Removing Barriers for Students From Groups Underrepresented in Medicine (URiM) and/or English Language Learners (ELLs) by Providing Equitable Assessments

Rebecca Petre Sullivan[1], Erin B. Bruce[2], Marisol E. Lopez[3]

[1]Lewis Katz School of Medicine at Temple University, Department of Biomedical Education and Data Science, Philadelphia, PA, USA; [2]University of Florida, College of Pharmacy, Department of Pharmacodynamics, Gainesville, FL, USA; [3]Boston University Chobanian and Avedisian School of Medicine, Department of Pharmacology, Physiology & Biophysics, Boston, MA, USA

Poster # 109

How do we insure MCQs are assessing student learning and not test taking strategy?

Guideline	No Flaw	Flaw	N/A
Can the question be answered without looking at the options? Avoids "Which of the following statements is true?" without context. Specific enough to be able to create a list of responses in our head. Could answer if it was a short answer.			
Is there a single best answer choice? Either more than one correct answer or no correct answer.			
If the answer choices are verbal, do they avoid long or complex options? If the answer choice is more than 1 line, we will choose No. Each answer choice should answer no more than one question. Meaning there should not be two parts to the answer that both have to be true.			
Are the stem and lead in present, focused, clear, and succinct? If the question includes a vignette: It should avoid superfluous language. Details should be guided by the level of the test taker.			
If the answer choices are numerical, are they listed in ascending or descending order?			
Do the item options avoid vague terms such as "usually" or "often" or absolute terms such as "always" or "never"?			
Do the item options avoid None of the above or All of the above?			
Are the item options parallel in grammatical form and/or structure and avoid grammatical clues? All answers should be either passive or active voice. One answer choice should not be structurally different from majority of other choices. Answers should have consistent word choices and avoid synonyms. Ex) all 'decreased' or all 'reduced,' not mixed a) reduced b) decreased			
Does the item avoid negative phrasing such as "not," "except," or "which of the following statements are false?"			
Do the item options avoid grouped outcomes (increase, decrease, no change along with two other distractors) and option pair(s)? If all answer choices have the same direction (all increase) it doesn't count as "grouping." The item writer should be able to rank order each option on the same dimension.			
Does the correct answer avoid repeating a word or phrase from the stem?			
Does the correct answer avoid having the most repeated terms among all options? (excluding terms that are repeated in every answer)			

Follow this QR code and try the new and improved rubric!

You can help us refine our guidelines with your feedback

In a healthy heart, which of the following statements is TRUE?

A. Depolarization of the interventricular septum occurs from right to left.

B. The PR interval is longer than the QT interval.

C. The wave of depolarization passes more quickly through the AV node than the Purkinje fibers.

D. The QRS complex is predominantly upright in Leads I and II.*

E. The mean electrical axis is generally between +100 and +150 degrees.

Acknowledgments: We would like to thank our external reviewers Dr. Jonathan Kibble (University of Central Florida), Dr. Judith Saide (Boston University), Dr. Shi Jin (University of Florida), and Dr. Jaya Kolli (University of Florida) for their time spent assessing the multiple-choice exam questions using our rubric. In addition, we acknowledge the help of Shiwen Yang of the Boston University MSSP statistical consulting service under the guidance of Professor Masanao Yajima.

