**Statistical Analysis Spring 2025**

**MSSP Consulting Services**

## MCQ Analysis Project

n= 31, there were 32 responses but one person did not complete the demographic survey. Question: Should we still include them in the analysis for the assessment?

Student populations: Dental and MAMS

1. **Spearman’s correlation**: (non-parametric) (presence or absence of a flaw)

Presence of item flaws and point biserial, and item difficulty.

Testing relationships between variables to see if we can predict behavior. Do the variables change in a related or independent fashion?

Point biserial is a correlation itself: Does the item predict how they did on the overall exam? how they did on that question compared to how they did on the exam. Measures the direction and strength of the linear relationship of one factor that is measured on a dichotomous scale (correct/incorrect) and the other one that is measured on a continuous scale (grade on exam).

Question: can you do correlation between item difficulty and point biserial?

There is also another way to determine the discrimination of an item: using the discrimination index. It distinguishes the performance of students who did well on the exam and students who did poorly for each item.

Question: would it make more sense to use the item discrimination index instead?

in addition to the relationship between the number of item flaws and the point biserial/item difficulty. We were wondering about the relationship between average time taken on the exam and item difficulty and the average time and number of flaws.

**Spearman’s correlation (2-tailed):** correlation among the presence of flaws, item difficulty, and average time time take on question

1. **T-tests**: to test differences in exam performance between each demographic group.

Independent variable: White vs URG

Dependent variable: exam performance

We are examining exam performance in the 2 groups. Is there a difference in performance between the two groups?

Demographics we are interested in:

| Accommodations | Sex | Race/Ethnicity | English Proficiency | Born USA | Home Language |  |
| --- | --- | --- | --- | --- | --- | --- |

We also collected data on age arrived in the USA. Question: how can we use this data to test differences based on the age they arrived? Because there are few from each group it may not make sense to include?

We used a likert scale - 1) 0-23 months 2) 2-5 years 3) 6-10 years 4) 11-17 years 5) >18 years

Question: Because the data may not be normally distributed, would it make sense to run these statistics using the Mann Whitney Test?

1. **Mann Whitney U test** (Non-parametric test, equivalent to t -test): compares two independent groups: (looking at specific flaws)

* Effect of presence of  a specific flaw on item difficulty
* Effect of presence of a specific flaw on point biserial

Question:

One thought for the analysis relates to the expected relationship between p value and rpb. It occurred to me that theoretically this must be a parabolic relationship because rpb = 0 both at a p value of 0% and 100%, so doing linear correlation is questionable. My hypothesis is we would expect a positive correlation between p and rpb for lower score ranges, then a negative correlation for higher score ranges. I vaguely recall the maximum point on a graph of rpb v p is at about p= 0.65 – 0.70 (which is why licensure exams set their passing score at this level). My question then is what the mean score was on the test for BU students versus UCF students? If it was lower, this might partly explain the positive correlation you are seeing at BU versus the one seen in the UCF data.

## Rubric Analysis from Spring 2024

The initial study looked at the interrater reliability (4 raters):

A couple of questions (17, 23, 28) showed low reliability

A couple of rubric items (2, 6, 12) showed low reliability

Now we have modified the rubric and will send it to external reviewers (2-3) once more.

Then we will need the analysis of inter-rater reliability (IRR).

## Differential Item Functioning

Quantitative Analysis Item Performance and Student Demographics: Logistic regression-based Differential Item Functioning (DIF) analysis will be conducted to determine the effects of several variables (sex, race, English proficiency, languages spoken, years in the United States) on answering exam questions correctly (Martinkova et al., 2017). Through this analysis we will be able to determine if students from different backgrounds perform differently on questions. Students from a focal group (the minority regarding race, sex, or linguistic background) will be compared to students from a reference group (the majority group). If DIF is identified, there is the potential that a particular question requires other knowledge besides from content knowledge for it to be answered correctly. If this is determined to be true, then the question can be classified as unfair. The presence of bias in a question would make the question not equitable since the actual student’s content knowledge cannot be properly determined.