.665	-0.5259 A	0.3398
MCQ10	.31	.20	.463	.401	0.497	1.5814 C	0.0019
MCQ11	.24	.26	.430	.440	0.522	0.3532 A	0.6147
MCQ12	.00	.00	.000	.000	0.569	NaN?	0.0000

Table 1 shows the findings of the Differential Item Functioning (DIF) analysis, which was used to test gender-based item bias in SSC Chemistry multiple-choice test items. The analysis was based on descriptive statistics (mean and standard deviation), point-biserial correlation coefficients (Rpbis), Mantel-Haenszel delta statistics (ΔMH), and the associated p-values to assess the possibility that individual items operated differently between male and female students despite controlling for the overall ability.

Descriptive statistics show that female and male students fit in the same category on the majority of the test items. The differences in means of the two groups were generally small, indicating that the overall level of difficulty of items was similar between genders. Indicatively, MCQ1, MCQ4, MCQ5, MCQ7, and MCQ8 indicate the same mean scores in both male and female students, indicating equal performance trends. This is further supported by the values of standard deviation, with the dispersion of scores being similar between the genders of the sample population in the majority of the items, which suggests the same variability of responses.

Discrimination of items, as measured by point-biserial correlations, was found to be acceptable in almost all items. Both male and female scores on the values of Rpbi were moderate to high, which

showed that these items were useful in differentiating higher- and lower- ability examinees. This is an indication that, in general, the chemistry test operated sufficiently as a gauge of student achievement between genders. Nonetheless, the MCQ12 item also had a zero-discrimination value for both groups, which means that the item was not able to discriminate between students with different levels of ability.

The Mantel-Haenszel delta (2MH) statistics give direct information about whether or not there is DIF and the extent of the same. According to regular DIF classification criteria, the majority of the items belong to Category A, which means no DIF. Questions like MCQ1, MCQ3, MCQ4, MCQ5, MCQ7, MCQ8, MCQ9, and MCQ11 had non-significant p-values (p >.05) and low values of 3MH, meaning they acted similarly to both male and female students. These results suggest that most SSC chemistry test items were non-gender-biased and did not favor either of the groups.

However, there were two questions MCQ2 and MCQ10 with statistically significant DIF. MCQ2 had a significant p-value (p =.0051) and, therefore, by definition, this constitutes a Category B DIF. Female students received a higher mean score on this item than male students, which means moderate DIF in favor of females. In the same manner, MCQ10 was reported to have a ΔMH of 1.5814 with a very strong p-value (p =.0019), indicating it is in Category C, which indicates large DIF. The difference in the means that was observed indicates that female students had a significant upper hand on this item compared to their male colleagues.

MCQ6 had a relatively higher negative ΔMH value; that is, there was a tendency of male advantage, but the p-value was not statistically significant. As a result, this item was categorized as Category A and cannot give adequate evidence of meaningful DIF, but can be considered as a suggestion of a qualitative inspection during item review.

MCQ12 is also a problematic item, with zero mean scores and zero discrimination indices obtained by both male and female students. The undefined DeltaMH statistic and significant P value indicate that the item was not attempted, constructed incorrectly, or that the item is always incorrectly constructed. In such a way, MCQ12 cannot be used in the interpretation of DI and should not be used in future tests or be modified significantly.

Overall, the DIF analysis shows that the SSC Chemistry examination is fair primarily in terms of gender because the majority of items showed insignificant DIF. Nevertheless, moderate to high DIF in MCQ2 and MCQ10 suggests that there can be an item bias, which could be due to content presentation, cognitive load, or acquaintance with context. These things need to be reviewed further qualitatively to determine the causes of gender-related disparate performance. The solution of these problems is critical in improving psychometric standards and equity of SSC chemistry testing in Pakistan.

Question B: Is there any Chemistry test item used in *the Board of Intermediate and Secondary Education that shows* differential item functioning among regular and private students?

Table 2: DIF Analysis among the Regular and Private Students

Item No.	Mean		S.D.		Rpbis	DeltaMH	P-value
	Regular students	Private students	Regular students	Private students			
MCQ1	.43	.47	.495	.502	0.603	-0.6913 A	0.3455
MCQ2	.44	.47	.497	.502	0.362	-0.5540 A	0.5166
MCQ3	.22	.15	.414	.355	0.391	1.3100 B	0.1055
MCQ4	.40	.33	.491	.471	0.487	0.8593 A	0.1781

MCQ5	.54	.54	.499	.501	0.559	0.0546 A	0.9374
MCQ6	.42	.30	.494	.462	0.676	1.2586 B	0.0443
MCQ7	.32	.34	.466	.475	0.625	0.3283 A	0.6741
MCQ8	.41	.43	.492	.497	0.393	0.3600 A	0.6941
MCQ9	.22	.19	.414	.395	0.665	0.5362 A	0.5315
MCQ10	.25	.33	.431	.471	0.497	1.0361 B	0.1488
MCQ11	.25	.24	.436	.427	0.522	0.2003 A	0.8850
MCQ12	.00	.00	.495	.000	0.569	NaN?	0.0000

