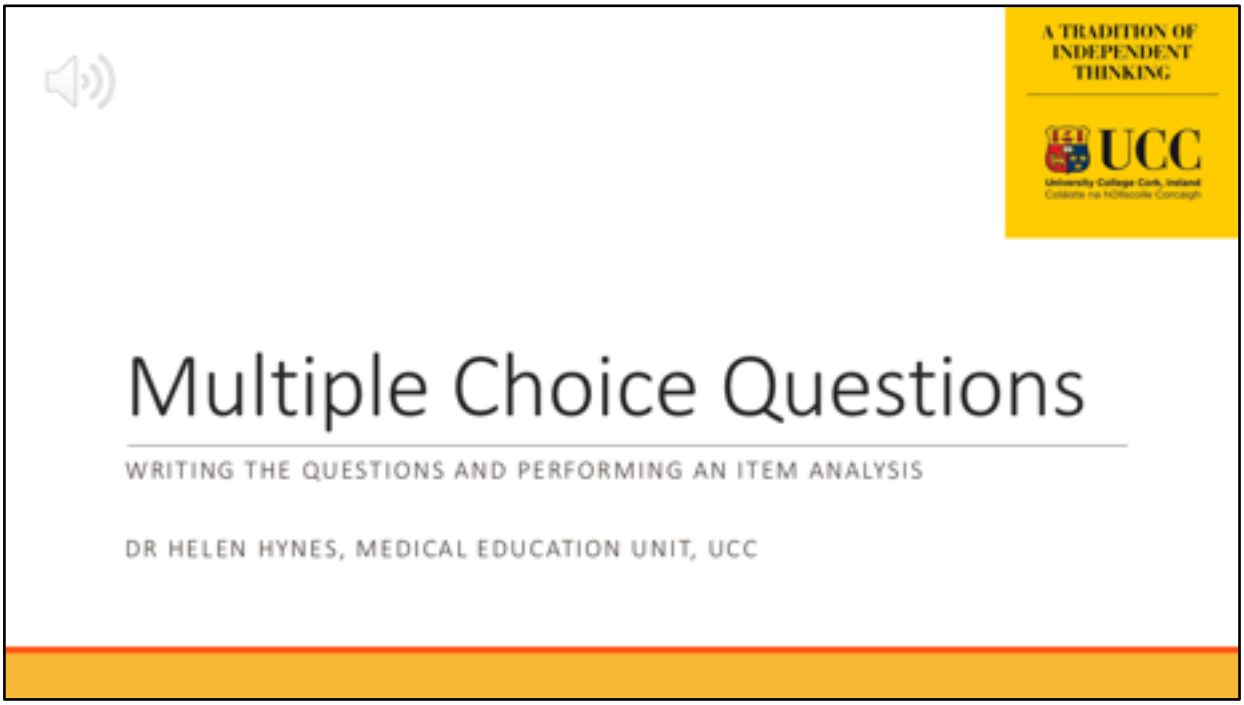
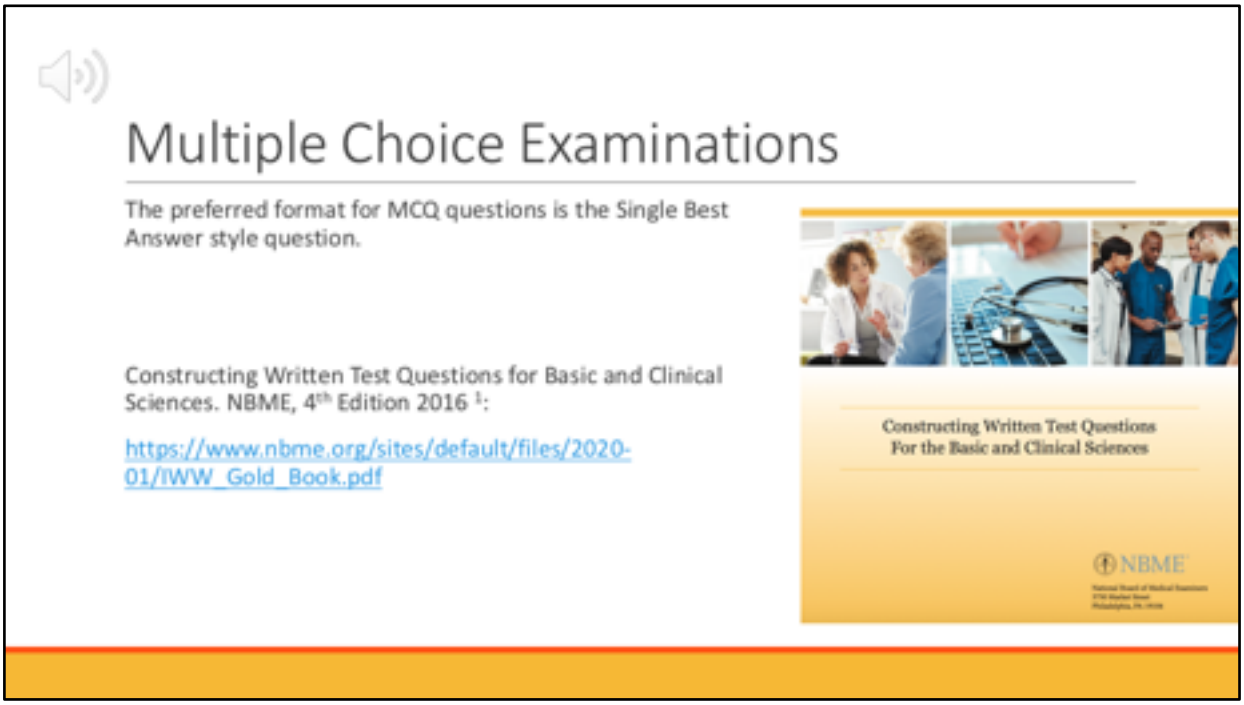


