Ensuring the Quality of Multiple-Choice Tests: An Algorithm to Facilitate Decision Making for Difficult Questions
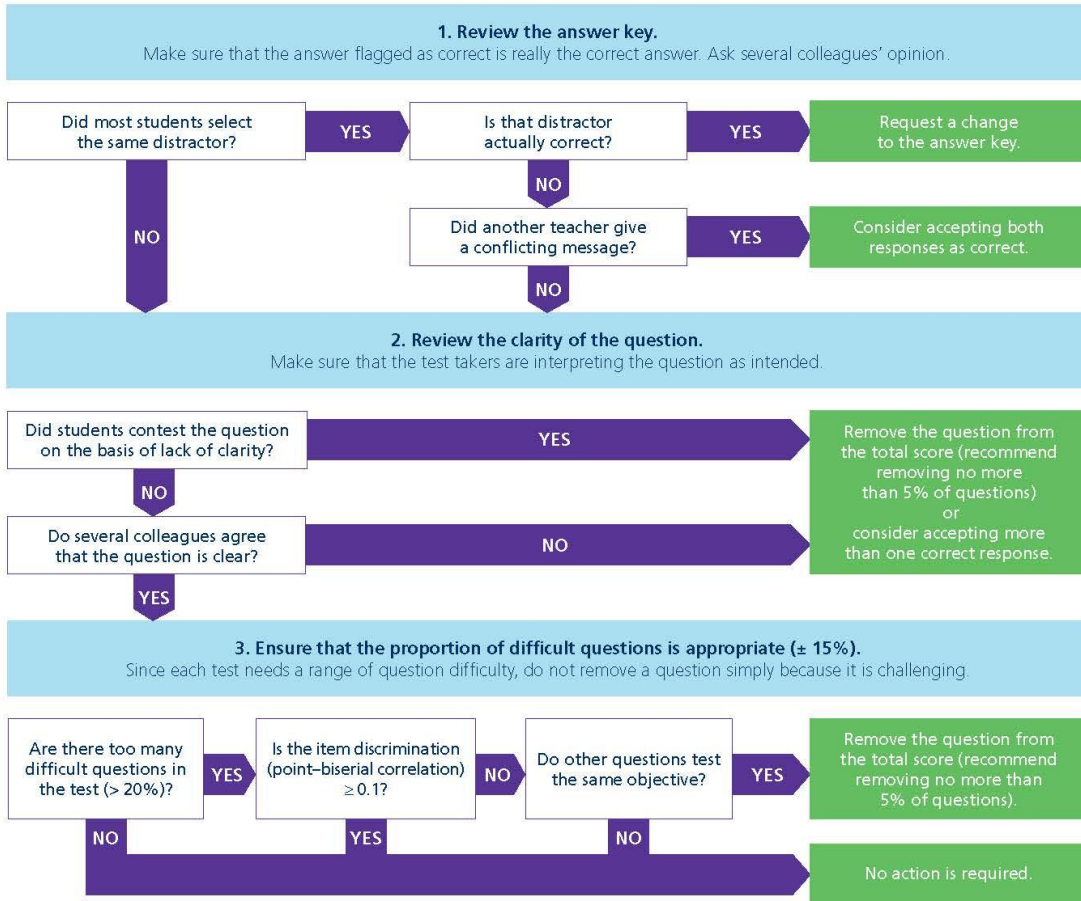
Valérie Dory, MD, MMedEd, PhD, assistant professor, Department of Medicine and Centre for Medical Education, Kate Allan, assessment administrator, Undergraduate Medical Education, Leora Birnbaum, MD, MHPE, assistant professor, Department of Medicine, Stuart Lubarsky, MD, MHPE, assistant professor, Department of Neurology & Neurosurgery and Centre for Medical Education, Joyce Pickering, MD, MSc, associate professor, Department of Medicine and Centre for Medical Education, and Meredith Young, PhD, associate professor, Department of Medicine and Centre for Medical Education, McGill University—All authors are members of the Student Assessment Subcommittee, Undergraduate Medical Education, McGill University

Processes for ensuring the quality of multiple-choice question (MCQ)-based tests occur before, during, and after test administration.

BEFORE THE TEST
Faculty development, policies, and item-vetting processes ensure the development of a good test

DURING THE TEST
Policies and processes ensure smooth and fair administration of the test

AFTER THE TEST
Policies, guidelines, and decision algorithms ensure good decisions, including regarding test rescoring

Despite the existence of guidelines, posttest analysis requires judgment and can be particularly challenging.[1-4] The Student Assessment Subcommittee (Undergraduate Medical Education, McGill University) developed user-friendly decision algorithms to support instructors in interpreting item difficulty and discrimination indices. Here we present our approach for what to do with **difficult questions** (i.e., items that less than 30% of students have answered correctly). In our experience these flowcharts—which can be adapted according to local needs and procedures—empower instructors to use item analysis reports judiciously to reach transparent and defensible decisions regarding examination scoring.

Decision Algorithm for Difficult Questions

1. Review the answer key.
Make sure that the answer flagged as correct is really the correct answer. Ask several colleagues' opinion.

- Did most students select the same distractor? → YES → Is that distractor actually correct? → YES → Request a change to the answer key.
- Is that distractor actually correct? → NO → Did another teacher give a conflicting message? → YES → Consider accepting both responses as correct.

2. Review the clarity of the question.
Make sure that the test takers are interpreting the question as intended.

- Did students contest the question on the basis of lack of clarity? → YES → Remove the question from the total score (recommend removing no more than 5% of questions) or consider accepting more than one correct response.
- Do several colleagues agree that the question is clear? → NO → Remove the question from the total score (recommend removing no more than 5% of questions) or consider accepting more than one correct response.

3. Ensure that the proportion of difficult questions is appropriate (± 15%).
Since each test needs a range of question difficulty, do not remove a question simply because it is challenging.

- Are there too many difficult questions in the test (> 20%)? → YES → Is the item discrimination (point–biserial correlation) ≥ 0.1? → NO → Do other questions test the same objective? → YES → Remove the question from the total score (recommend removing no more than 5% of questions).
- No action is required.

References:
1. Schmeiser CB, Welch CJ. Test development. In: Brennan RL, ed. Educational Measurement. 4th ed. Westport, CT: Praeger Publishers; 2006:307-353.
2. Tavakol M, Dennick R. Postexamination analysis: A means of improving the exam cycle. Acad Med. 2016;91:1324.
3. Tavakol M, Dennick R. Post-examination analysis of objective tests. Med Teach. 2011;33:447-458.
4. Schuwirth LW, van der Vleuten CP. A plea for new psychometric models in educational assessment. Med Educ. 2006;40:296-300.

Author contact: valerie.dory@gmail.com; **Twitter:** @valeriedory1