Understanding How Test Questions Measure Student Performance

Chang Lu, Qiuyi Feng, Beiming Yu
Teaching Fellow: Vindyani Herath
Faculty Supervisor: Masanao Yajima

Abstract

This report presents the results of statistical tests conducted to analyze item discrimination, differential item functioning, and the relationship between number fo flaws, item difficulty and point-biserial correlation in a sample of quiz items. We examined the suitability of parametric (T-tests) and non-parametric (Mann-Whitney U tests) statistical methods based on normality checks. The study's primary goal is to determine whether certain demographic factors or quiz characteristics significantly influence test performance.

Through a combination of correlation analysis and significance testing, we evaluate the effectiveness of quiz items in distinguishing between high- and low-performing students. Our results contribute to a better understanding of assessment fairness and item validity.

1 Introduction

Examinations are an effective way to evaluate whether students have achieved the intended learning outcomes in higher education. A well-designed test provides a fair measure of students' knowledge and skills while allowing teachers to make accurate judgments about their progress. In professional schools, assessments are often conducted through multiple-choice questions (MCQs), which offer a time-efficient grading process. However, poorly written questions can negatively affect students' performance, making it essential to ensure the quality and clarity of exam items.

Technical flaws in MCQs can either create unnecessary difficulty or give an unfair advantage to test-wise examinees. According to the National Board of Medical Examiners (NBME) guidelines, common flaws that add irrelevant difficulty include complex question stems with negative phrasing, unclear or overly wordy answer choices, inconsistent presentation of numerical data, nonparallel answer options, and the use of "none of the above." Ensuring that MCQs are clearly written and well-structured helps maintain fairness and accurately assess students' knowledge.

The primary goal of this report is to examine how the number of flaws, item difficulty, and point-biserial correlation impact a student's test performance. Additionally, it aims to assess whether demographic factors play a role in test performance.

2 Experiment Design

A total of 32 first-year dental students and graduate students in the medical science program at Boston University Chobanian & Avedisian School of Medicine completed a 20-question physiology assessment and demographic survey. Each question was allotted a total of two minutes, and for every incorrect response, students were asked a follow-up question to explain why they believed they answered incorrectly. An evaluation instrument based on the National Board of Medical Examiners (NBME) item writing guide was employed to identify the item writing flaws in each question to investigate the influence of technical item flaws on test outcomes across two schools of medicine.

3 Methodology

3.1 Selection of Sample Size

One student did not complete the demographic survey; therefore, complete case analysis will be used, and the analysis will be based on 31 observations. Complete case analysis ensures consistency by including only data points with complete information, reducing bias in correlation estimates, and simplifying interpretation. Given the minimal missing data and the potential application of multivariate analysis, this approach is appropriate for maintaining overall generalizability.

3.2 Different item analysis measures

Item analysis in a test refers to the process of examining individual questions (items) on an exam to evaluate their quality, difficulty level, and how well they differentiate between high and low-performing test-takers, ultimately helping to improve the overall quality of the test by identifying and addressing poorly performing questions; it involves analyzing student responses to each item to assess if they are effectively measuring the intended knowledge or skill.

This study uses four such measures:

1. **Number of flaws:** Each item was analyzed for the presence of flaws using a rubric. The total number of flaws for each question was then recorded.
2. **Item difficulty:** Percentage of students who choose an item correctly.
3. **Point biserial correlation:** A statistical measure that assesses how well a single test question (correct or incorrect) discriminates between students who perform well overall on the test and those who perform poorly. Measures the correlation between the overall test score and a student's answer to a single test question
4. **Item discrimination index:** A measure of how well an item differentiates between high-performing and low-performing test-takers.

The Spearman correlation between each of these measures is shown in Section (4). Then T-tests and Mann-Whitney U-tests were conducted to compare the differences between various demographic groups and the total test score.

3.3 Inter-Rater Reliability (IRR) Assessment

To evaluate the consistency of ratings assigned by multiple reviewers, we computed Inter-Rater Reliability using Light's Kappa, a generalization of Cohen's Kappa suitable for more than two raters. Four raters (Latchman, KVA, LY, and Wagner) independently scored responses across 12 rubric criteria and 30 questions. We filtered out incomplete entries to ensure valid comparisons and calculated Light's Kappa both by criterion and by individual question. This method captures agreement beyond chance, with Kappa values interpreted as follows: values ≥ 0.75 suggest excellent agreement, 0.4 to 0.75 indicate moderate agreement, and values below 0.4 imply poor or inconsistent agreement.

4 Results

Table (1) shows item analysis measures for each test question.

Table 1: Item analysis measures

Item	Number of flaws	Item difficulty	Point bi-serial correlation	Item discrimination index
Q1	0	0.548	0.448	0.575
Q2	2	0.323	0.439	0.400
Q3	1	0.290	-0.236	-0.200
Q4	1	0.355	0.261	0.275
Q5	0	0.419	0.307	0.475
Q6	0	0.710	0.491	0.500
Q7	2	0.323	0.706	0.850
Q8	1	0.484	0.339	0.475
Q9	0	0.484	0.464	0.575
Q10	0	0.258	-0.163	-0.213
Q11	1	0.161	0.267	0.225
Q12	1	0.484	0.250	0.263
Q13	1	0.194	0.361	0.375
Q14	2	0.581	0.614	0.763
Q15	3	0.484	0.589	0.775
Q16	3	0.613	0.604	0.700
Q17	1	0.645	0.298	0.325
Q18	5	0.419	0.614	0.825
Q19	3	0.581	0.542	0.638
Q20	5	0.645	0.354	0.525

A few questions had negative point bi-serial correlation values, suggesting that high-performing students tended to get those items wrong more often than lower-performing students, which may indicate problematic questions.

4.1 Relation to Point-Biserial Correlation

Both Point-Biserial Correlation and Item Discrimination Index (DI) measure how well an item differentiates high-scoring and low-scoring test-takers. However, they are calculated differently and provide unique insights into item performance.