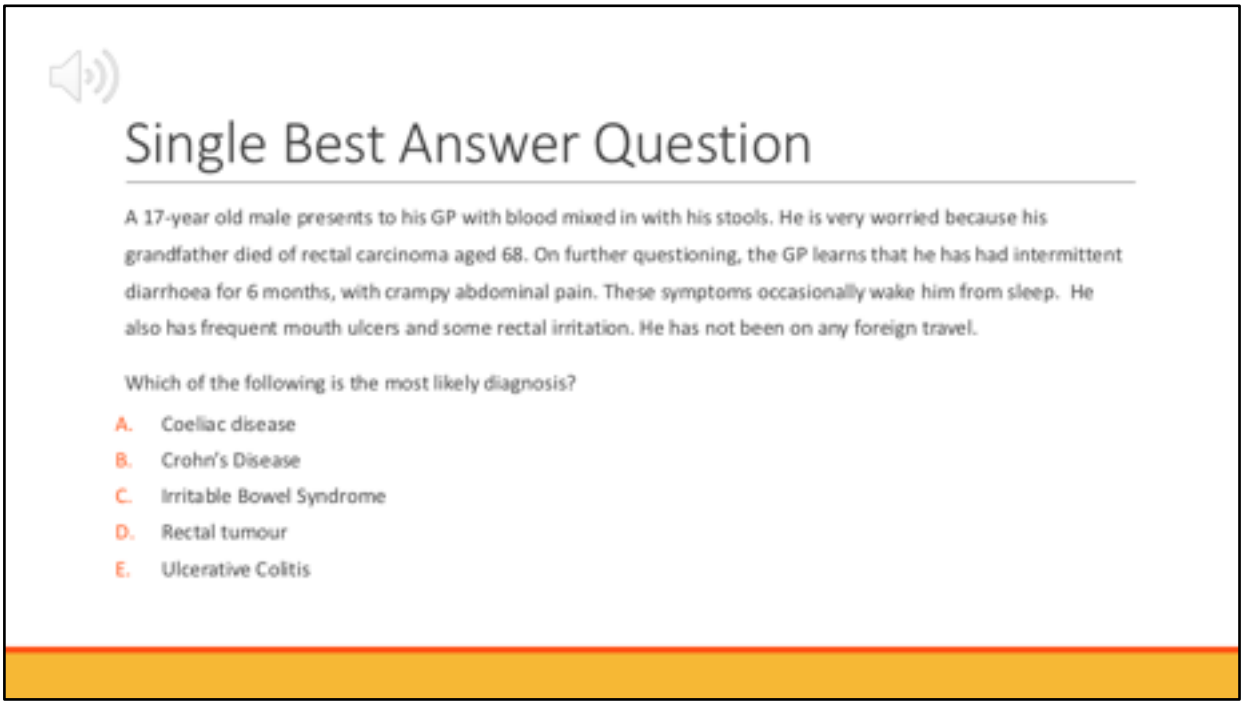
This presentation, prepared by the Medical Education Unit at University College Cork focuses on how to write good multiple choice questions, and how to carry out a post hoc item analysis to see how well the questions performed.



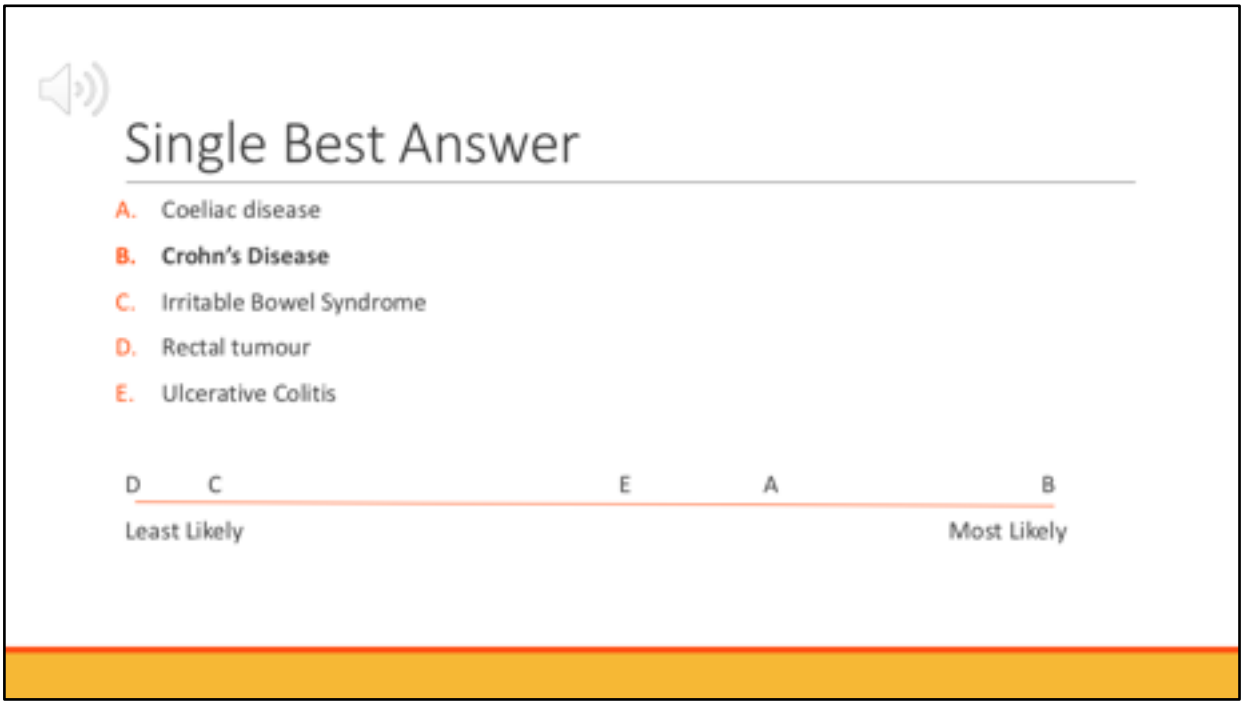
Multiple different formats of Multiple Choice Questions have been used in the past, but the recommended type for use in medical education is the single best answer format.

The American National Board of Medical Examiners published a guide on constructing written test questions. Based on over 30 years of experience with writing and evaluation of MCQ items, the NBME now recommends avoiding True / False style questions completely and instead uses single best answer style questions in its examination.

