Guidelines for reviewing multiple-choice questions on UGME examinations

Timothy J. Wood

March 2, 2020

Special thanks to Dr. Michelle Anawati, Dr. Isabelle Burnier, Dr. Heather MacLean, Dr. Laurie McLean, Dr. Chris Ramnanan for providing comments on earlier versions of this document.

Questions or suggestions for improvements can be sent to Tim Wood (twood@uottawa.ca).

Questions on an examination can perform poorly due to a number of reasons, including typographical errors, language issues, poor writing or even testing topics that were not covered in the course. After an examination, it is common practice to review the performance of questions and identify any that did not perform well. Identifying these questions and trying to resolve why they performed poorly helps ensure students' scores are accurate reflections of their skills and knowledge.

The purpose of this document is to provide guidance to unit leaders and content experts when they are reviewing the performance of multiple-choice questions that have been flagged as poor performing questions on an examination. The focus will be on the statistical indicators that are used to identify poor performing questions and on how that information should be used by unit leaders and content experts when determining what should be done with a flagged question.

Statistical indices used to flag poor performing questions

Table 1 displays an example of a multiple-choice question report that is provided to unit leaders and content experts after an examination. Within this table are two statistical indices that are used to flag poor performing questions. The first index is difficulty (highlighted in yellow), which captures how many students answered the question correctly. This index is reported as a proportion correct. For example, for question 1, 0.75 or 75% of the students answered the question correctly. For question 2, 0.96 or 96% of the students answered the question correctly. If a difficulty index is less than 0.30, then the question would be considered very difficult. If the difficulty index is greater than 0.95, then the question would be considered very easy.

Table 1. Example of the statistical information generated for multiple-choice questions

Question	Answer	Difficulty	Biserial	Point-biserial	Discrimination
1	E	0.75	0.45	0.38	0.47
2	D	0.96	0.11	0.05	0.08
3	C	0.25	0.22	0.15	0.12
4	E	0.25	0.12	0.07	0.07
5	D	0.25	-0.10	-0.12	-0.11

The second index displayed in Table 1 is discrimination, which captures the degree to which a question separates lower-scoring from higher-scoring students on that examination. There are actually two values displayed in the table that serve this purpose. The first value is the point-biserial correlation (highlighted in green). This correlation measures the relationship between students' score on the question and their total score. The second value is labelled discrimination (highlighted in grey). To generate this value, students with overall exam scores that fall in the top 25% and bottom 25% are identified. For any given question, the number of lower-scoring students who answered the question correctly is subtracted from the number of higher-scoring students who answered the question correctly and then divided by the total number of students in these two groups. Although the biserial correlation

is provided in the report, it is not used because multiple-choice questions do not meet the assumptions of this correlation.

Either the point-biserial or the discrimination value can be used as the measure of how well a question is discriminating between strong and weak students' scores and it is personal preference as to which is used. Both produce similar results and both values are interpreted in the same manner. If the discrimination index is above 0.10, it indicates that students with higher overall exam scores tended to answer the question correctly and students with lower examination scores tended to answer the question incorrectly. This would be considered the optimal level for a question. If the discrimination index falls between 0 and 0.09, then it suggests there is little difference in how lower-scoring and higher-scoring students performed on that question and it could indicate a potential problem with the question. If the discrimination index is ever negative, then it indicates that students with low scores on the examination answered the question correctly whereas students with higher scores answered the question incorrectly. A question with a negative discrimination index is always flagged as a poor-performing question. In the example displayed in Table 1, question 5 has a negative point-biserial correlation and therefore would be automatically flagged.

Difficulty and discrimination indices also interact. If a question is very difficult, then many higher-scoring students will answer incorrectly, and the discrimination index will tend to be low. If a question is very easy, then even lower-scoring students will answer correctly, and the discrimination index will also tend to be low. Question 2 in Table 1 illustrates this interaction. The difficulty index is high (0.96), and therefore the point-biserial correlation is less than 0.10, which indicates that even the lower-scoring students were answering the question correctly because it was relatively easy.

What to look for when reviewing a flagged question?

Flagging poor performing questions based on statistics is the first step in determining if a question performed poorly. All flagged questions should be reviewed because there may be an explanation for its statistical results. At this stage, it might be helpful to also review the feedback from students. This feedback could provide an explanation for the statistics that was not obvious to the person reviewing the question.

The following guidelines list the information that should be considered when reviewing a question.

1. If a question is difficult (proportion correct less than 0.30):
    - check the answer key to ensure the correct answer has been properly identified for scoring;
    - read the question to confirm that the correct answer is listed and, if so, check its accuracy and if any of the other answers would apply;
    - read the question to ensure there are no clues (e.g., typos, grammar) or other ambiguities that would encourage students to choose a wrong answer;
    - confirm that the question is testing an objective that was covered in the course;
    - check to ensure that the English and French versions of the question are equivalent; and
    - check if the students identified any issues with this question.