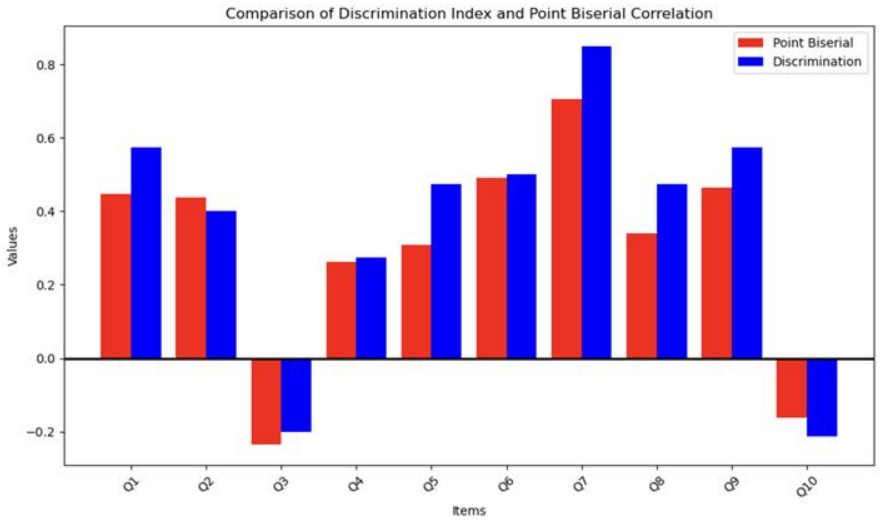

Figure 1: Point Biserial Correlation and Item Discrimination Comparison

Figure 1 compares two item-level metrics—Point-Biserial Correlation and Item Discrimination Index—for each quiz question. Both metrics assess how well an item differentiates between high- and low-performing students. The plot shows that most items have positive values for both metrics, suggesting they effectively discriminate between students.

A few questions (notably Q3 and Q10) exhibit negative values for both measures, meaning high-performing students were more likely to get them wrong—an indicator of potentially flawed or confusing items. Items like Q7 and Q6, on the other hand, show high positive values for both metrics, indicating that they strongly differentiate student performance and are likely well-constructed.

Overall, the alignment of the two metrics reinforces their consistency in evaluating item effectiveness.

4.2 Mann-Whitney U Test for Test Flaws

A Mann-Whitney U test comparison was done to check the distribution of item difficulty and point biserial correlation across all tested flaw categories. For item difficulty, we test the following hypothesis:

Null Hypothesis(H_0) : The distribution of item difficulty is the same across categories of each flaw type
Alternative Hypothesis(H_1) : The distribution of item difficulty differs between the flaw categories

Table 2: Item difficulty: Mann-Whitney U test

Flaw	p value	Decision
Flaw 1	0.489	Retain the null hypothesis
Flaw 2	0.931	Retain the null hypothesis
Flaw 3	0.962	Retain the null hypothesis
Flaw 4	0.081	Retain the null hypothesis
Flaw 5	NA	Unable to compute
Flaw 6	NA	Unable to compute
Flaw 7	NA	Unable to compute
Flaw 8	0.282	Retain the null hypothesis
Flaw 9	0.202	Retain the null hypothesis
Flaw 10	0.229	Retain the null hypothesis
Flaw 11	NA	Unable to compute
Flaw 12	0.537	Retain the null hypothesis

As shown in Table (2), for each flaw type, there was no evidence that the item difficulty differs significantly between flawed and non-flawed items at 5% significance level.

For point biserial correlation, we test the following hypothesis:

Null Hypothesis(H_0) : The distribution of point-biserial correlation is the same across categories of each flaw type
Alternative Hypothesis(H_1) : The distribution of point biserial correlation differs between the flaw categories

Table 3: Point-biserial correlation: Mann-Whitney U test

Flaw	p value	Decision
Flaw 1	0.751	Retain the null hypothesis
Flaw 2	0.386	Retain the null hypothesis
Flaw 3	0.321	Retain the null hypothesis
Flaw 4	0.049	Reject null hypothesis
Flaw 5	NA	Unable to compute
Flaw 6	NA	Unable to compute
Flaw 7	NA	Unable to compute
Flaw 8	0.235	Retain the null hypothesis
Flaw 9	0.204	Retain the null hypothesis
Flaw 10	0.231	Retain the null hypothesis
Flaw 11	NA	Unable to compute
Flaw 12	0.741	Retain the null hypothesis

As shown in Table (3), for all tested categories except Flaw 4, there was no statistical evidence that the distribution of point-biserial correlation differs between flawed and non-flawed items.

4.3 Differential Item Functioning Analysis

Differential item functioning (DIF) analysis evaluates whether individuals from different groups, such as males vs. females or native vs. non-native English speakers, respond differently to specific test items, after controlling for overall ability. This helps detect potential bias in the items.

To conduct DIF analysis, we applied the Mantel-Haenszel (MH) method. This method stratifies test-takers into ability levels based on total test performance and compares the odds of correctly answering each item between the reference and focal groups. The aggregated odds ratios across levels reflect whether one group consistently has an advantage on an item.

For each item and group comparison, we computed the MH effect size. A larger MH effect size indicates a larger differential functioning of the item. However, our results showed that none of the questions exhibited statistically significant DIF in any of the three demographic groups analyzed.

