Sari, H., Shenoy, V., VanBuren, V., Dubuisson, J., Dickey, D., & Johnston, J., (2026). Development of Standardized Guidelines for Post-Exam Adjustments in Medical Education Assessments. *Intersection: A Journal at the Intersection of Assessment and Learning, 7*(1), 105-117.

Development of Standardized Guidelines for Post-Exam Adjustments in Medical Education Assessment

Halil Sari, Ph.D., Vinayak Shenoy, Ph.D., Vincent VanBuren, Ph.D., James Dubuisson, B.A., Danielle Dickey, Ph.D., Jody Johnston, B.S.

Author Note

Halil Sari, https://orcid.org/0000-0001-7506-9000

"We have no conflicts of interest to disclose."

Intersection: A Journal at the Intersection of Assessment and Learning
Volume 7, Issue 1

Abstract: This study addresses the need for standardized guidelines for post-exam adjustments in medical education. It emphasizes the principles of validity and reliability, which are essential for evaluation of student performance. The research examines current practices and identifies key challenges: inconsistencies across curriculum tracks, grade inflation, and the unintended consequences of ad hoc decision-making. Limitations of existing approaches—such as awarding full credit for flawed items or making emotionally driven decisions—can compromise assessment integrity and equity. Therefore, we developed an evidence-based framework for grade adjustment. These guidelines aim to enhance consistency, fairness, and transparency of assessments. By anchoring decisions to psychometric data and expert consensus, the framework reduces subjective variability and improves grading defensibility. Implementation of guidelines led to a reduction in item-level adjustments and improved faculty and student confidence. This work refines assessment methodologies and has potential implications for improving student learning outcomes, clinical readiness, and overall healthcare quality.

Keywords: *post-exam adjustments, validity, reliability, multiple choice exams*

Introduction

Assessment in medical education plays a critical role in ensuring that students achieve the competencies required for safe and effective clinical practice (Newble et al., 1981; Raymond et. al., 2025). To maintain fairness and accuracy, assessment practices must adhere to principles of validity and reliability, which are widely recognized as essential for high-stakes examinations in health professions education (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014; Crocker & Algina, 1981, Messick, 1987). Validity refers to the degree to which evidence and theory support the interpretation of test scores for their intended purposes (AERA, APA, NCME, 2014), while reliability ensures that scores consistently reflect a learner's true performance rather than random error or subjective judgment (Crocker & Algina, 1981, Messick, 1987). When post-exam adjustments are made without standardized criteria, both validity and reliability can be compromised, leading to inconsistent grading and potential inequities (Aubin et al., 2020; Fullerton, 1993).

Despite the importance of these principles, many institutions lack formal guidelines for post-exam adjustments (Menon et al., 2021). Decisions are often made on an ad hoc basis, influenced by faculty discretion or student appeals, which can result in grade inflation, inconsistent practices across courses, and diminished confidence in the assessment process (Ginzburg et al., 2025; Reckase, 2023). Previous research in medical education has highlighted the need for structured, evidence-based approaches to grading and post-assessment review to ensure fairness and transparency (Aubin et al., 2020; Crocker & Algina, 1981; Raymond et al., 2024).