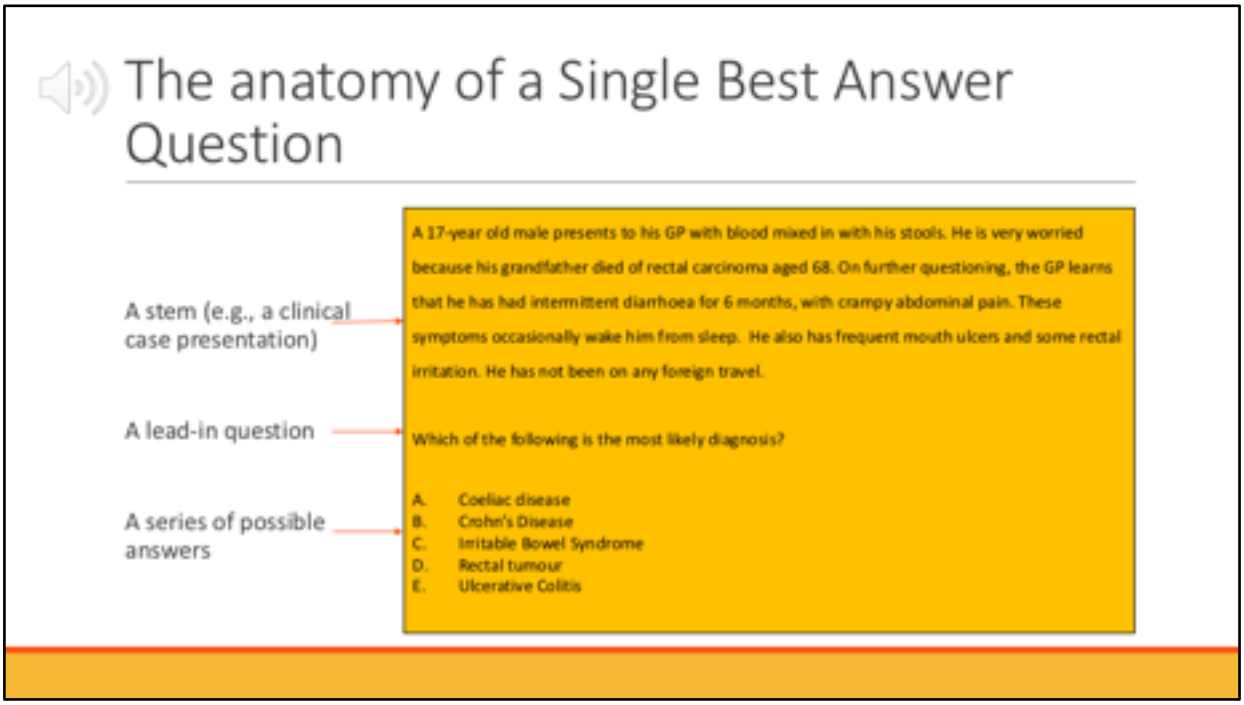
You can access the NMBE guide [1] by clicking on the link on this slide.



This is an example of a single best answer style question. All of the answers are feasible but some are less likely than others and 1 answer is definitely the best.

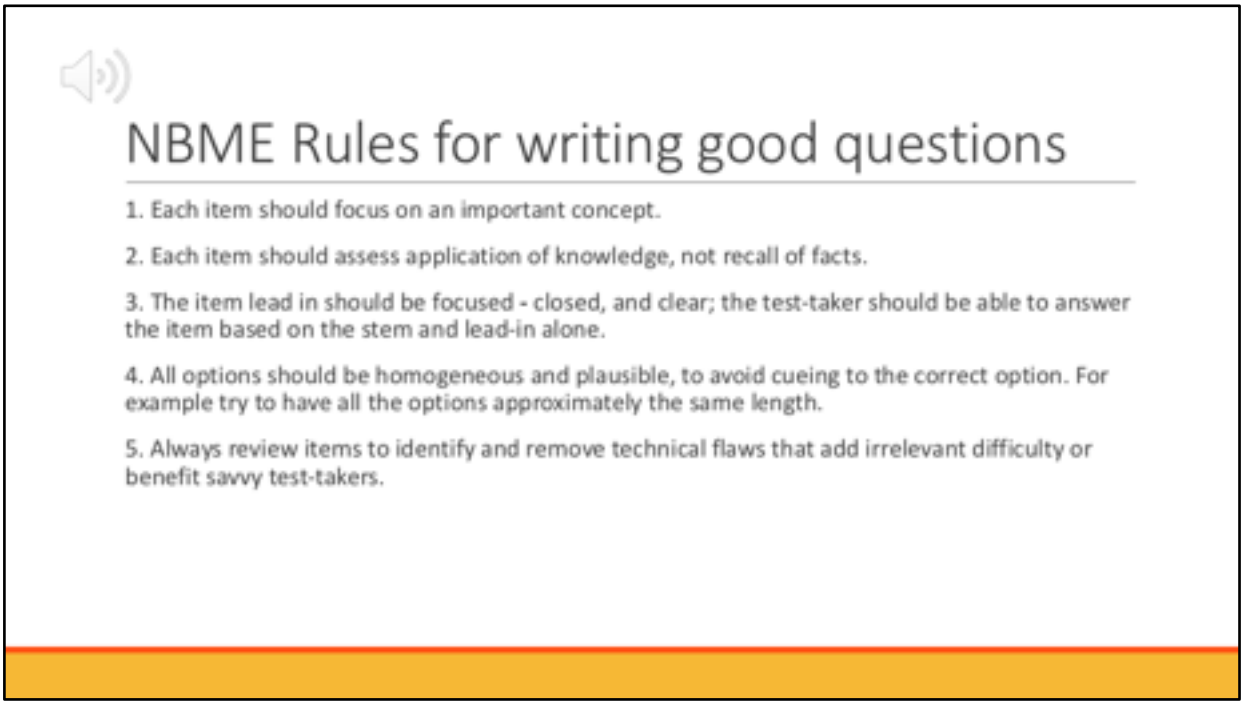


Considering the question on the last slide, we can diagram the possible options as shown here – with the most likely option, given the patient's symptoms and age, being Crohn's disease, and the remaining 4 options being possible but less likely when all the details in the question stem are taken into consideration.