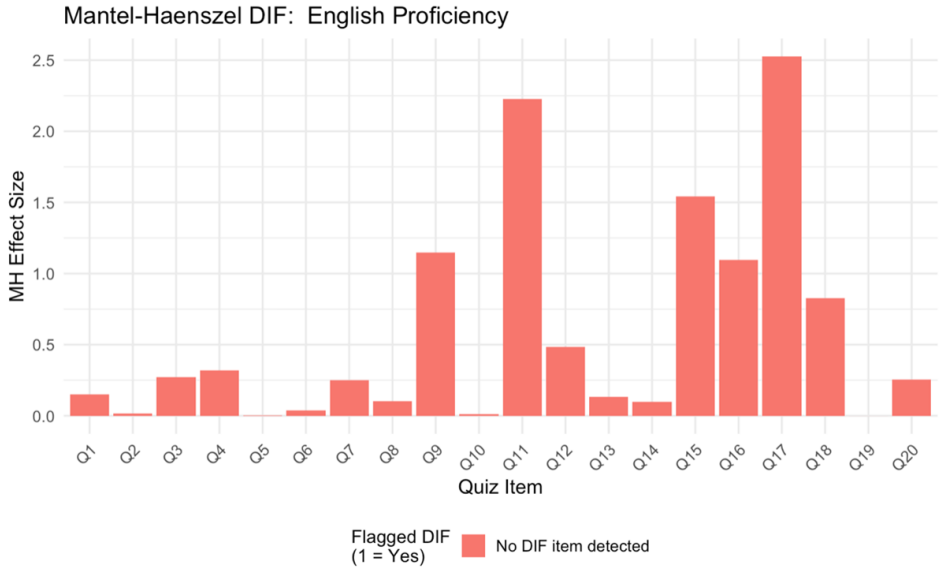

Figure 2: Mantel-Haenszel DIF Effect Sizes by Quiz Item — English Proficiency

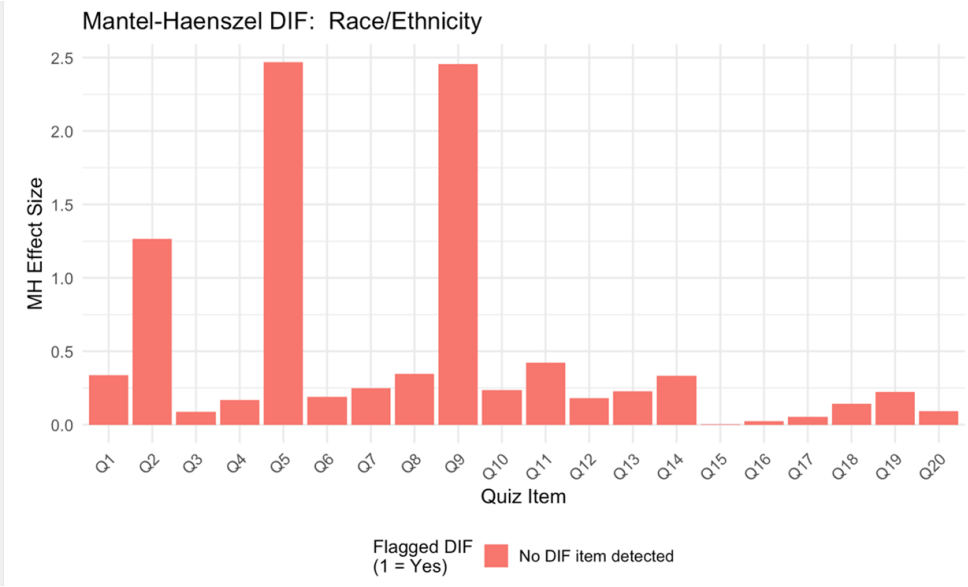

Figure 3: Mantel-Haenszel DIF Effect Sizes by Quiz Item — Race/Ethnicity

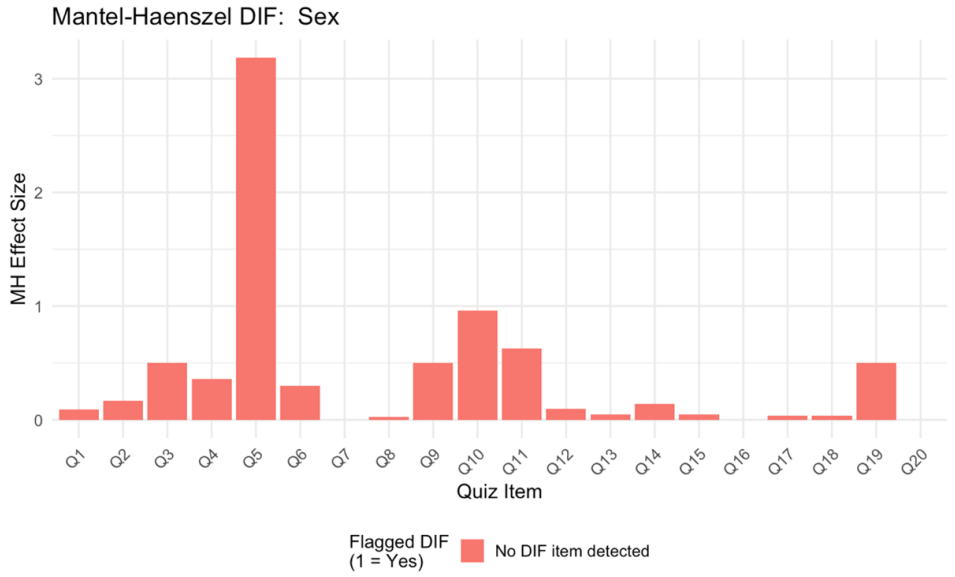

Figure 4: Mantel-Haenszel DIF Effect Sizes by Quiz Item — Sex

Each of the above figures illustrates the Mantel-Haenszel (MH) DIF effect sizes for 20 quiz items across three demographic groups: English proficiency, race/ethnicity, and sex.

For all three groups, the majority of items exhibited low MH effect sizes, indicating minimal differential functioning. A few items (e.g., Q11 and Q17 for English proficiency, Q5 and Q9 for race, and Q5 for sex) had relatively higher bars. These spikes suggest that test-takers in different groups had differing odds of getting the item correct, despite similar overall performance levels.

However, no item showed statistically significant DIF according to standard thresholds. This means the apparent differences are not large enough to conclude that the item is biased against any particular group. Overall, these findings support the fairness of the assessment, as items appear to function consistently across demographic groups.

4.4 Correlation Coefficient Analysis

To examine the relationship between each measure of item analysis, correlation coefficient was calculated. We compared Pearson correlation, Spearman's Rank Correlation and Kendall's Tau Correlation. Each measure provides different insights into the relationship.

