Multiple Choice IRR Project Report

2024-05-03

Faculty Supervisor: Masanao Yajima
Teaching Fellow: Shiwen Yang
Member: Tszwai Ng, Huaijin Xin, Tahuang Chen, Bolong Xian

1 Introduction

The topic of this consulting project is to develop and validate a rubric to assess the course exam's multiple-choice questions. Our client, Marisol Lopez, is a professor at the Department of Pharmacology, Physiology, and Biophysics at the School of Medicine at Boston University. In this project, her goal is to determine the best statistical analysis to validate the inter-rater reliability.

Our study's primary goal is to analyze and validate the inter-rater reliability (IRR) of multiple-choice questions to ensure consistent evaluation and fair outcomes. The inter-rater reliability (IRR) statistic is applied because it quantifies the extent of agreement among different raters evaluating the same subjects using a set of criteria. This metric is crucial for ensuring that the assessment tool yields consistent and reliable results, regardless of who is applying it. High IRR indicates that the tool is designed so that different raters perceive and score the criteria similarly, minimizing the influence of subjective judgments.

2 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. A higher level of agreement on one rubric will serve as evidence that the rubric is less ambiguous to the raters. A 30-question multiple-choice exam was used. We should note that two types of bias may affect the clarity of a rubric:

1. rater's bias – The creators of the rubric may hold inherent biases towards its criteria and potential outcomes. To mitigate this, creators are excluded from the pool of raters to ensure an unbiased assessment. Also, the proficiency of different raters may lead to this bias. Thus, those raters with no understanding relevant to the questions should also be excluded.
2. question's bias – The inherent clarity or complexity of the questions may influence the rater's judgment. Since it's unlikely to change those questions with low clarity or high complexity, an analysis involving controlled variables would be necessary to address this issue.

This is a fully crossed design where all subjects are rated by multiple raters. (Hallgren, 2012) In the context of IRR, fully crossed design is when subjects rated by multiple raters gets rated by the same set of coders. This design is chosen to ensure that each rater's assessments can be directly compared with every other one on the same subjects. This allows for a comprehensive analysis of agreement among raters.

Our clients planned to recruit six raters on a voluntary basis among the colleagues of our clients, and by the time we started the analysis, they had responses from 4 of the raters. Each rater was 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"s are also present in the matrix. For example, a criterion says, "If the answer choices are numerical, are they listed in ascending or descending order?" which would not be applicable to questions with non-numerical answers.

3 Methodology

This project uses the kappa statistic to calculate the inter-rater reliability. The kappa statistic is used in statistics to measure inter-rater reliability for qualitative items.

Cohen's Kappa is designed for situations where two raters are involved. It measures the degree to which two raters agree beyond what would be expected by chance alone. Fleiss' Kappa assesses the reliability of agreement between three or more raters. It is suitable when raters are randomly sampled from a larger population.

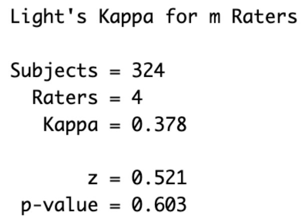
Light's Kappa addresses a fully crossed design where three or more raters assess all subjects. It calculates the kappa statistic for each pair of raters and then uses the mean of the kappa statistic to provide an overall index of agreement (Hallgren, 2012). Since its assumptions match those of our study design, we shall proceed with Light's Kappa.

Our data comes in as a 30 × 12 × 4 array − 4 raters, each filling out a 30 × 12 binary matrix. To quantify the clarity of each criterion, we compute the kappa statistic using the slice of the array that corresponds to the criterion, which has dimension 30 × 4. In this case, the subjects are the 30 questions. Then, to assess the potential bias from the questions, we find the slice of the array that corresponds to the question, which has dimension 12 × 4. In this case, the subjects are the 12 criteria. Note that fixing criterion/question, a rater might rate all questions/criterion the same way by chance, thus resulting in a constant column in the data matrix. Such columns are problematic because Cohen's kappa requires the ratings from each rater to have some level of randomness.

Light's kappa is the average pair-wise Cohen's kappa. In Cohen's kappa implementation in R, specifically, 0 is the default value when one of the raters gives the same rating for every subject. So, if we include such columns, Light's kappa will be a lower score than if we exclude such columns. We decided to maintain such a column to protect the authenticity in the end. This shouldn't matter too much since there are very few of these columns.

4 Result

Here is the result of running the Light's Kappa function to our final data frame:



```
Light's Kappa for m Raters

Subjects = 324
  Raters = 4
   Kappa = 0.378

       z = 0.521
 p-value = 0.603
```

The results show that the Light's Kappa statistic, which measures the agreement among multiple raters beyond chance, is 0.378 for four raters evaluating 324 subjects. This kappa value indicates fair agreement but suggests that the consistency across raters is moderate and not particularly strong. The associated z-value of 0.521 and a p-value of 0.603 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.