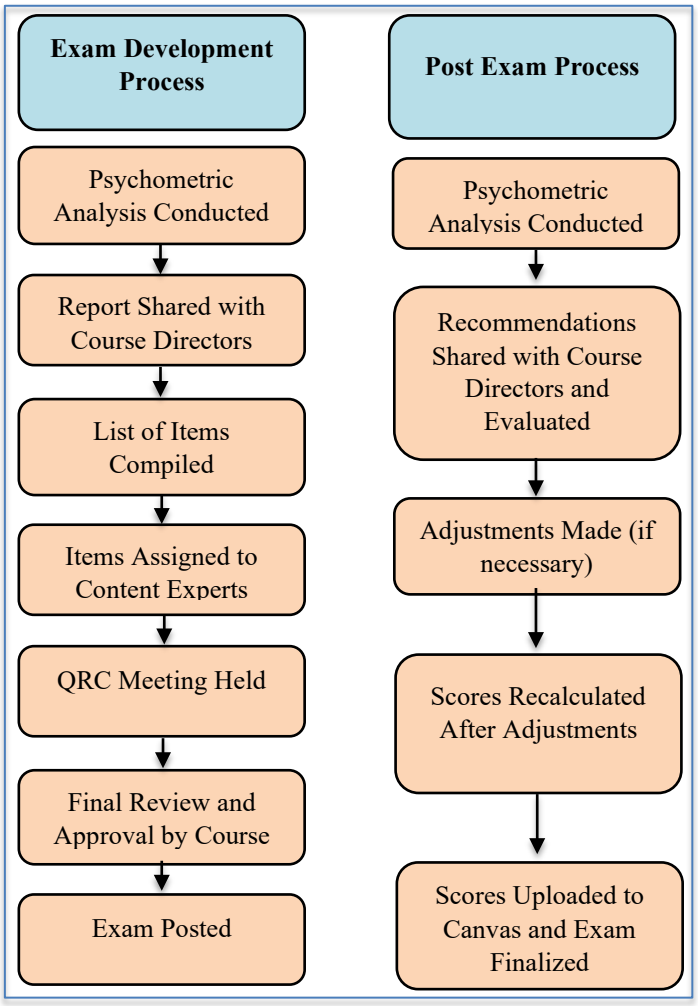
The purpose of this work was to develop and implement standardized guidelines for post-exam adjustments in a medical school setting and to examine their impact on grading consistency and assessment quality. These guidelines aim to reduce variability in decision-making, promote data-driven practices, and uphold the integrity of assessments. By providing a structured framework, this work addresses a common challenge in health professions education and offers a model that can be adapted by other institutions to enhance fairness and defensibility in grading decisions.

Exam Development Process

The Office of Evaluation and Assessment (OEA) serves as the central coordinating body for exam development within Naresh K. Vashisht College of Medicine at Texas A&M University. The OEA builds about 155 exams per academic year, including written exams, practical exams, objective structured clinical examination (OSCEs), quizzes, and remediation exams. The exam development team within OEA consists of four members, including a psychometrician and three program managers (experienced in assessment administration and curriculum logistics). The exam building process for a pre-clerkship in-house exam follows a structured approach. First, the psychometrician conducts an item analysis and flags questions that did not perform well in the past (e.g., item discrimination was very low or negative, item was very difficult or easy). The psychometrician then shares the item analysis report with the course directors and the exam development team. Next, the course directors compile a list of items they want to include in the exam, which is subsequently shared with the exam development team. This list includes both brand-new items and previous questions flagged either by the course directors or the psychometrician. The exam development team then forwards these items to the Question Review Committee (QRC). The QRC consists of six content experts from various disciplines, the course director(s) for that exam, two members from the exam development team, and one psychometrician. About two weeks prior to the QRC meeting, the exam development team assigns items to specific content experts for review. This review includes evaluating both previously flagged questions as well as new items for content accuracy, clarity, and the appropriateness of response options. During the meeting, the QRC members collectively discuss potential issues with both flagged items and new questions, suggesting necessary modifications to the item stems and response options. These meetings typically last for about an hour, during which the committee reviews an average of 5 to 10 items per exam. Finally, the exam development team uploads the revised exam for the course directors' final review. Upon approval, the exam is finalized and prepared for administration.

Figure 1

Flowchart of Exam Development and Post



This structured review process aims to minimize flaws that could compromise validity and reliability. While rigorous development reduces the likelihood of problematic items, some issues inevitably emerge during administration, making post-exam review and adjustment necessary. Understanding this development process provides context for why standardized post-exam guidelines are critical—they complement, rather than replace, robust pre-exam quality assurance.