- **Spearman's Rank Correlation:** Measures the strength and direction of a monotonic relationship between two variables.
- **Pearson Correlation:** Measures the strength and direction of a linear relationship, assuming normality in both variables.
- **Kendall's Tau Correlation:** Non-parametric measure, robust to small sample sizes and tied ranks.

The Shapiro-Wilk normality test was conducted to assess the normality assumption for the item analysis measures (see Appendix). Since some measures were not normally distributed, Spearman's correlation was used to evaluate the associations between variables.

	Total flaws	Item difficulty	Point bi-serial	Item discrimination
1. Total flaws		0.173	0.493*	0.523*
2. Item difficulty			0.361	0.426
3. Point-biserial correlation				0.953*
4. Item discrimination				

* Correlation is significant at the 0.05 level

Table 4: Spearman's correlation coefficient between item analysis measures

The interpretation of Table 7 is as follows:

- **Total flaws and point biserial correlation** ($\rho = 0.493, p < 0.05$): A moderate positive correlation suggests that as the number of flaws in an item increases, its point biserial correlation tends to increase as well. This indicates that flawed items may still differentiate between high- and low-performing students, though the relationship is not strong.
- **Total flaws and item discrimination** ($\rho = 0.523, p < 0.05$): A moderate-to-strong positive correlation suggests that an increase in total item flaws is associated with higher item discrimination. This might indicate that some flawed items still perform well in distinguishing high- and low-scoring students, though further investigation is needed to understand whether this trend is due to specific types of flaws.
- **Item difficulty and point biserial correlation** ($\rho = 0.361$): A moderate positive correlation suggests that as item difficulty increases (i.e., as questions become harder), the ability of the question to differentiate high- and low-performing students also increases slightly. However, this relationship is not strong, meaning that factors beyond difficulty contribute to item discrimination.
- **Item difficulty and item discrimination** ($\rho = 0.426$): A moderate positive correlation suggests that more difficult items tend to have higher discrimination values, meaning they better distinguish between high- and low-performing students. This aligns with the expectation that medium-difficulty items often provide better discrimination.
- **Point biserial correlation and item discrimination** ($\rho = 0.953, p < 0.05$): A very strong positive correlation suggests that these two measures are almost identical in how they assess an item's ability to distinguish between high- and low-scoring students. This is expected since both metrics are designed to evaluate an item's effectiveness in differentiating student performance.

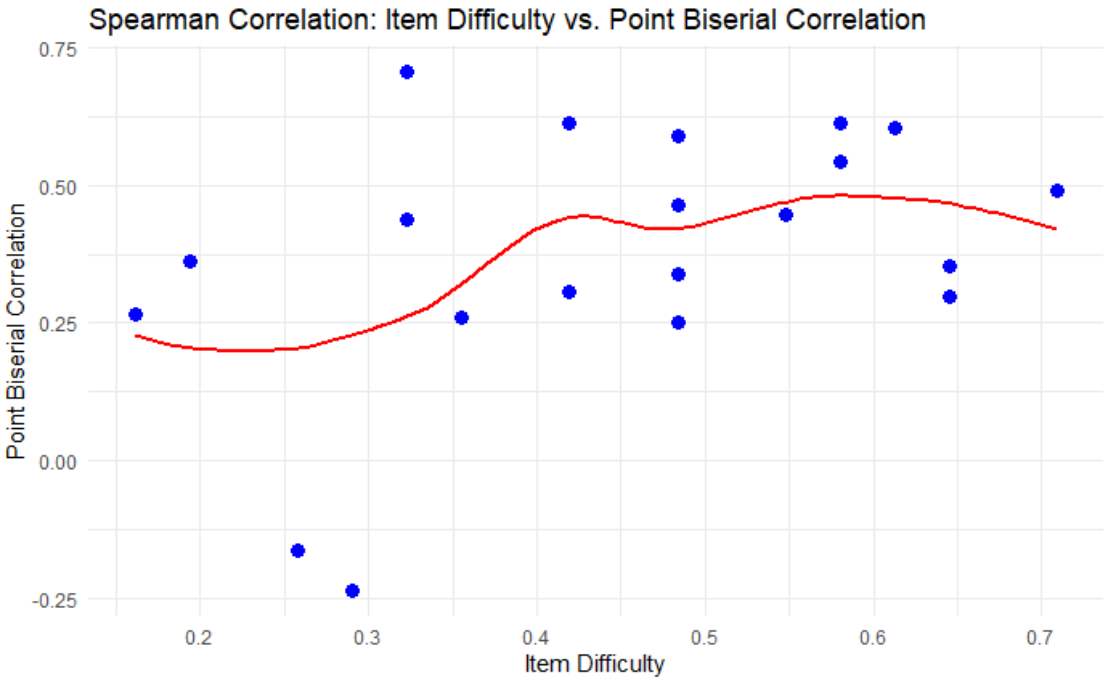

Figure 5: Relationship between Item Difficulty and Point Biserial Correlation

Figure (5) shows how item difficulty (x-axis) relates to point biserial correlation (y-axis), which measures how well a question distinguishes between high- and low-performing students. Since the relationship is not strictly linear, Spearman's correlation provides the most appropriate measure of association.

4.4.1 Summary of Findings

- Items with more flaws tend to have slightly better discrimination power and higher point biserial correlations, though the strength of these relationships is moderate.
- Item difficulty is positively associated with both point biserial correlation and item discrimination, meaning that harder items generally differentiate students better, but the effect is not very strong.
- The very strong correlation between point biserial correlation and item discrimination confirms that both measures capture similar aspects of item performance.