The Differential Item Functioning (DIF) analysis to compare the performance of regular and private students on the SSC Chemistry multiple-choice items is shown in Table 2. Statistical tests include the use of descriptive statistics (mean and standard deviation), point-biserial correlation coefficients (Rpbis), Mantel-Haenszel delta statistics (ΔMH), and p-values to determine whether individual items worked equally well with both categories of students when independent variables (overall ability) were controlled.

The average results indicate that regular and private students have shown overall similarity in performance on most items, with a slight difference in observed difficulty. The mean value of some of the items, e.g., MCQ5, MCQ7, MCQ8, and MCQ11, is pretty similar, which means that there are similar patterns of responses between the two groups. The values of standard deviation also tend to be similar, which indicates the same level of variability among regular and private candidates.

Most of the point-biserial correlation coefficients between the two groups are within an acceptable range, and hence, the item is well discriminated. The argumentation gives the impression that most of the items were successful in separating students with better and lower ability, regardless of whether they were enrolled or not. However, again, MCQ12 has an anomalous pattern, where the mean scores are zero, and the levels of discrimination are negligible or zero, indicating an ineffective item.

The Mantel-Haenzel delta (ΔMH) statistics show that most of the items lie under the category A, showing no significant DIF between the regular and the private students. Questions like MCQ1, MCQ2, MCQ4, MCQ5, MCQ7, MCQ8, MCQ9, and MCQ11 show a low value of 0MH and a non-significant p-value (p >.05), and this indicates that these questions work identically with the different categories of student enrollments.

However, some evidence of DIF is seen in only a few items. MCQ6 has ΔMH value of 1.2586 and p-value is statistically significant (p =.0443), which means that it lies in the category of B DIF, implying moderate DIF but in favor of regular students, as shown by the higher mean score (.42), than the mean score of private students (.30). MCQ3 and MCQ10 also indicate ΔMH with values greater than the threshold of Category B, but the p-values were not significant which can be viewed as indicative of unstable or sample-dependent DIFs.

MCQ12 once more becomes a problematic item, and the MEAN scores of both groups are zero, and the value of 8(ΔMH) is undefined. This trend indicates that the item was either keyed wrongly, or it was answered with high accuracy, or not responded to by the test takers. In this way, MCQ12 cannot be used to interpret DIF, and it is not to be continued in the psychometric analysis.

The findings, in general, indicate that the SSC Chemistry examination is fairly so in terms of regular and private student status, since not a lot of items are showing any significant level of DIF. However, the moderate strength of DIF in MCQ6 suggests that some of the items might be favorable to students

whose school-based education is more formal, which might be due to variations in curriculum coverage or instructional resources, and laboratory exposure. These results indicate why DIF analysis should be incorporated in regular examination review procedures so that there is fairness in the evaluation of regular and private candidates in the Pakistan SSC system.

Question C: Is there any Chemistry test item used in *the Board of Intermediate and Secondary Education that shows* differential item functioning for students' institution type?

Table 3: DIF Analysis among the Public and Private Institutes

Item No	Mean		S.D.		Rpbis	DeltaMH	P-value
	Public school	Private school	Public school	Private school			
MCQ1	.44	.42	.497	.495	0.603	0.0792 A	0.9837
MCQ2	.45	.44	.498	.498	0.362	-0.0210 A	0.9363
MCQ3	.20	.22	.401	.417	0.391	-0.2516 A	0.7365
MCQ4	.39	.40	.488	.490	0.487	-0.0414 A	0.9900
MCQ5	.53	.55	.500	.499	0.559	-0.4474 A	0.4891
MCQ6	.39	.41	.489	.494	0.676	-0.1744 A	0.7861
MCQ7	.33	.30	.471	.458	0.625	0.5255 A	0.3475
MCQ8	.42	.38	.495	.488	0.393	0.5687 A	0.3986
MCQ9	.21	.23	.404	.425	0.665	-0.4300 A	0.4955
MCQ10	.25	.27	.436	.446	0.497	-0.4102 A	0.5162
MCQ11	.27	.22	.443	.413	0.522	0.7757 A	0.2514
MCQ12	.00	.00	.000	.000	0.569	NaN?	0.0000

The findings were represented in Table 3, which shows the Differential Item Functioning (DIF) analysis of the performance of students in public schools and those in private schools in SSC Chemistry multiple-choice items. Descriptive statistics (mean and standard deviation), point-biserial correlation coefficients (Rpbis), Mantel-Haenszel delta statistics (ΔMH), and p-values are used to identify whether individual test items performed in the same way across school sectors, once all tests were controlled by overall ability.