Post Exam Process

After each exam, preliminary scores, item analysis, and student comments are compiled and shared with course directors. A detailed psychometric analysis follows, and faculty engage in a structured

review to determine whether the assessment met its intended objectives. In this context, assessment objectives refer to the competencies and learning outcomes defined during the exam blueprinting phase, where items are mapped to course and session-level objectives. These objectives ensure alignment between instruction and assessment and serve as benchmarks for evaluating exam performance.

During post-exam review, faculty verify that item performance supports these objectives and that any necessary adjustments are evidence-based. This process includes evaluating flagged items, considering student feedback, and applying standardized guidelines to maintain fairness and consistency. Finalized scores are then uploaded to the learning management system, and a comprehensive report summarizing strengths and areas for improvement is distributed to faculty.

Problem Statement

Despite robust pre-exam quality assurance, a subset of items periodically fails to perform as intended during administration (e.g., extreme difficulty or poor discrimination). Historically, our institution lacked standardized, data-anchored guidelines to adjudicate such items post-exam. In practice, ad hoc decisions—often varying by curricular track (i.e., distinct academic pathways within the College of Medicine) and course—produced inconsistent outcomes and, at times, unintended grade inflation. The College of Medicine, regardless of the source of the measurement problems, has often tended to award full credit to all students when a test item was questioned. Some of these problematic test items were common across both, yet decisions regarding the same item sometimes differed between the tracks. Furthermore, inconsistent decisions have also been made across different courses within the same track. In some instances, post-exam decisions appeared to be made based on emotional responses to students rather than by objective data. For example, full credit was awarded for an entirely incorrect response because a large number of students chose that answer.

Another issue our institution has experienced is that when a high number of post-exam adjustments are made, students often receive too many points back, resulting in grade inflation[8]. Furthermore, due to inconsistent decision-making for each item, students may receive a different number of points, leading to disparities in their final grades. In some cases, students' grades remained unchanged, while others experienced a decrease in their exam scores. For example, when full credit is awarded for a response option that differs from the correct answer listed in the answer key, students who answered correctly receive no additional benefit. Conversely, when an item is removed from grading, students who answered correctly are penalized, while those who answered incorrectly benefit from a reduced denominator. Consider a scenario where a student answers 7 out of 10 items correctly, earning a score of 70%. After post-exam adjustments, one of the correctly answered items is removed from grading, lowering the total number of questions to 9. Now the student's adjusted score becomes 6 out of 9 (66.6%). As a result, the student who initially passed the exam before the adjustment, now fails if the passing threshold is 70%. Such scenarios foster student skepticism about the legitimacy of exams and undermine confidence in the reliability of in-house examinations. Therefore, there is a clear need for

standardized decision-making guidelines to ensure consistent and theoretically reasonable judgments based on data, rather than on emotional responses.

Development of the Grade Adjustment Guidelines

The guidelines presented in Table 1 were created through a systematic and iterative process. First, we analyze historical post-exam decisions by reviewing grading records and adjustment outcomes with the aid of our test development software, which automatically displays item properties from previous administrations. This allows us to compare historical and current item statistics, identifying patterns of problematic questions and their impact on grade reliability and fairness. Next, we conduct a psychometric review, including calculations of item difficulty (proportion correct, flagged at $p < .50$), discrimination (point-biserial correlation flagged at $rpbis < .10$), and distractor analysis (evaluating whether incorrect options functioned as plausible alternatives). These metrics follow established thresholds and provide actionable insights into item performance and quality. Each recommendation is grounded in classical test theory, which offers a practical and widely accepted methodology for evaluating item reliability and validity in educational assessment contexts (Crocker & Algina, 1981).

Consensus is achieved by sending flagged items and recommendations to the two course directors assigned to each course, who make final adjustment decisions. In rare cases of disagreement, the assistant dean of pre-clerkship makes the final decision. This process ensures that recommendations reflect both psychometric evidence and expert content review. The resulting guidelines map common item issues to evidence-based actions, prioritizing solutions that correct construct-irrelevant variance while minimizing unintended score distortion (AERA, APA, NCME, 2014; Aubin et al., 2020; Crocker & Algina, 1981; Kane, 2016).