4.5 Test Scores and Demographic Group Analysis

As the next step in the analysis, the client was interested in examining differences in exam performance across demographic groups. To assess this, we test the following hypothesis:

$$\text{Null Hypothesis}(H_0) : \text{There is no significant difference between the means of the two groups}$$
$$\text{Alternative Hypothesis}(H_1) : \text{There is a significant difference between the mean values of the two groups}$$

To compare exam performance across demographic groups, we will first check whether the assumptions for a t-test are met. The assumptions for conducting an independent t-test include:

- Normality: The exam scores within each demographic group should be approximately normally distributed. This will be assessed using the Shapiro-Wilk test (or Kolmogorov-Smirnov (KS) test or Lilliefors test) and visual inspections such as histograms and Q-Q plots.
- Homogeneity of Variance: The variance of exam scores should be equal across groups. This will be tested using Levene's test.

- Independence: The observations (exam scores) must be independent within and between groups.

If these assumptions are satisfied, we will use an independent t-test to compare the means of the groups. However, if the normality assumption is violated, we will use the Mann-Whitney U test, a non-parametric alternative that does not assume normality. The following figures illustrate the normality test results for Total Score across various demographic and quiz-related factors.

- **Shapiro-Wilk Test Results**:
  - **Sex**: $p < 0.05$ (non-normal)

Sex <dbl>	Shapiro_Statistic <dbl>	P_Value <dbl>
1	0.9033553	0.0296610
0	0.8413044	0.0776804

Figure 6: Results for Total Score by sex

  - **Race/Ethnicity**: $p > 0.05$ (normal)

Race/Ethnicity <dbl>	Shapiro_Statistic <dbl>	P_Value <dbl>
1	0.8984191	0.0539308
0	0.9320853	0.3628041

Figure 7: Results for Total Score by Race/Ethnicity

  - **English Proficiency**: $p > 0.05$ (normal)

English Proficiency <dbl>	Shapiro_Statistic <dbl>	P_Value <dbl>
0	0.9102287	0.08683599
1	0.9054876	0.15904845

Figure 8: Results for Total Score by english proficiency

  - **Born in USA**: $p < 0.05$ (non-normal)

Born USA <dbl>	Shapiro_Statistic <dbl>	P_Value <dbl>
0	0.9138409	0.04276168
1	0.8991240	0.32570961

Figure 9: Results for Total Score by whether the student was born in USA or not

  - **Home Language**: $p > 0.05$ (normal)

Home Language <dbl>	Shapiro_Statistic <dbl>	P_Value <dbl>
0	0.9069693	0.06514482
1	0.9192120	0.27945384

Figure 10: Results for Total Score by home language

- **Kolmogorov-Smirnov Test Results**:
  - Accommodations group had sample size issues ($N < 3$ for one category). Hence the Kolmogorov-Smirnov test was used.

Accomodations <dbl>	Sample_Size <int>	KS_Statistic <dbl>	P_Value <dbl>
2	29	0.1447509	0.5778148
1	2	*NA*	*NA*

Figure 11: Results for Total Score by accomodations

- **Lilliefors Test Results**:
- **Quiz Time**: We converted the quiz time data from seconds to minutes and categorized it into three groups based on 10-minute intervals: 0-10, 10-20, 20+ $p > 0.05$ (normal)

Quiz Time Group <fctr>	Sample_Size <int>	Lilliefors_Statistic <dbl>	P_Value <dbl>
0-10 min	7	0.2434260	0.2385870
20+ min	17	0.1664781	0.2399435
10-20 min	7	0.1779499	0.7211381

Figure 12: Results for Total Score by Quiz Time

- **Age Arrive USA**: Age arrive USA group had the same problem as Accommodations group, which indicated Mann-Whitney U test.

Age arrive USA <dbl>	Sample_Size <int>	Lilliefors_Statistic <dbl>	P_Value <dbl>
0	24	0.1631931	0.09782839
2	2	*NA*	*NA*
3	2	*NA*	*NA*
4	2	*NA*	*NA*
5	1	*NA*	*NA*

Figure 13: Age Arrive USA

The results below present the t-test and Mann-Whitney U test outcomes, applied based on the normality check. If the normality assumption was met, a t-test was conducted; otherwise, the Mann-Whitney U test was used to compare exam performance between groups.

A significance threshold of $p < 0.05$ was used to determine significant differences.

4.5.1 T-Test Results

Variable	Test Statistic (t)	p-value
English Proficiency	1.0918	0.2839
Race	-0.0973	0.9232
Home Language	-0.0986	0.9222
Sex	1.4789	0.1499
Born in USA	0.5773	0.5682

4.5.2 Mann-Whitney U-Test Results

Variable	Test Statistic (W)	p-value
Sex	125	0.1390
Born in USA	96.5	0.5675
Accommodations	29	1.0000
English Proficiency	144.5	0.2758
Race	114.5	0.9357
Home Language	112	0.9511

4.6 Quiz Time and Demographic Group Analysis

To investigate whether total quiz time varied across different demographic groups, we initiated a new analysis and performed a combination of normality checks, t-tests, and Mann–Whitney U tests to compare quiz times across groups.

Due to the relatively small sample sizes and normality violations in some groups, we used non-parametric Mann–Whitney U tests where appropriate. In other cases, both t-tests and U-tests were conducted to ensure robustness. A significance threshold of $p < 0.05$ was used.

- **English Proficiency (Fluent vs Native)**: No significant difference in total quiz time ($t = 0.79$, $p = 0.44$; $U = 113$, $p = 0.89$).
- **Race (White vs Non-White)**: No significant difference found ($t = -0.01$, $p = 0.99$; $U = 142$, $p = 0.33$).
- **Born in U.S.**: No significant difference in time between students born in the U.S. and those born elsewhere ($t = -0.40$, $p = 0.69$; $U = 76$, $p = 0.72$).
- **Mother Born in U.S.**: Showed a trend toward significance, but not statistically significant ($t = -0.50$, $p = 0.62$; $U = 72$, $p = 0.092$).
- **Father Born in U.S.**: No significant difference found ($t = -0.30$, $p = 0.77$; $U = 85$, $p = 0.21$).
- **Home Language (English vs Other)**: No statistically significant difference ($t = -0.93$, $p = 0.36$; $U = 76$, $p = 0.13$).
- **Gender (Female vs Male)**: No significant difference ($t = -0.60$, $p = 0.56$; $U = 66$, $p = 0.25$).