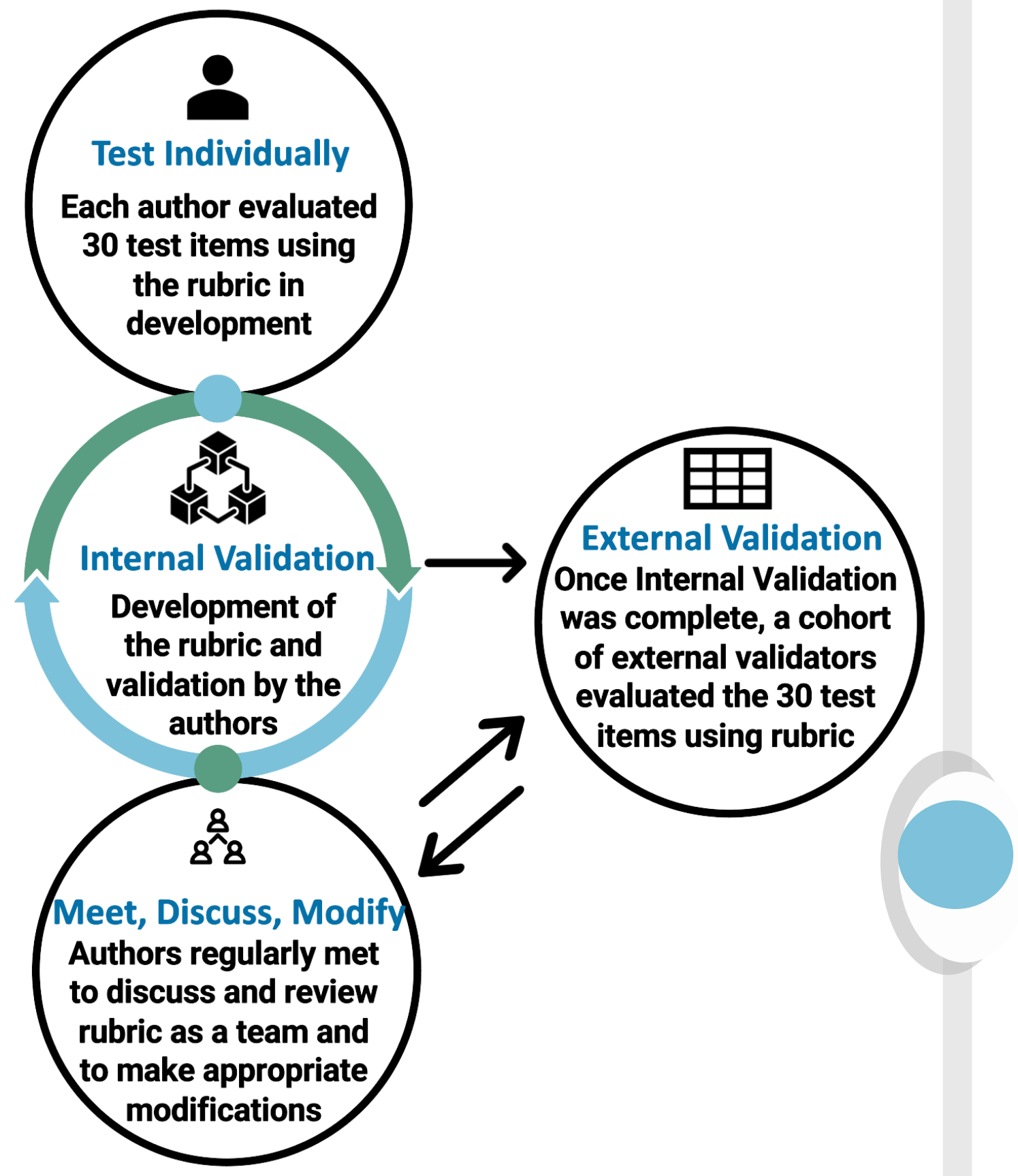
How it Started

The authors met at the **American Physiological Society Summit 2023 Pre-Summit Center for Physiology Education workshop**. We found we each have a similar problem with our **High-Stakes MCQ exams**. Students are consistently **failing**. What we don't know, is **"Why?"** Are they failing ***because they aren't learning the material***, or ***because the exam questions are flawed?***

Pre-Summit Center for Physiology Education Workshop
Thursday, April 20, 1:30–3:45 p.m. PDT

Title: Navigating Educational Research
Chair: Rob Carroll, PhD, FAPS, *East Carolina University, Greenville, North Carolina and Center for Physiology Education Advisory Board Member*

Marisol Lopez (she/her/hers)
Rebecca Petre Sullivan (She/Her)
Erin Bruce (She/Her)