Table 1

The Proposed Grade Adjustment Guidelines

Source of the Problem	Example Scenario	Recommendation(s)	Reasoning
1. Item was written in a way such that there is no clear right answer, or the item was poorly written	For an item, the stem stated, *"Which of the following is the most important function of the liver?"* However, multiple listed options—such as metabolism, detoxification, and protein synthesis—could all reasonably be argued as "most important," and the stem failed to specify a clinical context or domain.	Option 1: Remove the item from the grading with or without giving any credit to all students. Option 2: Award full credit to all students. Option 3: The course director may decide to keep the item as is (no change).	Poorly written items can confuse students and fail to accurately assess their knowledge, leading to unfair grading.
2. The answer key is mis-keyed.	For an item, the answer key was option B, however, content experts/course directors realized that the best answer should be option C.	Replace the answer key with the correct response and recalculate scores (i.e., the course directors will determine the best answer).	A mis-keyed answer can lead to incorrect scoring, misrepresenting students' true understanding.

3. There are two best answer choices.	For an item, the answer key was option B, however, content experts/course directors realized that option C is ALSO the best answer.	Option 1: Award full credit for both options (e.g., option B and C) and re-calculate scores. Option 2: The course director may decide to keep the item as is (no change).	Sometimes, an item may have more than one plausible correct answer, causing confusion and unfair grading.
4. Students choose one or more response options that are totally incorrect (e.g., A and B).	For an item, most of the students selected option A and B. However, neither was correct responses. The keyed correct answer was C.	Option 1: Review the answer key to determine if one of the most commonly chosen incorrect options (either A or B) could potentially be considered correct. Then, award full credit for correct responses only, while not awarding credit for entirely wrong response options. Option 2: The course director may choose to keep the item as is (no change).	If a significant number of students choose incorrect options, it may indicate that the item was misleading or not well understood.
5. Students accumulated in two or more response options	For an item, most of the students selected option A and B. However, B was the best answer item in the answer key.	Review whether the option that is marked incorrectly in the answer key could be considered a correct response. If so, award full credit for both answers; otherwise, leave as is (i.e., the course directors will decide what the best answer(s) is/are).	When students' responses are evenly split between two options, it may suggest that both options are plausible.
6. Content was not covered.	Course directors reviewed lecture material and other course materials and realized that the topic was not covered adequately.	Option 1: Eliminate the item and give extra credit to those who answered correctly, while ensuring the maximum score remains capped at 100%. The total number of items will be the same for all students. Option 2: No grade adjustment is recommended if the question was performed well (e.g., greater than 50% correct).	If the content of an item was not adequately covered in the course material, students are unlikely to answer it correctly, leading to unfair grading.

7. An item has a correct response rate of less than 50% but greater than or equal to 30%.	For an item, the difficulty value was .45.	Option 1: Check the source of problems using scenarios #1 through #6, then apply the appropriate recommendation. Option 2: Check the item's discrimination index. If it is less than 0.10, consider removing the item and awarding credit to students who answered correctly, while ensuring the maximum score remains capped at 100%. Option 3: Review historical item properties and examine the item stem for potential confusion. Consider awarding everyone full credit if there is a problem with the item stem, while ensuring the maximum score remains capped at 100%. Option 4: If there is no problem at all, then leave as is.	Items with low correct response rates may indicate problems with the item itself or its alignment with the course content.

8. An item is less than 30% correct regardless of the discrimination parameter.	For an item, the difficulty value was .25.	Option 1: Check the source of problems using scenarios #1 through #6, then apply the appropriate recommendation. Option 2: Review historical item properties and try to identify the source of the problem. Depending on the source of the problem, consider awarding full credit to all students, while ensuring the maximum overall score remains capped at 100%. Option 3: If no problem is detected, remove the item from grading; however, award credit to those who got the item correct, while ensuring the maximum score is capped at 100%. Option 4: Check the item discrimination index. If it is less than 0.10, consider removing the item and awarding credit to those who answered it correctly, while ensuring the maximum score is capped at 100%. Option 5: The course director may want to keep the item as is (no change).	A very low correct response rate suggests significant issues with the item, such as poor wording or misalignment with the course content.