This analysis highlights the need to revise the rating criteria or train raters to achieve more consistent evaluations.

Then, we got some more detailed results to analyze where the potential problems were.

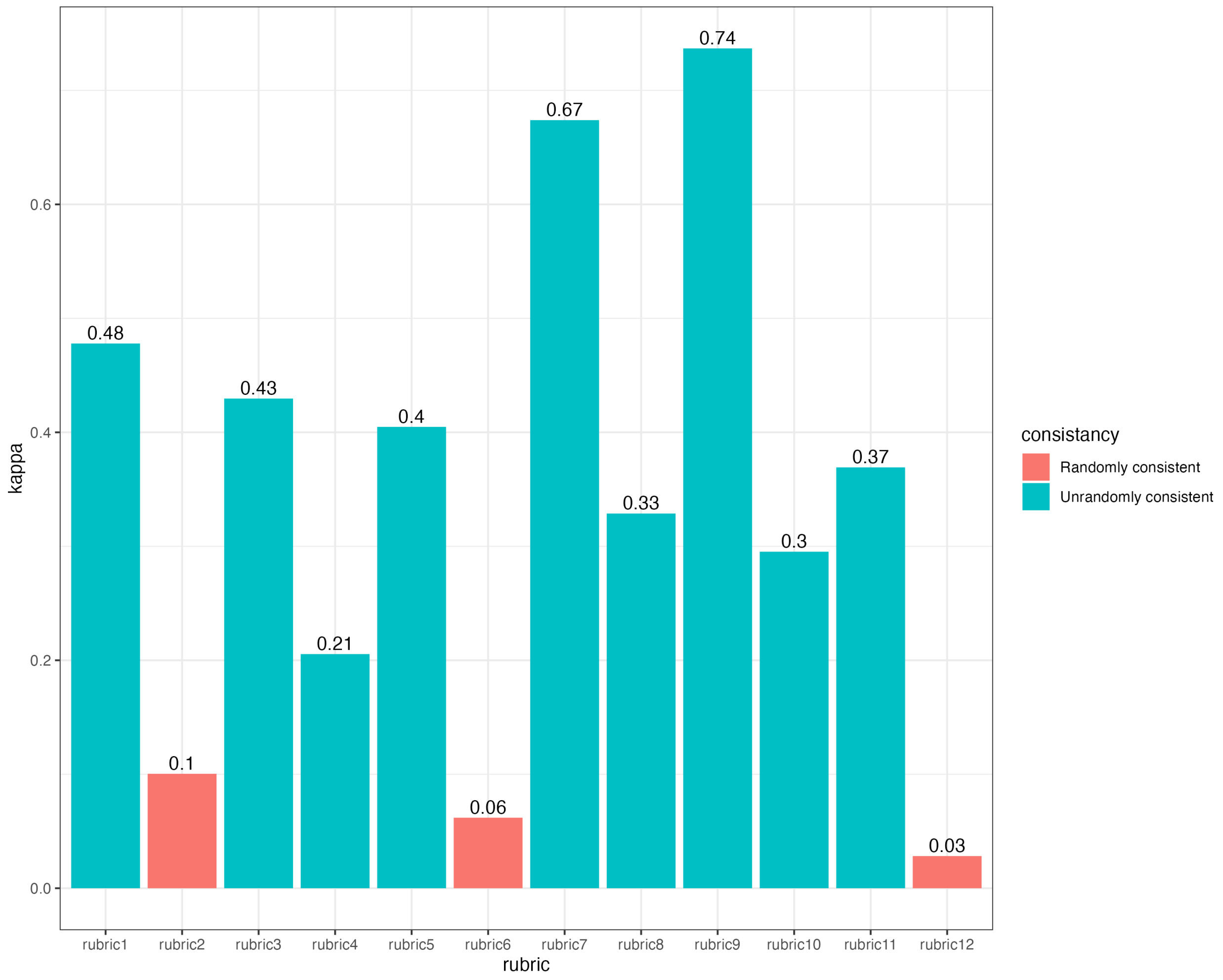


Figure 1: Kappa value by rubrics

Figure 1 illustrates the kappa values for various rubrics used in the evaluation process, revealing considerable variation. Rubric 5 shows a deficient kappa value of 0.06, indicating negligible agreement among raters, likely suggesting that the rubric is either poorly understood or not consistently applied. On the other hand, rubric 10 and rubric 9 exhibit high kappa values of 0.74 and 0.67, respectively, suggesting strong agreement among raters for these criteria. Very low kappa values for specific rubrics such as rubric2, rubric6, and rubric12 indicate almost no agreement beyond chance, highlighting potential issues in the clarity or relevance of those specific evaluation criteria.

