Report 1

Chang Lu

February 2025

1 Introduction

This study investigates the relationship between **item difficulty** and **point biserial correlation (r_pb)** to assess how well different test items discriminate between high- and low-performing students. Additionally, normality tests were conducted on various demographic and quiz-related factors to determine the suitability of parametric or non-parametric statistical tests for further analysis.

2 Part 1: Correlation Between Item Difficulty and Point Biserial Correlation

2.1 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 1: 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 2: Q11 to Q20

2.2 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 3: 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 4: Point Biserial Correlation: Q11 to Q20

2.3 Correlation Analysis

To examine the relationship between item difficulty and point biserial correlation, we computed three different correlation coefficients:

- **Spearman's Rank Correlation:** $r = 0.361$ (moderate positive correlation, preferred method)
- **Pearson Correlation:** $r = 0.423$
- **Kendall's Tau Correlation:** $r = 0.244$

Each measure provides different insights into the relationship:

- **Spearman's Rank Correlation** is the most appropriate method because it does not assume normality. It assesses whether an increase in one variable is generally associated with an increase (or decrease) in the other, even if the relationship is not strictly linear. A Spearman correlation of 0.361 suggests a moderate positive monotonic association, meaning that as item difficulty increases, point biserial correlation also tends to increase in a general trend.
- **Pearson Correlation** measures the strength and direction of a linear relationship, assuming normality. While Pearson's correlation ($r = 0.423$) is slightly stronger, it assumes that both variables follow a normal distribution, which is not the case here. Therefore, its interpretation should be made with caution.

```
        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 5: PBC is not normally distributed

- **Kendall's Tau Correlation** is another non-parametric measure, robust to small sample sizes and tied ranks. The lower correlation value ($r = 0.244$) suggests that while the monotonic association exists, it is weaker than indicated by Spearman's correlation.

2.4 Findings

- A **moderate positive correlation** ($r = 0.361$) was found between item difficulty and point biserial correlation, with Spearman's correlation being the most reliable measure.
- **Medium-difficulty questions** (0.4 - 0.6) were generally the **best discriminators**.
- Some **easy questions** (difficulty ≥ 0.8) had high discrimination power, but **outliers** were present.
- A few questions had **negative r_pb values**, suggesting that high-performing students tended to get those items wrong more often than lower-performing students, which may indicate problematic questions.

2.5 Figures and Tables

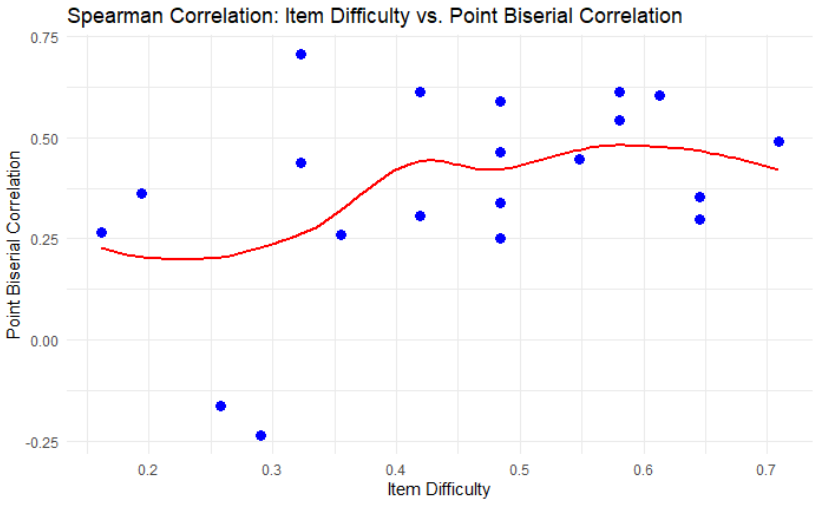

Figure 6: Relationship between Item Difficulty and Point Biserial Correlation

This plot 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.

3 Part 2: Normality Check

3.1 Methodology

- Data was cleaned and formatted to numeric values for normality testing.
- Three normality tests were conducted:
  - **Shapiro-Wilk Test**
  - **Kolmogorov-Smirnov (KS) Test**
  - **Lilliefors Test**
- Normality was tested for **Total Score** across different demographic and quiz-related factors.

3.2 Findings

- **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 7: 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 8: Race

– **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 9: 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 10: Born In USA

– **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 11: Home Language

- **Kolmogorov-Smirnov Test Results**:

– Accommodations group had sample size issues ($N < 3$ for one category). So we used another test.

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

Figure 12: Accomodations

- **Lilliefors Test Results**:
  - **Accomodations** Accommodations group had sample size issues ($N < 3$ for one category).

Accomodations <dbl>	Sample_Size <int>	Lilliefors_Statistic <dbl>	P_Value <dbl>
2	29	0.1447509	0.1246735
1	2	*NA*	*NA*

Figure 13: Accomodations

  - **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 14: Quiz Time

  - **Age Arrive USA**: Age arrive USA group had same problem as Accomodations 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 15: Age Arrive USA

3.3 Statistical Test Selection

- **Use T-Test (parametric)** for:

  - Quiz Time (categorized by 10 minutes)
  - English Proficiency
  - Race/Ethnicity
  - Home Language
- **Use Mann-Whitney U Test (non-parametric)** for:
  - Accommodations
  - Age Arrive USA
  - Sex
  - Born in USA

4 Conclusion

- There is a moderate positive correlation between item difficulty and point biserial correlation, but some items behave unexpectedly.
- Medium-difficulty questions tend to be the best discriminators, but some easy questions also have high r_pb.
- Normality tests suggest that some demographic groups require non-parametric tests, while others meet parametric test assumptions.