MSSP Consulting: Limited Duration Project Description and Intended Scope

Client: Dr. Marisol Lopez
Graduate student: Gloria Urrutia
School / Department: Chobanian & Avedisian School of Medicine
MSSP Team: Vindyani Herath, Chang Lu, Qiuyi Feng, Beiming Yu
Intake Date: Jan 31, 2025
Faculty Supervisor: Masanao Yajima

Project Description and Goals:

Ultimate Goals: The primary objective of this study is to analyze the impact of technical item flaws on test performance among students and to validate a rubric to assess the course exam's multiple-choice questions across two schools of medicine.

Experiment Setups:
The participants included 32 students, consisting of first-year dental students and graduate students from the medical science program at BU. Each participant completed a 20-question assessment designed to evaluate their knowledge in the subject and identify technical item flaws, as well as a separate 30-question assessment used to validate a rubric for evaluating question quality. The National Board of Medical Examiners (NBME) item writing guide was used to evaluate flaws in the questions, and the association between the number of flaws answered and participant demographics was analyzed.

Experiment Descriptions & Procedures:

1. Data Collection

A total of 32 students participated in the study, including first-year dental students and graduate students from the medical science program at Boston University. Each participant completed two assessments:

1.1 A 20-question multiple-choice test designed to evaluate subject knowledge and identify technical item flaws based on the National Board of Medical Examiners (NBME) item writing guide.

1.2 A 30-question assessment used to validate a rubric for evaluating question quality. In addition, participants who are willing to share their information provided demographic information, including race, sex, English proficiency, home language, and test accommodations.

2. Evaluation metrics:

The test questions were analyzed based on item difficulty, point-biserial correlation , and the number of flaws in each question.

3. Improvement

Additionally, four raters reviewed the test questions using a rubric, and inter-rater reliability (IRR) was measured to see if the raters agreed. Based on the findings, flawed questions will be fixed or removed, the rubric will be improved, and further research will explore how test strategies affect student performance to ensure fairness in medical education assessments.

Definitions, Vocabularies, Basic Understandings:

Item difficulty - The proportion of test-takers who answer a particular item correctly.

point-biserial correlation coefficient (rpb) - Measures the correlation between a dichotomous (right/wrong) test item and the total test score (continuous variable).

Technical Item Flaws – Mistakes in question design that may make it harder for students to answer correctly, such as unclear wording, double negatives, or misleading answer choices.

Differential Item Functioning (DIF) – A method to check if a test question is fair by analyzing whether different groups (e.g., native vs. non-native English speakers) perform differently on the same question, even if they have the same knowledge level. If a question is harder for one group but not the other, it might be biased.

Inter-Rater Reliability (IRR) – A measure of how well different reviewers (raters) agree when evaluating the same test questions. High IRR means the rubric used to score the questions is clear and consistent.

Expectations of the Experiments / Desired Outcomes:

The primary audience includes educators, medical school faculty, exam designers (NBME), and education researchers who aim to improve the fairness and effectiveness of multiple-choice exams. Additionally, students may benefit from a better understanding of how test flaws impact their performance.

The primary goal is to identify the impact of test question flaws on student performance and determine if flawed questions make exams more difficult or unfair for certain groups.
Ensure fairness in medical education assessments – Analyze whether different student demographics (e.g., race, English proficiency) are affected by item flaws or biased questions.

Data Descriptions:
Variable names:
The data set includes the correctness of Q1-Q20 for each student, the total score and the time they use. It also collected the demographic of the students including age, sex, native language, race, age when arriving in the USA.

Improve test design and grading rubrics – Provide recommendations to remove or fix flawed questions and enhance the rubric to ensure clear, unbiased, and valid test questions.

Additional Comments:
The client has already run several statistical analyses such as independent sample t-tests, Spearman's rank correlation, and Mann-Whitney U tests to compare performance between different student groups, and logistic regression (DIF analysis) to check if some questions were unfair to specific groups. The client needs help in checking the models she fitted and determining their reliability.

Initial milestone and Deliverables:

To address the Client's needs, we will use reasonable efforts to provide the following initial work product for Client's review within 2 weeks from the intake meeting.

1. Exploratory data analysis to understand the data. Once we have a clear understanding of the data, we will start checking on the client's analysis.

Purpose of the consulting: The purpose of this consulting arrangement is 1) to train both Boston University (BU) MSSP and PhD students on the process of statistical consultation and 2) to provide a service to improve the quantitative aspects of research/projects on BU's campus. Student consultants will be monitored by MSSP faculty in order to help the consultants provide their best possible service. Nevertheless, we ask that you understand that students are not (yet!) professional consultants, and we thank you for your patience and cooperation in this regard.

Scope of the project: The project described below has been approved for what we call "limited duration" consultation. As the name suggests, such projects are intended to be fairly focused, with a total duration of no more than 10 student hours' worth of work. Consultation will involve MSSP student consultants and PhD leaders, supervised by MSSP program faculty. An initial intake meeting should have been held and additional meetings, as needed, will be arranged at the mutual convenience of the Client and the student consultants.

Attribution: When the scope of the work is of 'limited duration' but nevertheless is used in any sort of academic product (e.g., presentation, project, abstract, or publication), we would ask that you please include an appropriate acknowledgment. For example, Clients may include "We acknowledge the help of [NAME] of the MSSP statistical consulting service under the guidance of Professor [NAME]" as a footnote accompanying grant support and related uses[1].

Modification of Intended Scope of work: At any time during this process, if both the Client and the MSSP Team mutually agree that the above proposal does not adequately address the Client's needs, we leave open the possibility of holding additional discussions to alter the above proposal. Consultations anticipated to require substantially more than 10 hours of work may be recommended for collaborative consulting. Evaluation of such recommendations typically will include a discussion of co-authorship on resulting end-products.

Disclaimer: Client acknowledges the work will be performed by current students in BU's MSSP and PhD Program with some faculty supervision. As such, we make no representations or warranties, express or implied, of any nature, including without limitation, warranties of merchantability or fitness for a particular purpose. Client acknowledges and agrees that all work performed is provided on an "as-is" basis only. We will use reasonable efforts to preserve the privacy and security of data provided to BU or work product generated by BU and its students and faculty that includes material identified by Client as confidential, but we shall have no liability for any loss or release of data or data use.

[1] We explicitly do not insist upon co-authorship for projects of `limited duration' type -- although, of course, if the conventions of your field permit, co-authorship for at least a subset of the consulting team is always greatly appreciated. This MSSP policy is intended to allow clients maximum flexibility in seeking out and benefiting from our statistical consulting work while, at the same time, promoting best practice regarding academic intellectual property. Our perspective is informed by the American Statistical Association's guide *When You Consult a Statistician … What to Expect*, which states that any such product "should acknowledge the participation of the statistician, consistent with the value of that participation."