Test Individually
Each author evaluated 30 test items using the rubric in development

Internal Validation
Development of the rubric and validation by the authors

External Validation
Once Internal Validation was complete, a cohort of external validators evaluated the 30 test items using rubric

Meet, Discuss, Modify
Authors regularly met to discuss and review rubric as a team and to make appropriate modifications

Rubric Development

We began developing a rubric based on Dr. Lopez's previous work[3] that would encompass **NBME Best Practices** in question writing, as well as **avoiding** factors that add **irrelevant difficulty** for English Language Learners and/or those who have not had the opportunity to develop **"test taking strategies."**

Experimental Design

The **rubric** consists of 12 guidelines. Four raters assessed 30 MCQs using each of the 12 guidelines in the rubric. They could enter "Yes", "No" or "N/A" for each response.

Two potential types of **bias**:

1. Rater bias
   - Creators of the rubric may hold inherent biases
     - Exclude creators from validation
   - Different proficiency of raters
     - Exclude raters with no subject expertise
2. Question bias
   - Inherent clarity or complexity of question may influence rater's judgment

Statistical Methods

Inter-rater reliability was assessed using **Light's kappa** to determine the level of agreement among the raters using the rubric. A higher level of agreement on one question serves as evidence that the question is less ambiguous to the raters and that the raters agree more than by chance alone.