9. An item has a low discrimination parameter (point-biserial <0.10 or Discrimination Index <0.10).	For an item, the discrimination parameter was around .00 (or negative).	Option 1: Check the source of the problem using scenarios #1 through #6, then apply the appropriate recommendation. Option 2: If scenarios #1 through #6 do not apply and the item has an acceptable correct response rate (e.g., higher than or equal to 30%), no action is needed. Leave as is. Option 3: If the correct response rate is low (e.g., less than 30%), remove the item from grading and award full credit to those who answered it correctly, while ensuring the maximum score remains capped at 100%.	A low discrimination index indicates that the item does not effectively differentiate between students who understand the material and those who do not.

Operational Definitions and Thresholds

Item difficulty (p value): items with $p < 0.30$ (very difficult) and $p < 0.50$ (difficult) or $p > 0.90$ (very easy) are flagged for review. Discrimination (point biserial): items with $rpbis < 0.10$ are flagged; $rpbis \leq 0.00$ indicates a defective item requiring priority review. Distractor analysis: non functioning distractors are rarely selected (< 5%) or show positive discrimination. Thresholds are interpreted considering cohort size and content considerations (Aubin et al., 2020; Crocker & Algina,1981; Kane, 2016).

Interpretation and Theoretical Basis

These guidelines standardize decision making while allowing faculty flexibility within evidence based boundaries. By reducing subjective variability, they enhance reliability; by aligning actions with psychometric and content evidence, they strengthen the validity argument for resulting scores. Options such as “credit all” or “remove item” are reserved for cases where no defensible key exists or technical errors occur, because they can inflate scores or penalize correct responders. When content evidence supports it, key changes or crediting multiple defensible options are preferred. This approach reflects a contemporary validity argument (e.g., Messick; Kane, 2016), ensuring that scores represent standing on the intended construct (AERA, APA, NCME, 2014; Messick, 1987; Newble et al., 1981; Kane, 2016).

Practical Implications and Study Outcomes

After developing the guidelines, we conducted a “behind the scenes” pilot during the subsequent exam review cycles. This pilot involved applying the guidelines to post-exam item analysis without formally announcing the changes to faculty or students. Specifically, we flagged problematic items according to the new criteria and used the guidelines to recommend actionable steps for adjustment. During this period, decisions were tracked and compared to previous ad hoc practices to assess consistency and reduction in the number of post-exam adjustments. Based on these findings, we iteratively refined the guidelines, adding further sources of item issues and corresponding recommendations to cover a broader range of scenarios. Implementation was monitored by recording the number and types of item-level adjustments made each semester. For example, in Fall 2023, there were 68 adjustments across pre-clerkship exams. After guideline refinement and approval from the curriculum and assessment committees, this number dropped to 39 in Fall 2024—a 42.65% reduction. This process demonstrated improved item alignment, greater consistency in decision-making, and a more structured, transparent post-exam review process. Beyond reducing the number of adjustments, the guidelines improved consistency in decision-making. Previously, similar scenarios could lead to different actions across courses or tracks (e.g., full credit in one course vs. key change in another). The standardized framework now anchors decisions in psychometric evidence and content review, reducing subjective variability and supporting reliability. While the guidelines allow flexibility for context-specific judgment, they prioritize actions that preserve validity (e.g., key changes or crediting multiple defensible options) over those that distort score meaning (e.g., credit all).

Faculty feedback indicated increased awareness of fairness and defensibility in grading, and students expressed greater confidence in the transparency of the process. Although these guidelines were developed for one institution, they address a common challenge in health professions education and can serve as a model for adaptation elsewhere.