    If the above checks are all negative, then look at the discrimination index. If it is above 0.10, then the question is difficult, but higher-scoring students tended to answer correctly and therefore the question is likely acceptable. If the discrimination index is between 0 and 0.09,

then there is likely something wrong with the question and it may need to be removed from scoring and/or sent to a content expert who will revise it before it is used again.

2. If a question is extremely easy (difficulty index above 0.95):
   - read the question to ensure there are no clues (e.g., typos, grammar) that would identify the correct answer.

   If the above check does not indicate an obvious reason for the question being easy, then it is likely acceptable to keep it as a scored item, but it should be sent to a content expert to revise before it is used again.

3. If the discrimination index is negative:
   - check the answer key to ensure the correct answer has been properly identified for scoring;
   - read the question to confirm that the correct answer is listed and, if so, check its accuracy and if any of the other answers would apply;
   - read the question to ensure there are no clues (e.g., typos, grammar) or other ambiguities that would encourage higher-scoring students to choose an incorrect answer;
   - confirm that the question is testing an objective that was covered in the course;
   - check to ensure that the English and French versions of the question are equivalent; and
   - check if the students identified any issues with this question.

   If the above checks are all negative, then there is something else wrong with the question. Under rare circumstances, a question with a negative discrimination index could be kept on the examination (e.g., the value was close to 0 with a small number of students tested), but in most cases the question should be removed from scoring and sent to a content expert who will revise it before it is used again.

What decisions should be made regarding a question?

There are four options available when reviewing a question that was flagged as poor performing.

1. Delete the question
   Deleting a question removes it from scoring so that even if a student answers correctly, they will not get credit for it. This option is used if a) there is no correct answer in the question (i.e., all answers are incorrect) or b) the question is testing an objective that was not covered in the course. The deleted question should be sent back to the content expert to either improve it or remove it from the bank.
2. Credit the question
   Crediting a question involves removing the question from the total score, but students who answered the question correctly still get a point. This option is used when the review has identified an issue with the question (e.g., typos, ambiguity, translation issue). Credited questions should be returned to a content expert for editing in order to address any issues before they are used again.
3. Change the answer key

On occasion, the answer that was identified as correct does not apply, but another answer does. In this case, the answer should be changed, and the examination should be rescored using the correct answer key. Note that the person revising the question may decide that there is more than one correct answer in which case, the exam should be rescored using all identified correct answers. In the latter case, the question should be sent to the content expert who will address the issue of having more than one correct answer before it is used again.

4. Keep the question on the examination but send it back to the content expert for editing
   This option is used when an issue has been identified with a question, but it is not deemed a major question flaw.
5. Keep the question on the examination
   This option is used when no content or structural issues with the question are identified and when the question is relevant to the course.

Examples of common statistical patterns and what should be done in each case

The following examples display common situations that could occur with questions and serve to help illustrate how the statistics should be interpreted. All examples refer to the questions and statistics displayed in Table 1.

Question 1: difficulty index = 0.75, discrimination index (point-biserial) = 0.38
Because the difficulty index is between 0.30 and 0.95 and the discrimination index is above 0.10, this would be considered a question that is performing well and would not be flagged as poor performing.

Question 2: difficulty index = 0.96, discrimination index (point-biserial) = 0.05
The difficulty for this question is above 0.95, and therefore this would be considered a very easy question. The discrimination index is less than 0.10, which indicates that there was little difference in how higher-scoring and lower-scoring students answered the question. Review the question to ensure there wasn't a typo or other clue as to the correct answer. If there weren't any, then leave it on the examination as a scored item but send it to a content expert to revise before it is used again.

Question 3: difficulty index = 0.25, discrimination index (point-biserial) = 0.15
This would be considered a very difficult question but, because the discrimination index is above 0.10, it indicates that higher-scoring students were performing better on this question than lower-scoring students. Make sure the key is correct and check if another option is being frequently chosen by students. Check the wording of the question to ensure it is well written and check the translation to determine if there is an issue in the language. Confirm that the question is testing an objective that was covered in the course. If there are no obvious problems that would account for the low difficulty index, this question could be kept on the examination.

Question 4: difficulty index = 0.25, discrimination index (point-biserial) = 0.07
This would be considered a very difficult question. Because the discrimination is between 0 and 0.10, it indicates that higher-scoring students are answering the same as the lower-scoring students, and it could indicate that there is a problem. Make sure the key is correct and check if another option is being frequently chosen by students. Check the wording of the question to ensure it is well written and check

the translation to determine if there is an issue in the language. Confirm that the question is testing an objective that was covered in the course. Check the students' feedback to see there were any comments. If no problem is found, then this question could be kept on the examination if it is deemed to be crucial, otherwise consider crediting it.

Question 5: difficulty index = 0.25, discrimination index (point-biserial) = -0.12
This is a question that was either scored with the wrong answer key, did not have a correct answer, tests an objective that was not covered in the course, is poorly written, or had a translation issue. If it was scored with the wrong answer key, then it should be rescored using the right answer. In other cases, this question should either be deleted or credited.