Results

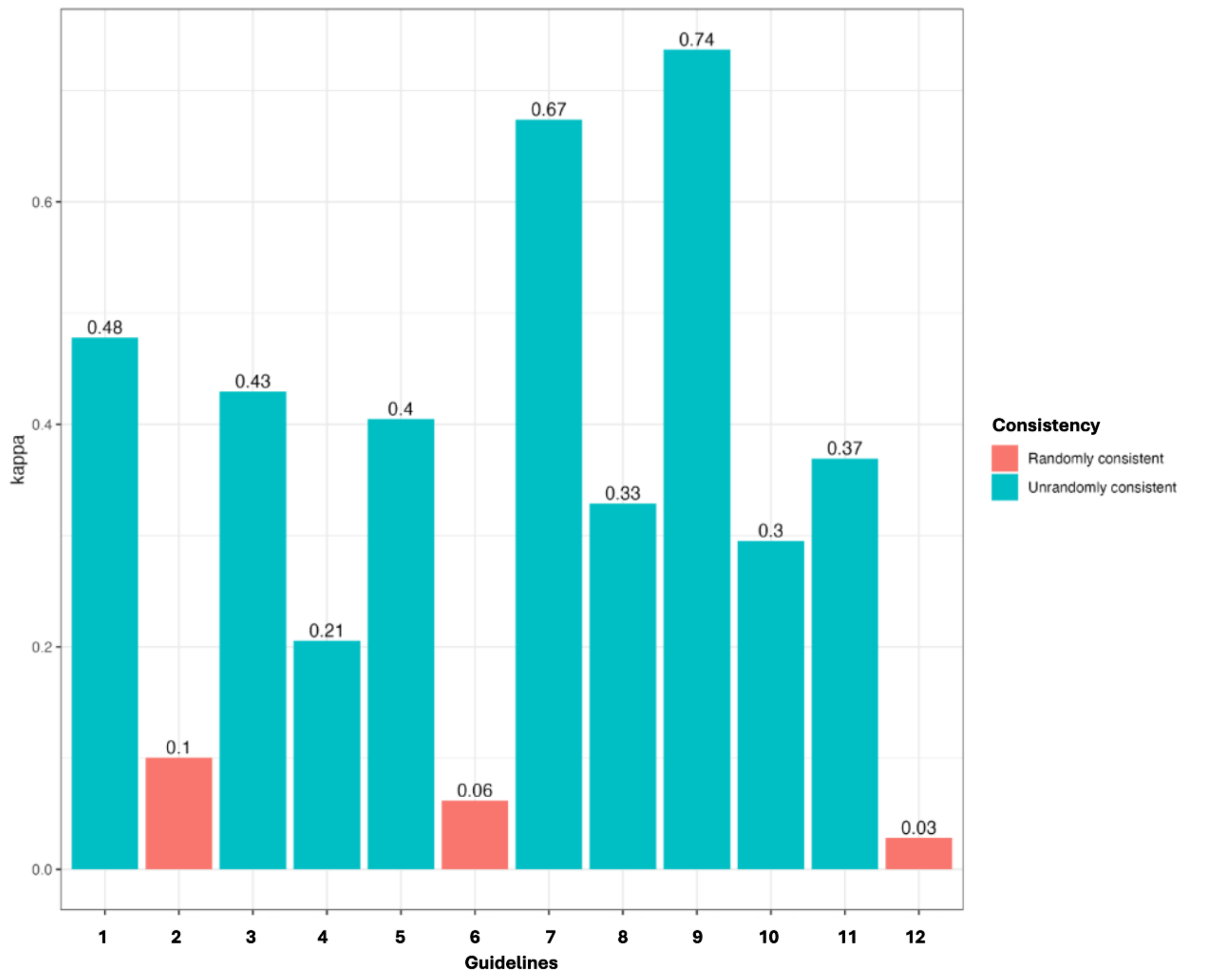


Figure 1.
Assessment of Inter-Rater Reliability (IRR) for each of the 12 guidelines using Light's Kappa. Guidelines 2, 6, and 12 (κ = 0.1, 0.06, and 0.03, respectively) show the level of agreement between the raters can be attributed to random chance. Guidelines 7 and 9 have the highest Kappa values, suggesting strong agreement among raters. All other guidelines are determined to have a moderate level of agreement that cannot be attributed to chance alone.