Conclusion Remarks

The development and implementation of standardized post-exam adjustment guidelines were driven by the goals of fairness, accuracy, consistency, and transparency. By reducing arbitrary decisions and anchoring adjustments in psychometric and content evidence, the guidelines strengthen the validity argument for resulting scores and enhance reliability.

Recurring item-level issues also provide actionable insights for curriculum improvement. Items with low discrimination or extreme difficulty often signal misalignment between instruction and assessment, prompting targeted revisions to content and teaching strategies. Thus, post-exam analysis serves not only as a quality control mechanism but also as a driver for continuous improvement in assessment and instruction.

While automation was beyond the scope of this evaluation, emerging technologies such as AI could complement these guidelines by streamlining item analysis and flagging problematic items in real time. Future research should explore how AI-driven tools can integrate with evidence-based frameworks to further enhance fairness and efficiency, while maintaining faculty oversight.

Acknowledgements

We thank all faculty members from the Texas A&M University Naresh K. Vashisht College of Medicine for their valuable feedback and contributions to this work.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Aubin, A. S., Young, M., Eva, K., & St-Onge, C. (2020). Examinee cohort size and item analysis guidelines for health professions education programs: A Monte Carlo simulation study. *Academic Medicine, 95*(1), 151–156. https://doi.org/10.1097/ACM.0000000000002888

Crocker, L., & Algina, J. (1981). Introduction to classical and modern test theory. Philadelphia, PA: Harcourt Brace Jovanovich College Publishers.

Fullerton, J. T. (1993). Evaluation of research studies. Part IV: Validity and reliability—concepts and application. *Journal of Nurse-Midwifery, 38*(2), 121–125. https://doi.org/10.1016/0091-2182(93)90146-8

Ginzburg, S. B., Sein, A. S., Amiel, J. M., Auerbach, L., Cassese, T., Konopasek, L., Ludwig, AB., Meholli, M, Ovitsh, R., & Brenner, J. (2025). An examination of grade appeals via a root cause analysis. *Academic Medicine, 100*(6), 662–672. https://doi.org/10.1097/ACM.0000000000006000

Kane, M. T. (2016). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy & Practice, 23*(2), 309–311. https://doi.org/10.1080/0969594x.2016.115664

Menon, B., Miller, J., & DeShetler, L. M. (2021). Questioning the questions: Methods used by medical schools to review internal assessment items. *MedEdPublish*, 10, 37. https://doi.org/10.15694/mep.2021.000037.1

Messick, S. (1987). Validity. *ETS Research Report Series, 1987*(i), 1–208. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x

Newble, D. I., Hoare, J., & Elmslie, R. G. (1981). The validity and reliability of a new examination of the clinical competence of medical students. *Medical Education, 15*(1), 46–52. https://doi.org/10.1111/j.1365-2923.1981.tb02315.x

Raymond, J., Dai, D. W., & McAllister, S. (2025). The interpretation-use argument–the essential ingredient for high quality assessment design and validation. *Advances in Health Sciences Education, 30*(4), 1313-1332. https://doi.org/10.1007/s10459-024-10392-6

Reckase, M. (2023). The psychometrics of standard setting: Connecting policy and test scores. Boca Raton, FL: CRC Press.

About the Authors

Halil Sari, Director of Evaluation and Assessment, Instructional Assistant Professor, Texas A&M University, hisari@tamu.edu

Vinayak Shenoy, Assistant Dean of Evaluation and Assessment, Instructional Associate Professor, Texas A&M University, vshenoy@tamu.edu

Vincent VanBuren, Instructional Associate Professor, Texas A&M University, vanburen@tamu.edu

James Dubuisson, Program Manager, Texas A&M University, jdubuisson@tamu.edu

Danielle Dickey, Senior Associate Dean of Academic Affairs, Instructional Associate Professor, Texas A&M University, dmdickey@tamu.edu

Jody Johnston, Program Manager, Texas A&M University, jping@tamu.edu