The average scores suggest that the performance of the students of both the public and the private schools was quite similar, with almost all the items. The difference in the mean values was also not very significant, which implied that the item difficulty levels were very similar across the two groups. As an illustration, MCQ1, MCQ2, MCQ4, MCQ5, and MCQ6 of the questionnaire exhibit a strongly corresponding mean score between the public and private school students. The values of standard deviation are also similar, and this means that there were similar variations in the responses of students in school sectors.

Correlation coefficients in point-biserial have an acceptable level of item discrimination in both groups in most items. The P-values of Rpbis are mostly in their moderate range, which implies that the items were useful in separating higher and lower ability students across school types. As it has been noted in the previous analyses, MCQ12 is zero-discriminative in both groups, which means that it did not add any useful information to assessing the ability of students.

The Mantel-Haenszel delta (ΔMH) statistic indicates that all items except MCQ12 are in Category A; this implies that the DIF is insignificant between the students of public and private schools. The ΔMH values are low in value with no significant p-values ($p > .05$), which implies that none of the items depict significant differences in functioning in the school sectors. These findings show that there is no difference between the chemistry test items used with students studying in private and in public institutions.

MCQ12 is also a problematic item where the mean score is zero, the discrimination index is zero, and the 0.25 value is not defined. This trend indicates that this item was misconstructed, all respondents responded erroneously, or did not attempt it, by the people taking the test. In this way, MCQ12 cannot be used to interpret DIF and cannot be further analyzed psychometrically.

With the overall results, it is seen that the SSC Chemistry examination shows great fairness when applied to both a public and a private school setting, and no substantive DIF is identified in the studied items. This implies that institutional type differences did not matter much in the performance of the items after the ability of the students had been adjusted. These findings can qualify the conclusion that the SSC Chemistry test is a fair measure of student achievement in school sectors in the secondary education system in Pakistan.

Discussions

The results of the Differential Item Functioning test show that the SSC Chemistry test is rather fair among the genders since most of the items showed insignificant DIF, implying that both males and females with similar levels of abilities were evaluated equally. This general trend is consistent with studies conducted internationally that well-designed science tests can be used equally in both genders when they are content-represented and cognitively challenged (Liu & Wilson, 2023). Nevertheless, the moderate and large DIF on a small number of items indicates the relevance of item analysis with high-stakes tests. The same tendencies have been seen in the international and local arena, whereby a smaller set of items performed differentially in spite of the overall fairness of the tests (OECD, 2023). The items that favor female students in the current research could be viewed as the differences in the focus of instruction, familiarity of contexts, or cognitive processing strategies related to chemistry learning, in line with other previous studies on science education research (Taber, 2021). The fact that a non-functional item has been identified further supports the importance of the systematic post-examining review since the low-performing items may compromise the score validity and blur the actual differences in ability (Wainer, 2012, and Chan, 2022). Taken together, these results highlight the importance of considering the use of DIF analysis as a part of the common examination review procedure within the Pakistani system of SSC in order to increase the validity of tests, decrease the occurrence of unintended bias, and promote the fairness of assessment results among different groups of students.

Results of the Differential Item Functioning analysis found that the SSC Chemistry examination tends to be fairly fair to both regular and private students since the majority of the items were found to have negligible DIF, which implied that the assessment of students with similar ability levels was conducted equally, irrespective of whether they are regular or private students. The research results are aligned with the international literature, which highlights that standardized tests have the potential to support a sense of fairness in a diverse group of candidates in cases where the items are properly correlated with the curricular goals and targeted cognitive skills (Zumbo, 2019). However, the moderate DIF in a few items, especially MCQ6, indicates that some chemistry items can favor typical students who are exposed to organized classroom teaching, laboratory studies, and tutored

examination preparation, over private exam takers, many of whom have a tendency of learning by themselves or through informal coaching arrangements. The same trends have also been observed in the local SSC-based DIF research, in which enrollment status and instructional context were found to have a contribution to the variation in item performance. However, the overall test fairness was observed (Islam et al., 2025). Also, the presence of non-functional items demonstrates the necessity of systematic post-examination item review because non-functional items may suggest the effect of poor performance on the score validity as well as the distortion of actual differences in achievement (Holland & Wainer, 2012). All of these results point to the necessity to include DIF analysis in the process of SSC examination review in Pakistan to improve the validity of the tests, minimize unwanted bias, and provide fair results of the assessment to regular and private students.