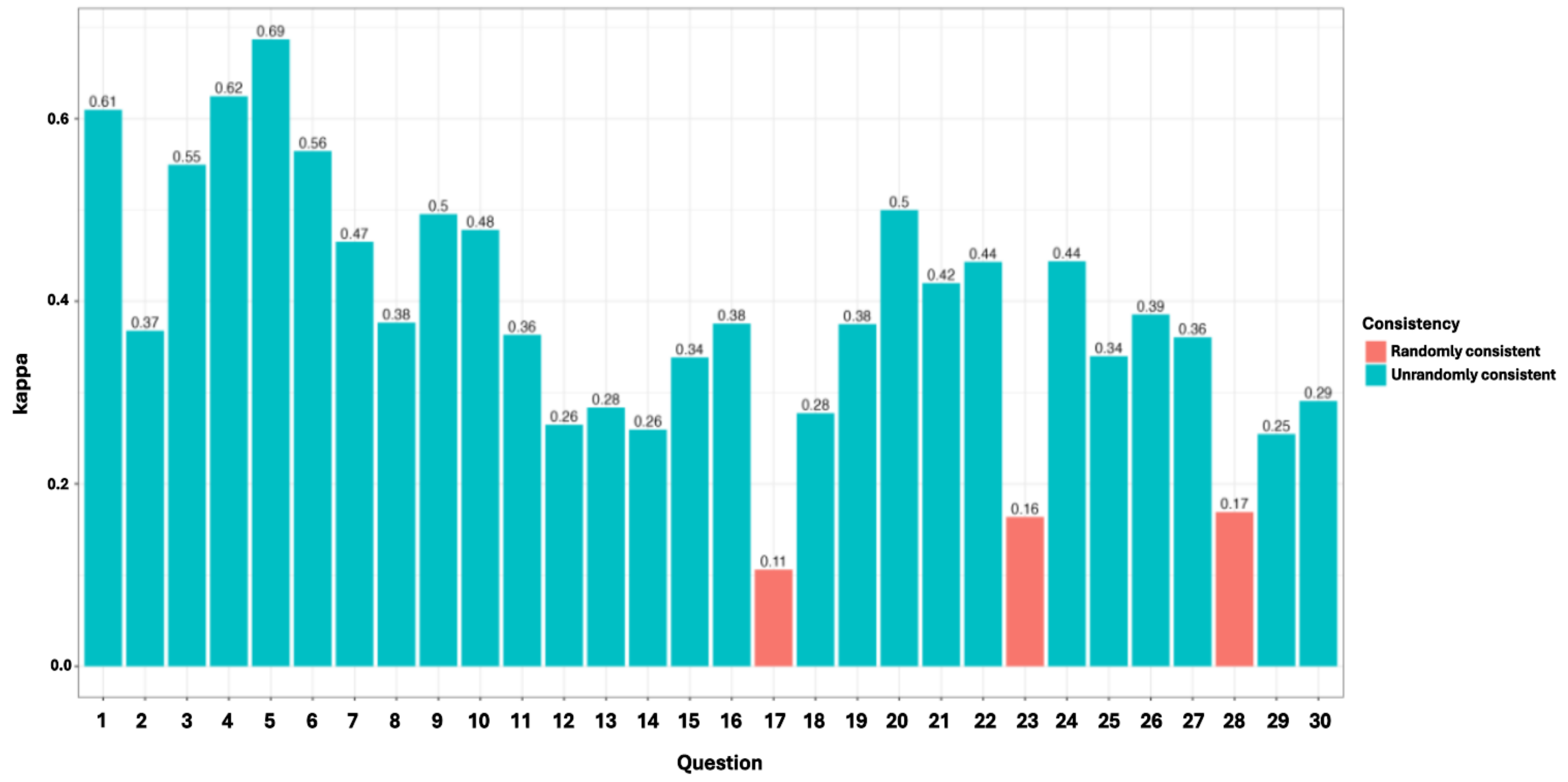


Figure 2.
Assessment of IRR for each of the 30 multiple choice questions using Light's Kappa. Questions 17, 23, and 28 (κ = 0.11, 0.16, and 0.17 respectively) show the level of agreement between the raters can be attributed to random chance. Whereas questions 4 and 5 have the highest Kappa values among raters (κ = 0.62 and 0.69, respectively). All other questions are determined to have a level of agreement that cannot be attributed to chance alone.

Feedback from Validators

- Change the instructions on how to complete the rubric by asking the validators to provide an answer that matches the question on the rubric (Yes, No, or N/A) rather than whether it has a flaw or not.
- It is easy to assess Guideline 1 when questions fall on the extremes of being either very broad ("Which of the following…") or being very specific ("Where is the greatest blood pressure drop?"). However, this guideline is more difficult to assess for those questions that fall somewhere in between this range.
- If the stem is not specific enough, some questions could have more than one correct answer choice when items are assessed with Guideline 2.

Summary

□ There was a moderate agreement among the four raters' assessment of the MCQs, κ= .378, p = 0.603. This means that raters agree more than would be expected by chance alone.

□ The guidelines that were found to have very low kappa scores (2, 6, and 12) have been modified to improve their clarity. In addition, an example has been added to rubric 12 to aid in understanding. The modified rubric will be provided to the raters for reevaluation, and the Inter-Rater Reliability will be reassessed.

Future Directions

- **Assess** question banks for physiology exams at each institution.
- **Disseminate** to educators in any program or discipline.
- **Support Best Practices** by eliminating flaws that provide irrelevant difficulty or benefit test wise students.

References:

1. Breakall, J., Randles, C., & Tasker, R. (2019). Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry. *Chem. Educ. Res. Pract. 20*(369), 369-382. https://doi.org/10.1039/c8rp00262b
2. Lee, E.N. & Orgill, M. (2022). Toward equitable assessment of English language learners in general chemistry: Identifying supportive features in assessment items. *J Chem. Educ. 99*, 35-48. https://doi.org/10.1021/acs.jchemed.1c00370
3. Lopez, M.E. (2023). Assessing multiple-choice questions based on language precision and best practices to promote equity in the Dental Physiology course. *Physiology 38*(S1). https://doi.org/10.1152/physiol.2023.38.S1.5731455