Variable	**T-test P-value**	**U-test P-value**
English Proficiency	0.4377	0.8886
Race	0.9885	0.3267
Country (Self)	0.6903	0.7231
Mother's Country	0.6234	0.0924
Father's Country	0.7697	0.2073
Home Language	1.0000	0.2073
Gender	0.5550	0.2497

Table 5: P-values from T-tests and U-tests for total quiz time by demographic group

Across all demographic groups, no statistically significant differences in total quiz time were detected. Although there was a trend toward significance in quiz time based on the mother's country of birth, this did not meet the standard threshold. These results suggest that students spent similar amounts of time on the quiz regardless of background, which further supports fairness in test-taking conditions.

4.7 Inter-Rater Reliability

The goal of this section is to analyze and validate the inter-rater reliability (IRR) of multiple-choice questions to ensure consistent evaluation and fair outcomes. It quantifies the extent of agreement among different raters evaluating the same subjects using a set of criteria. High IRR indicates that the tool is designed so that different raters perceive and score the criteria similarly, minimizing the influence of subjective judgments.

4.7.1 Experiment Design

The rubric consists of 12 binary questions aiming to assess qualities of a multiple question, like clarity, ambiguity, etc. Since the rubrics are designed to assess the quality of multiple-choice exam questions, the rubrics will be assessed by the rater's agreement when raters apply the rubric to the multiple-choice questions. 4 raters were asked to rate all 30 questions using all 12 binary criteria in the rubric on a "question-by-criteria" matrix. Note that not every question-criterion pair is meaningful, so "NA" are also present in the matrix. Light's Kappa was used to implement IRR based on prior work done by a different MSSP group.

4.7.2 Results

```
Light's Kappa for m Raters

Subjects = 270
  Raters = 4
   Kappa = 0.321

       z = 0.0939
 p-value = 0.925
```

Figure 14: Overall Kappa

The overall Light's Kappa across all items and raters was $\kappa = 0.321$, for four raters evaluating 270 subjects. It indicates fair to moderate agreement among the raters. While it shows some level of consistency, it is not particularly strong. The associated z-value of 0.094 and a p-value of 0.925 further indicate that the observed agreement level is not significantly different from chance, suggesting that any agreement among the raters could largely be coincidental and does not demonstrate strong inter-rater reliability.

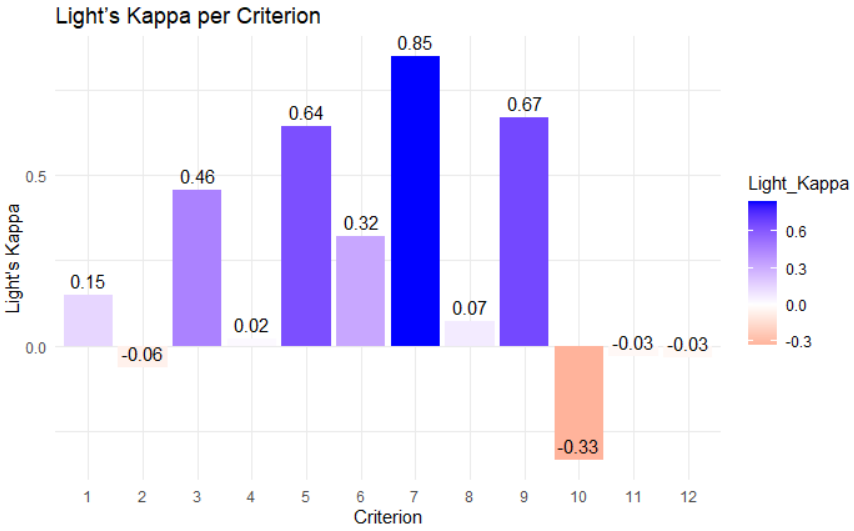

Figure 15: Kappa Per Criterion

By Criterion: Kappa scores ranged widely across criteria. For instance, Criterion 7 exhibited strong agreement ($\kappa = 0.848$), followed by Criterion 5 ($\kappa = 0.644$) and Criterion 9 ($\kappa = 0.670$). Conversely, Criterion 10 showed notably poor agreement with a negative Kappa of -0.333.