The outcomes of the Differentiating test will show that the SSC Chemistry test operates fairly between the public and private school settings because all the items, except one that was not a functioning item, exhibited insignificant DIF. This result implies that students in the public and private institutions had similar chances of getting the individual test items right after the ability had been adjusted, which indicates that there is fairness and validity of the assessment in both sectors of the school. These results are consistent with the findings of international research indicating that measurement invariance through standardized research is possible with instructional goals and content coverage being similar across the institutional type (Zumbo & Chan, 2022). The fact that there were no significant differences in DIF also points to the fact that the differences in the instructional materials, teacher credentials, or even learning conditions between the two schools (public and private) did not affect the performance of items in this test in a systematic way. The same has been indicated in local SSC-based studies where comparisons between public and private schools showed a negligible amount of item bias in spite of the wide-ranging structural differences within the education system (Islam et al., 2025). Nevertheless, the fact that a non-functional item of a test has been present highlights the need for regular post-examination review of items because poorly functioning items may invalidate a test and conceal the actual differences in achievement (Holland & Wainer, 2012). All in all, the results support the idea that the use of DIF analysis should be one of the parts of the examination quality assurance practice in Pakistan that will guarantee the consideration of the fair outcomes of the assessment in different institutions.

References

Ahmed, S., & Khan, M. (2019). *Gender-based differential item functioning in secondary school mathematics assessments in Pakistan: A case study of the 2018 SSC Examination*. Journal of Educational Assessment, 27(2), 45–61.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Press.

Basman, M., & Kutlu, O. (2020). *Identification of Differential Item Functioning on mathematics achievement according to the interactions of gender and affective characteristics by the Rasch Tree Method. International Journal of Progressive Education, 16*(2), 205–217.

Chen, L., Zhang, S., & Liu, J. (2025). Reducing differential item functioning via process data. *Journal of Educational Measurement*, 62(1), 45–67.

Eshun, P. (2025). Assessing differential item functioning in core educational courses: A psychometric analysis. *Scimundi Journal*.

Geary, D. C. (2015). *Mathematical disabilities: Cognitive, neuropsychological, and genetic components*. Elsevier.

Gierl, M. J., & Lai, H. (2018). *Examining differential item functioning in mathematics tests: Gender and content bias*. Journal of Educational Measurement, 55(2), 1-17. https://doi.org/10.1111/jedm.12167
Hambleton, R. K., & Rogers, H. J. (2015). *Differential item functioning: A primer for assessing fairness in educational testing*. Springer.
Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.
Huang, T.-W. (2022). Examination of gender-related differential item functioning in mathematics achievement tests. *Frontiers in Psychology.*
Islam, M. U., Naveed, H. M. S., & Amjad, A. (2025). A differential item functioning analysis of mathematics test items in the 2018 Secondary School Certificate Examination. *Journal of Current Studies, 3*(1), 1–18.
Kartianom, K. (2024). Assessing the fairness of mathematical literacy tests: Evidence from gender-based DIF approaches. *International Journal of Educational Research.*
Linn, R. L. (2015). *Assessments and accountability: The role of differential item functioning in evaluating fairness in educational assessments*. Educational Assessment, 20(2), 1-13. https://doi.org/10.1080/10627197.2015.1070773
Liu, Y., & Wilson, M. (2023). Investigating gender-related differential item functioning in science achievement tests using item response theory. *Educational Measurement: Issues and Practice, 42*(2), 34–45. https://doi.org/10.1111/emip.12506
Lyu, M., & Chen, L. (2020). *Differential item functioning in high school mathematics assessments: A gender-based analysis*. Educational Assessment, 25(3), 234–249. https://doi.org/10.1080/10627197.2020.1776451
Martinková, P. (2017). Checking equity: Why differential item functioning matters. *Journal of Educational Measurement.*
OECD. (2023). *Education at a glance 2023: OECD indicators*. OECD Publishing. https://doi.org/10.1787/69096873-en
Opesemowo, O. A. G. (2025). Exploring differential item functioning in high-stakes assessments among demographic variables. *ScienceDirect.*
Penfield, R. D. (2017). *Differential item functioning in educational testing: An overview*. Journal of Educational Measurement, 54(4), 390–412. https://doi.org/10.1111/jedm.12134
Shah, F., & Shah, Z. (2021). *Investigating school-based differential item functioning in the SSC mathematics examinations of Pakistan*. Pakistan Journal of Educational Research, 33(1), 12–28.
Structural inequality in education. (2025). *Educational Equity Studies.*
Taber, K. S. (2021). *Progressing science education: Constructing the scientific research programme into the contingent nature of learning science*. Springer.
Zumbo, B. D. (2019). *Understanding and using test scores: Measurement and assessment in education*. Routledge.
Zumbo, B. D., & Chan, E. K. H. (2022). Validity and fairness in educational assessment: Addressing bias through differential item functioning. *Assessment in Education: Principles, Policy & Practice, 29*(3), 315–332.