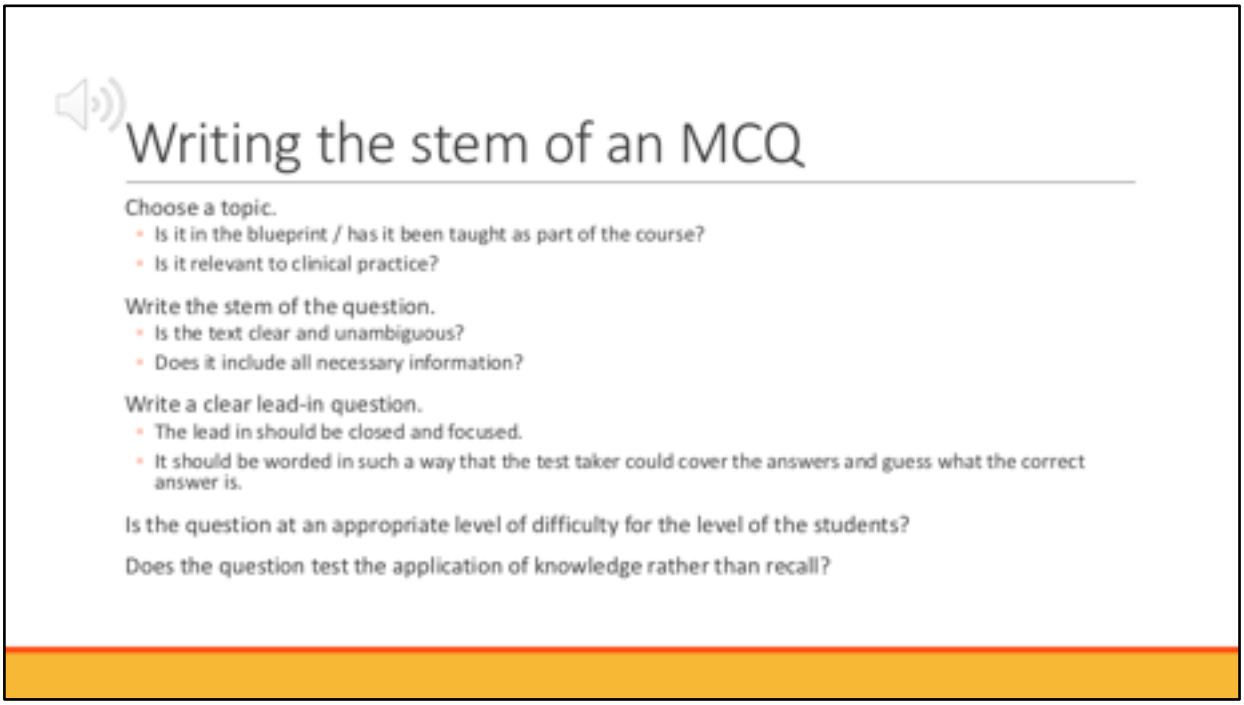
This is the what a single best answer style question should look like:

- There is a long stem containing clinical information.
- There is a short but clear lead in question.
- Then there are a number of possible answers from which the candidate must pick the best.



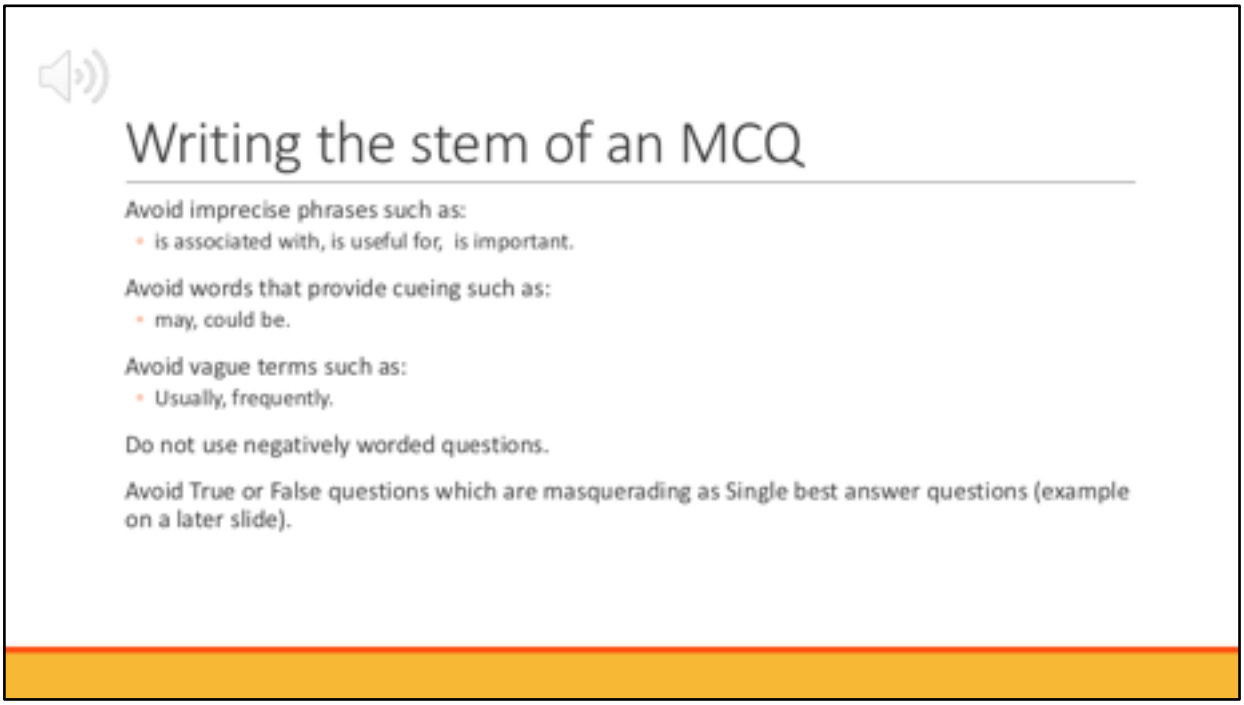
The National Board of Medical Examiners suggests a series of rules for writing good MCQ questions. These are:

1. Each item should focus on an important concept;
2. Each item should assess application of knowledge, not recall of facts;
3. The item lead in should be focused - closed, and clear; the test-taker should be able to answer the item based on the stem and lead-in alone;
4. All options should be homogeneous and plausible, to avoid cueing to the correct option. For example try to have all the options approximately the same length; and
5. Always review items to identify and remove technical flaws that add irrelevant difficulty or benefit savvy test-takers.