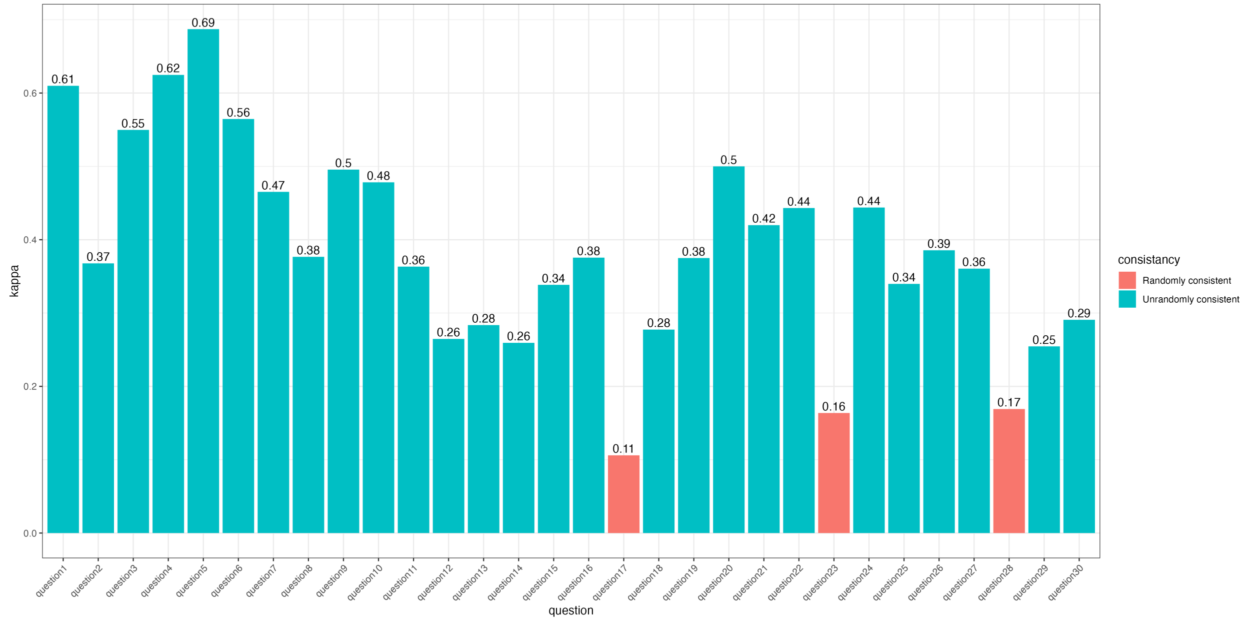


Figure 2: Kappa value by questions

Figure 2 presents the kappa values for inter-rater agreement across 30 questions, showing varying degrees of agreement among raters. The kappa values range widely, with some questions displaying higher levels of agreement indicative of more precise or more uniformly interpreted criteria (e.g., Question 4 with a kappa of 0.62 and Question 5 with 0.69). In contrast, some questions show significantly lower kappa values, such as Question 17, Question 23, and Question 28, which have kappa values of 0.11, 0.16, and 0.17, respectively, suggesting that any agreement among raters on these questions might be largely due to chance rather than a consistent application of evaluation criteria. There would be potential to highlight ambiguities or inconsistencies in how the questions are understood or applied.

The variability in kappa values across both individual questions and criteria indicates areas where rater agreement is solid and other areas where it is notably weaker, which may necessitate a review of the clarity and specificity of the evaluation criteria. Lower kappa values point to potential ambiguities or subjective interpretations that could be mitigated by unifying raters' understanding of questions and criteria, thus improving the ratings' reliability.

5 Conclusion

The study evaluated 324 subjects and employed Light's Kappa to provide a view of the inter-rater reliability among four raters. The Light's Kappa analysis indicates a fair level of consensus among raters, with an overall Kappa score of 0.378. While this level of agreement is not high, it suggests that the raters align more than would be expected by chance alone.

The detailed analysis of Kappa values by question and criteria reveals significant differences in inter-rater reliability, suggesting that some questions or criteria may be more straightforward or aligned with the raters' shared understanding of the evaluation criteria. Conversely, questions and criteria with notably low Kappa values point to potential ambiguities or subjective interpretations. Clarification and reevaluation should be made to the questions and criteria that show lower Kappa scores. Ensuring a common framework of understanding would enhance the consistency and fairness of the evaluations.

Reference

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. Tutoring Quantitative Methods in Psychology, 8(1), 23-34.