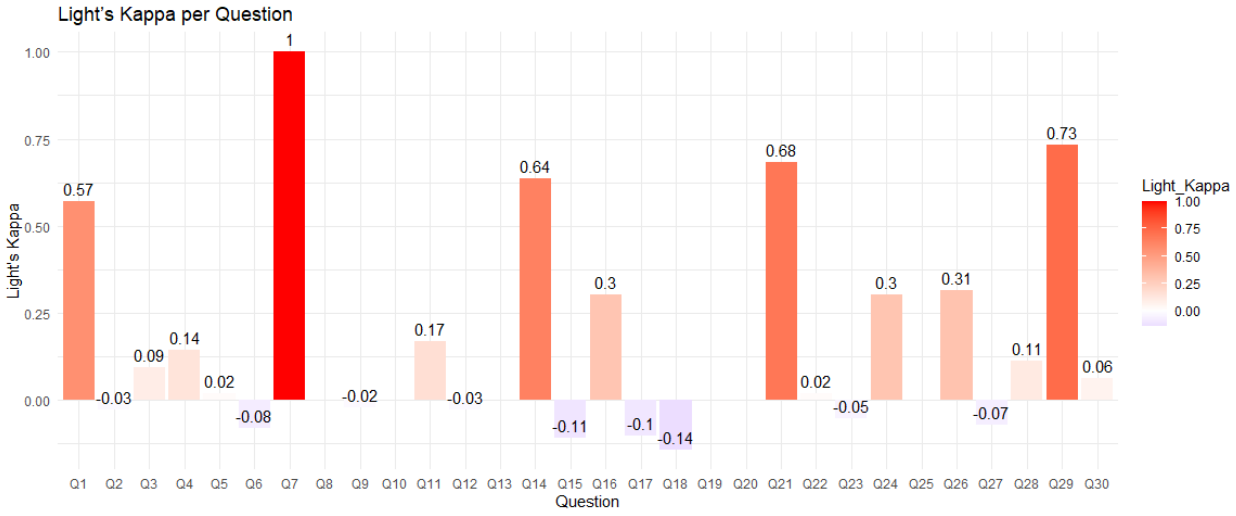

Figure 16: Kappa Per Question

By Question: Inter-rater reliability also varied at the question level. Question 7 achieved perfect agreement ($\kappa = 1.000$), while several others (e.g., Q18: $\kappa = -0.143$; Q15: $\kappa = -0.111$) had negative values, suggesting major inconsistencies among raters. Only a subset of questions (e.g., Q1: $\kappa = 0.570$, Q29: $\kappa = 0.739$) reached the moderate to high agreement threshold. Several NaN values were observed, primarily due to a lack of response variability, an indication, to some extent, of perfect agreement among the raters.

These findings indicate that while some criteria and items were rated reliably, others require clearer definitions or additional rater training to improve consistency.

5 Summary

This study aimed to evaluate the effectiveness of quiz items in distinguishing between high- and low-performing students while assessing whether demographic factors influence test performance. Through a combination of correlation analysis and statistical testing, we identified key patterns in item discrimination, item difficulty, and technical item flaws.

5.1 Key Findings

- **Item Analysis and Correlation:**
  - A moderate positive correlation ($\rho = 0.493, p < 0.05$) was found between the number of flaws and point-biserial correlation, indicating that flawed items may still differentiate student performance.
  - A strong correlation ($\rho = 0.953, p < 0.05$) between point-biserial correlation and item discrimination index suggests that both measures capture similar aspects of item effectiveness.
  - A moderate correlation ($\rho = 0.426$) between item difficulty and item discrimination supports the expectation that moderately difficult questions tend to provide better discrimination.
- **Demographic Analysis of Test Scores:**
  - No statistically significant differences ($p > 0.05$) were found between student performance and demographic factors (sex, race, English proficiency, birthplace, or home language).
  - Normality tests determined the appropriate use of parametric (T-tests) or non-parametric (Mann-Whitney U tests), ensuring valid comparisons.
- **Differential Item Functioning (DIF):**
  - DIF analysis using the Mantel-Haenszel method revealed no statistically significant differential item functioning across groups (sex, race, and English proficiency).
  - Although some items showed slightly elevated effect sizes, the overall results support item-level fairness across demographic groups.
- **Quiz Time and Demographic Analysis:**
  - No statistically significant differences were found in total quiz time between groups based on gender, race, birthplace, home language, or English proficiency.
  - A trend toward significance was observed in quiz time by mother's country of birth ($p = 0.092$), though it did not reach conventional thresholds.
- **Statistical Methods and Test Fairness:**
  - The selection of statistical tests was guided by normality assessments, ensuring the robustness of results.
  - The lack of significant demographic differences in both performance and time-on-task suggests that the assessment process is fair and unbiased.
- **Inter-Rater Reliability (IRR):**
  - The overall Light's Kappa value of $\kappa = 0.321$ indicates fair agreement among the four raters.
  - Agreement varied notably across both rubric criteria and individual questions, with some items (e.g., Criterion 7, Question 7) showing excellent consistency, while others showed negative Kappa values indicating disagreement.
  - Several `NaN` values were observed, resulting from a lack of variability in responses. These reflect instances of perfect agreement that render Kappa mathematically undefined.
  - These findings suggest a need to refine certain rubric items and improve rater training and guidelines to enhance scoring consistency and reliability.

5.2 Conclusion

This study provides evidence that item characteristics, particularly flaws and difficulty, moderately influence test performance, while demographic factors do not significantly affect outcomes. These results reinforce the fairness of the assessment process. However, inter-rater reliability findings indicate variability in scoring consistency, highlighting the importance of refining rubric definitions and enhancing rater calibration.

6 Appendix

6.1 Item Difficulty

Item Difficulty: Computed as the mean proportion of correct responses for each question.

Item <chr>	Item_Difficulty <dbl>
Q1	0.5483871
Q2	0.3225806
Q3	0.2903226
Q4	0.3548387
Q5	0.4193548
Q6	0.7096774
Q7	0.3225806
Q8	0.4838710
Q9	0.4838710
Q10	0.2580645

Figure 17: Q1 to Q10

Item <chr>	Item_Difficulty <dbl>
Q11	0.1612903
Q12	0.4838710
Q13	0.1935484
Q14	0.5806452
Q15	0.4838710
Q16	0.6129032
Q17	0.6451613
Q18	0.4193548
Q19	0.5806452
Q20	0.6451613

Figure 18: Q11 to Q20

6.2 Point Biserial Correlation

Point Biserial Correlation (r_pb): Computed using Pearson correlation between each test item and the total test score.

Item <chr>	Point_Biserial_Correlation <dbl>
Q1	0.4477848
Q2	0.4385671
Q3	-0.2356495
Q4	0.2608162
Q5	0.3070852
Q6	0.4909364
Q7	0.7055209
Q8	0.3388955
Q9	0.4637517
Q10	-0.1629667

Figure 19: Point Biserial Correlation: Q1 to Q10

Item <chr>	Point_Biserial_Correlation <dbl>
Q11	0.2665869
Q12	0.2497125
Q13	0.3609875
Q14	0.6141703
Q15	0.5886080
Q16	0.6038991
Q17	0.2980756
Q18	0.6141703
Q19	0.5419150
Q20	0.3539648

Figure 20: Point Biserial Correlation: Q11 to Q20

```
        Shapiro-Wilk normality test

data:  combined_IDPB$Item_Difficulty
W = 0.96532, p-value = 0.6546


        Shapiro-Wilk normality test

data:  combined_IDPB$Point_Biserial_Correlation
W = 0.87626, p-value = 0.01516
```

Figure 21: PBC is not normally distributed

7 Analysis using new data

Table (6) shows item analysis measures for each test question for 43 students after obtaining new data.