When writing an MCQ question it is useful to follow these steps:

- Choose a topic.
    - Is it in the blueprint?
    - Is it relevant to clinical practice?
- Write the stem of the question.
    - Is the text clear and unambiguous?
    - Does it include all the information that the candidate would need to figure out the answer?
- Next write a clear lead-in question.
    - The lead in should be closed and focused.
    - It should be worded in such a way that the test taker could cover the answers and guess what the correct answer is.
- Is the question at an appropriate level of difficulty for the level of the students?
- Does the question test the application of knowledge rather than recall?



When writing the stem of an MCQ, avoid imprecise phrases such as:

- is associated with;
- is useful for;
- is important.

Avoid words that provide cueing such as:

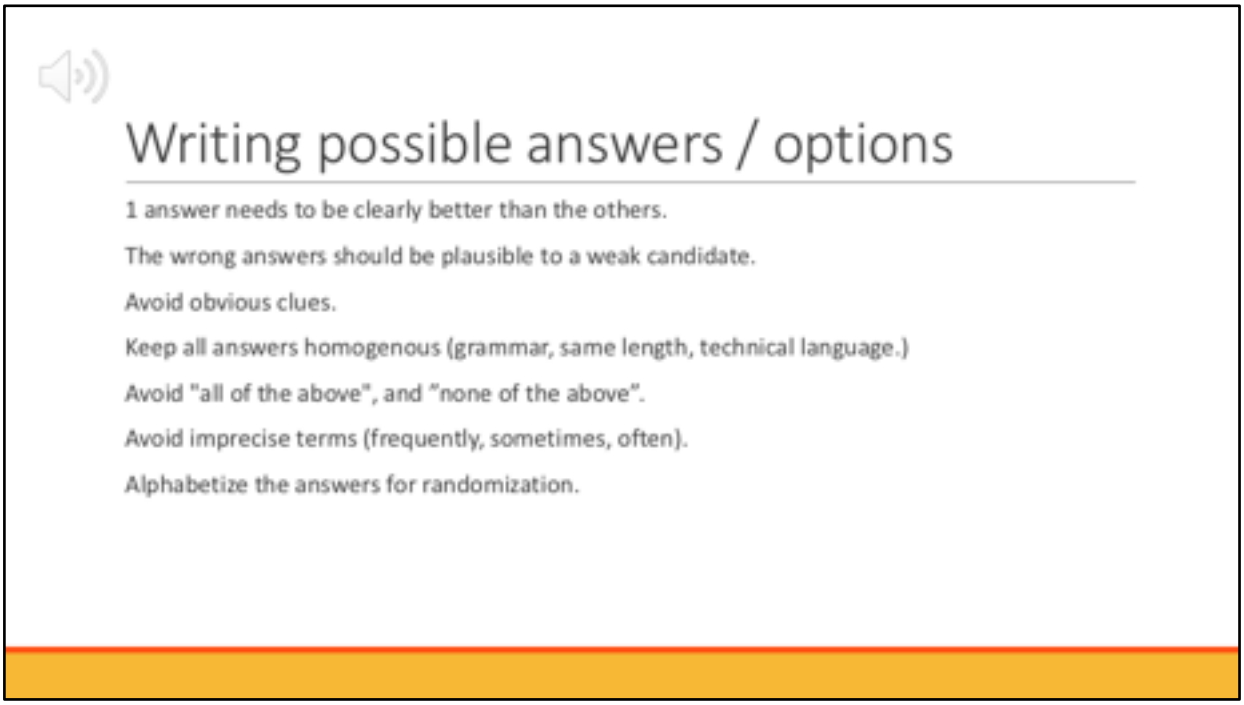
- May;
- could be.

Avoid vague terms such as:

- Usually;
- frequently

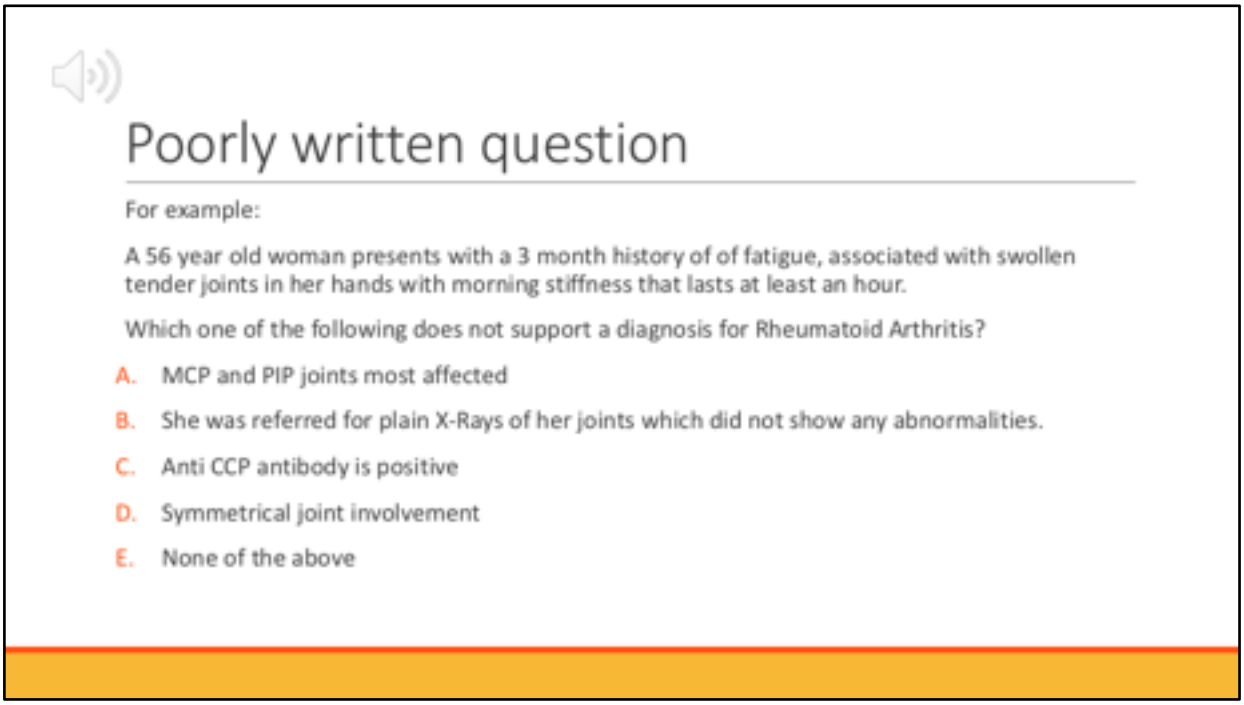
Do not use negatively worded questions.

Avoid True or False questions which are masquerading as Single Best Answer style questions (an example of this will be shown in a later slide).



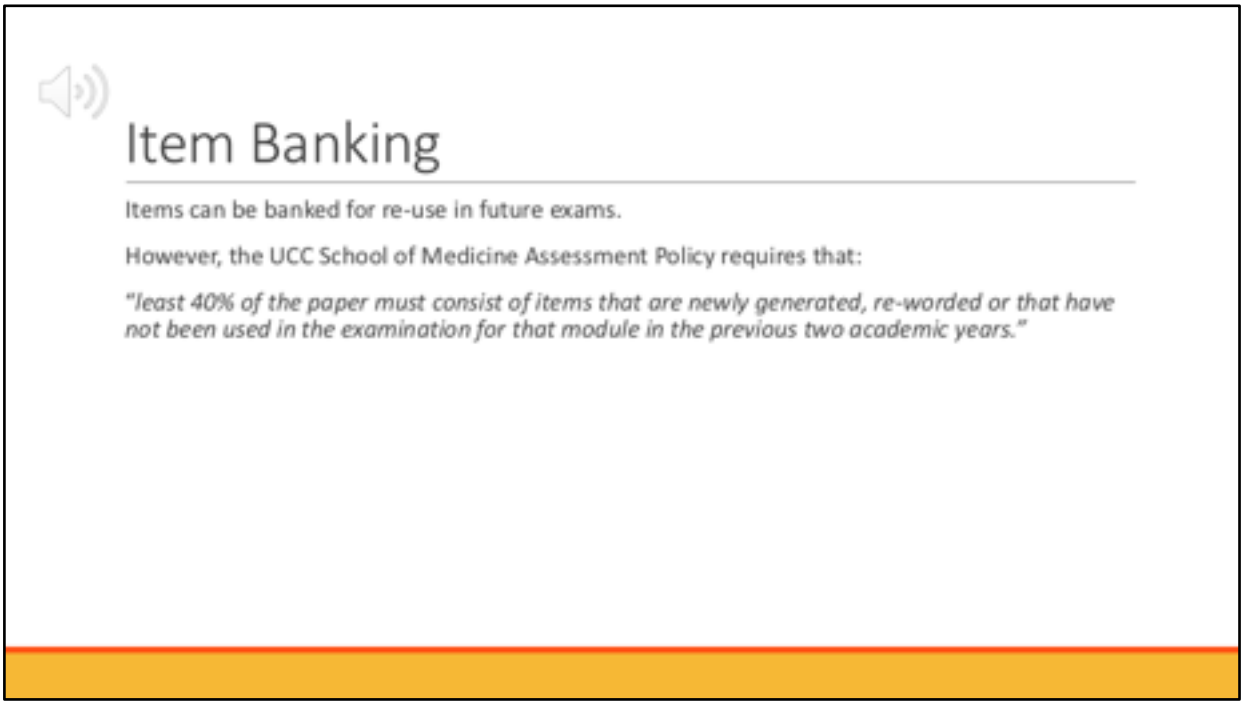
When writing a list of possible answers:

- 1 answer needs to be clearly better than the others but the wrong answers should be plausible to a weak candidate.
- Avoid obvious clues or anything that might allow the candidate automatically eliminate some of the answers.
- Keep all the answers homogenous (grammar, same length, technical language).
- Avoid "all of the above", "all of the following except", and "none of the above".
- Avoid imprecise terms (frequently, sometimes, often).
- We usually alphabetize the answers for randomization purposes.



This is a poorly written question for multiple reasons.

- Firstly it is a negative question –"Which of the following does not support a diagnosis of RA?" Negatively worded questions should not be used.
- Next it is actually a true or false type question which is masquerading as a Single Best Answer style question. 4 of the answers here are wrong. In a true single best answers questions all of the answers should be plausible to some degree, but 1 answer should be clearly better than the rest.
- Next the answers are not homogenous – some relate to clinical signs and others to investigations. The answers or options should all be homogenous to avoid giving the students cues which might help them to eliminate some of the possible answers.
- One of the answers here is significantly longer than the others. Again this can cause cueing which might help the test wise student to correctly guess the answer.
- One of the options here is "None of the above" which should be avoided.
- Finally the answers should be listed in alphabetical order to properly randomise them.