Table 6: Item analysis measures

Item	Number of flaws	Item difficulty	Point bi-serial correlation	Item discrimination index
Q1	0	0.581	0.422	0.450
Q2	2	0.372	0.436	0.483
Q3	1	0.326	-0.049	-0.033
Q4	1	0.419	0.118	0.15
Q5	0	0.535	0.266	0.433
Q6	0	0.698	0.492	0.500
Q7	2	0.349	0.723	0.850
Q8	1	0.488	0.330	0.433
Q9	0	0.442	0.474	0.533
Q10	0	0.326	0.018	0.017
Q11	1	0.163	0.287	0.267
Q12	1	0.419	0.336	0.367
Q13	1	0.140	0.288	0.250
Q14	2	0.605	0.607	0.700
Q15	3	0.535	0.470	0.567
Q16	3	0.605	0.594	0.783
Q17	1	0.558	0.382	0.45
Q18	5	0.395	0.651	0.867
Q19	3	0.535	0.470	0.533
Q20	5	0.558	0.331	0.483

Table 7: Spearman's correlation coefficient between item analysis measures for new data

	Total flaws	Item difficulty	Point bi-serial	Item discrimination
1. Total flaws		0.068	0.439	0.564*
2. Item difficulty			0.441	0.477*
3. Point-biserial correlation				0.954*
4. Item discrimination				

* Correlation is significant at the 0.05 level

7.1 Item Analysis Results

Table 6 summarizes the item analysis metrics for 20 test questions based on responses from new dataset of 43 students. The key measures include item difficulty, point biserial correlation, and item discrimination index, as well as the number of item writing flaws identified per question.

Overall, most items showed moderate to high levels of item discrimination and point biserial correlations, suggesting that most of the questions were effective in distinguishing between students with higher and lower scores. In particular, items Q7, Q14, Q16, and Q18 stood out with especially strong discrimination indices (above 0.7), indicating they did a great job at identifying differences in student ability. On the other hand, item Q3 had a negative point biserial correlation, which indicates it may have been confusing or misleading, especially for stronger students, and might need to be revised or replaced.

Table 7 shows the Spearman's rank correlations between different item analysis measures. Interestingly, the number of item flaws had a significant positive correlation with the item discrimination index ($\rho = 0.564$, $p < 0.05$), which suggests that even flawed questions can still do a good job of telling strong and weak students apart. On the other hand, the link between item flaws and item difficulty was quite weak ($\rho = 0.068$), meaning that the presence of flaws didn't really make questions harder or easier.

We also found that point biserial correlation and item discrimination were moderately correlated with item difficulty ($\rho = 0.441$ and $\rho = 0.477$, respectively), suggesting that as questions become more challenging or easier, their ability to distinguish between students also changes. The strongest relationship was between point biserial correlation and item discrimination ($\rho = 0.954$, $p < 0.05$), which makes sense, since both metrics are designed to capture how well a question distinguishes stronger and weaker students.

These findings highlight that, while technical flaws do not necessarily undermine item performance, consistent monitoring of item statistics remains critical to ensure fair and effective assessments.

8 New Data Analysis

8.1 Overview

The following analysis incorporates additional data from dental students to supplement the original MAMS and Dental student dataset. This expanded dataset was analyzed using the same statistical procedures to determine whether the inclusion of new observations alters the original conclusions. All tests were rerun on the combined data set using consistent methods.

The purpose of this extended analysis is to assess whether relationships between total score and demographic factors change as sample size increases.

8.2 Normality Check

To determine the appropriate statistical tests for comparing total score across different demographic groups, the Shapiro-Wilk test was applied within each subgroup to assess normality.

Demographic Factor	**Group(s)**	**Shapiro-Wilk p-value**
Sex	Female	0.047
	Male	0.173
Race/Ethnicity	White	0.250
	Non-white	0.268
English Proficiency	Native	0.262
	Non-native	0.282
Born in USA	No	0.177
	Yes	0.493
Home Language	English	0.204
	Non-English	0.506

Table 8: Shapiro-Wilk test results for total score by demographic subgroup using the combined MAMS and Dental student data.

The results indicate that normality was satisfied for all demographic groups except for the female subgroup under sex ($p = 0.047$). Therefore, independent-samples t-tests were used for all comparisons except for sex, where the non-parametric Mann–Whitney U test was applied.

8.3 Non-parametric Comparisons (Mann–Whitney U Test)

Due to the violation of normality in the female group under sex and small group sizes in accommodations, Mann–Whitney U tests were conducted:

- **Sex**: There was no statistically significant difference in total score between male and female students, $W = 260.5$, $p = 0.084$, although the result is marginal and may indicate a possible trend with a larger sample.

- **Accommodations**: There was no significant difference in total scores between students with and without accommodations, $W = 40$, $p = 0.977$, suggesting accommodations did not impact performance in this sample.

8.4 Parametric Comparisons (Welch Two-Sample *t*-Tests)

For all other demographic variables, Welch two-sample t-tests were used. Results are summarized below:

- **Race/Ethnicity (White vs Non-White)**: No significant difference in total score was found ($t = -0.489$, $p = 0.628$; 95% CI: [–0.145, 0.089]).
- **English Proficiency (Native vs Non-Native)**: No meaningful difference in performance between groups ($t = 0.397$, $p = 0.694$; 95% CI: [–0.096, 0.143]).
- **Born in U.S. (Yes vs No)**: No significant difference was observed between students born in or outside the U.S. ($t = -0.047$, $p = 0.963$; 95% CI: [–0.164, 0.157]).
- **Home Language (English vs Non-English)**: No statistically significant difference in scores based on home language ($t = -0.521$, $p = 0.606$; 95% CI: [–0.148, 0.087]).

Conclusion: Across all demographic factors evaluated, the new data did not show any statistically significant conclusions. None of the demographic variables showed a statistically significant relationship with students' total score in the combined dataset.