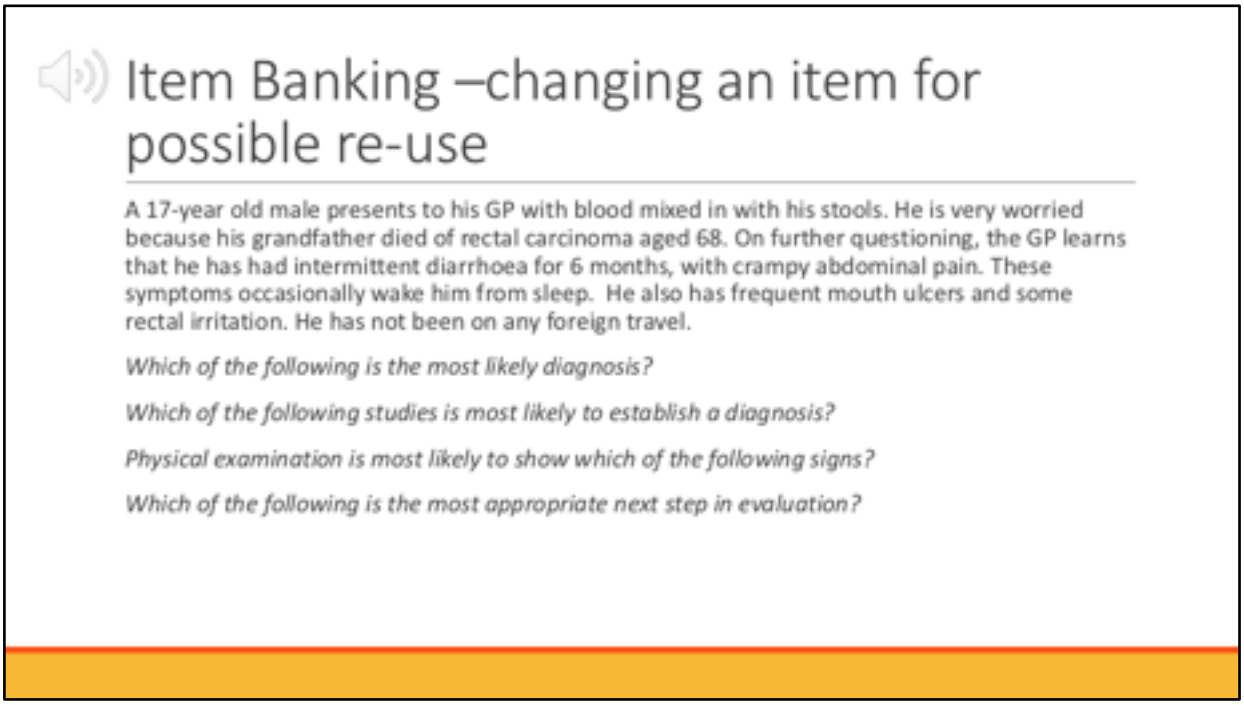


We have described how to write good questions.

Ideally you will have a bank of good questions that you can select from. However, we recommend that at least 40% of the questions in your paper should not have been used in the past 2 years.

This will mean writing new questions each year, and also reviewing and editing old questions so that they can be used again in a new format.

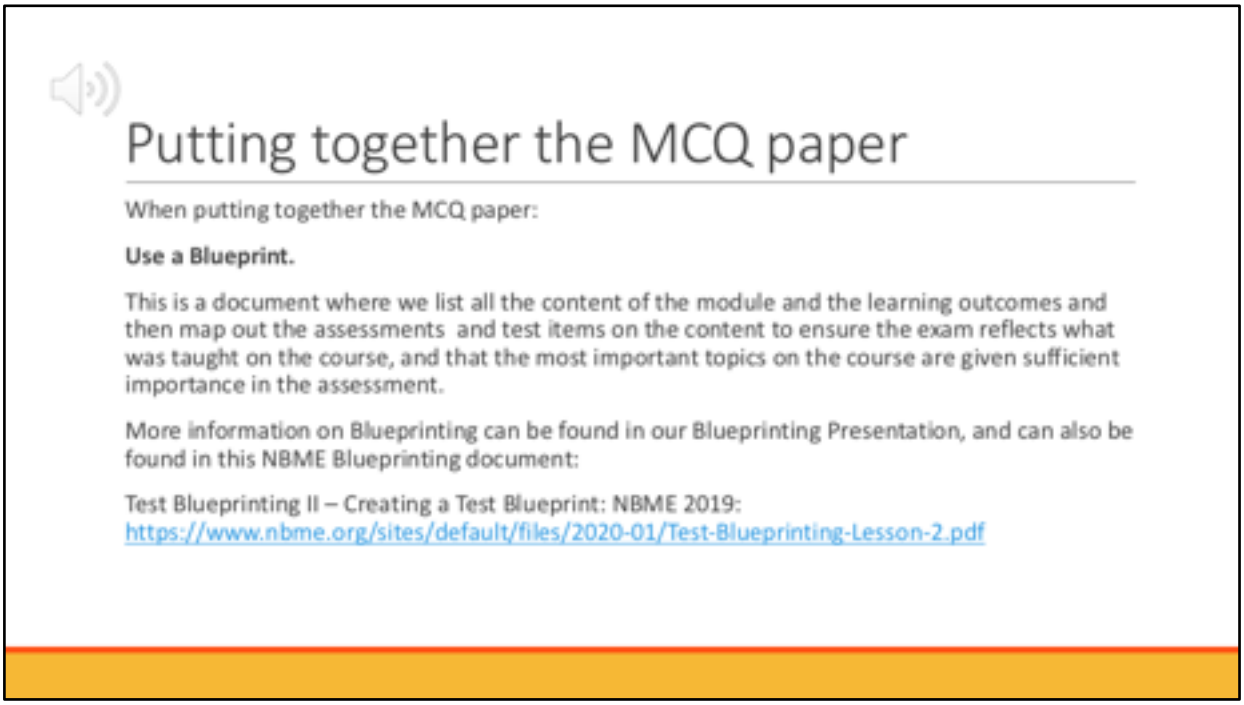
The NBME guide already referenced is an excellent guide for writing questions and editing old questions so that an old stem can be used in new questions. For example we can re-purpose an old stem by changing the focus of the lead in question and answers as we see on the next slide.



So using the sample question that we examined at the beginning of the presentation, we can re-use an existing stem and change the lead in question and the list of possible options in many different ways as shown here.

The lead in question for this stem could be any of the following:

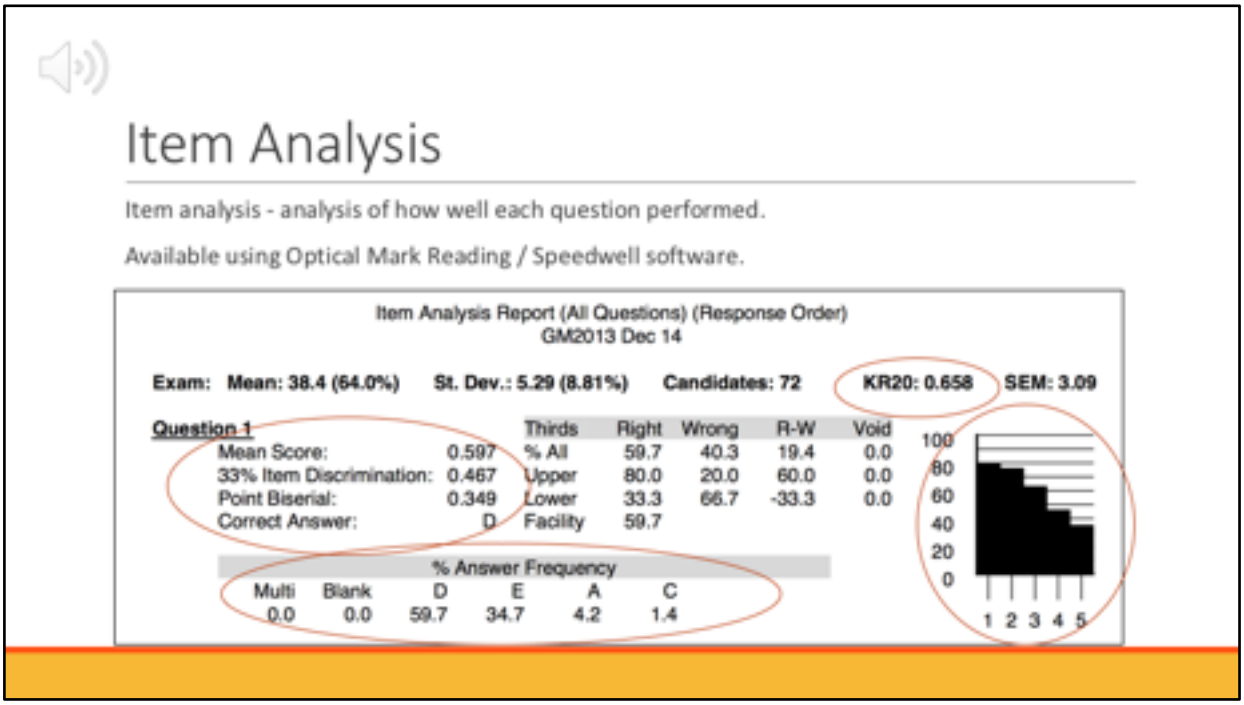
- *Which of the following is the most likely diagnosis?*
- *Which of the following studies is most likely to establish a diagnosis?*
- *Physical examination is most likely to show which of the following signs?*
- *Which of the following is the most appropriate next step in evaluation?*



When putting together the MCQ paper, use a Blueprint.

This is a document where we list all the content of the module and the learning outcomes and then map out the assessments and test items on the content to ensure the exam reflects what was taught on the course, and that the most important topics on the course are given sufficient importance in the assessment.

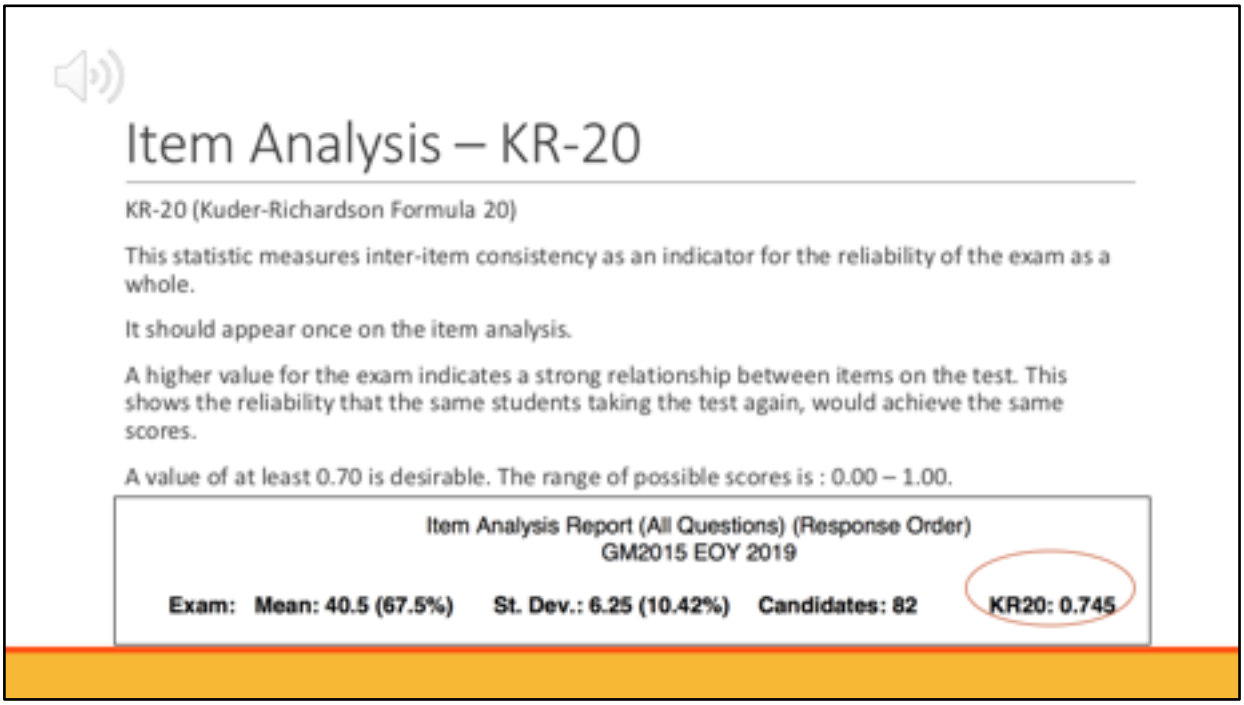
More information on Blueprinting can be found in our Blueprinting Presentation, and can also be found in this NBME Blueprinting document:
Test Blueprinting II – Creating a Test Blueprint: NBME 2019:
https://www.nbme.org/sites/default/files/2020-01/Test-Blueprinting-Lesson-2.pdf [2].



After the exam has been administered to the students, the module coordinator should carry out an item analysis (analysis of how well each question performed.)

This is available using the Optical Mark Reading software used by the university and also in the Speedwell software systems used by some modules in the School of Medicine.

The parameters we are going to consider here are KR20 Score, Mean Score, 33% item discrimination, point Bi-serial and also the graph shown on the right hand side of the image. We can also see the correct answer, which in this case was D and the frequency with which students selected each of the 5 possible answers.



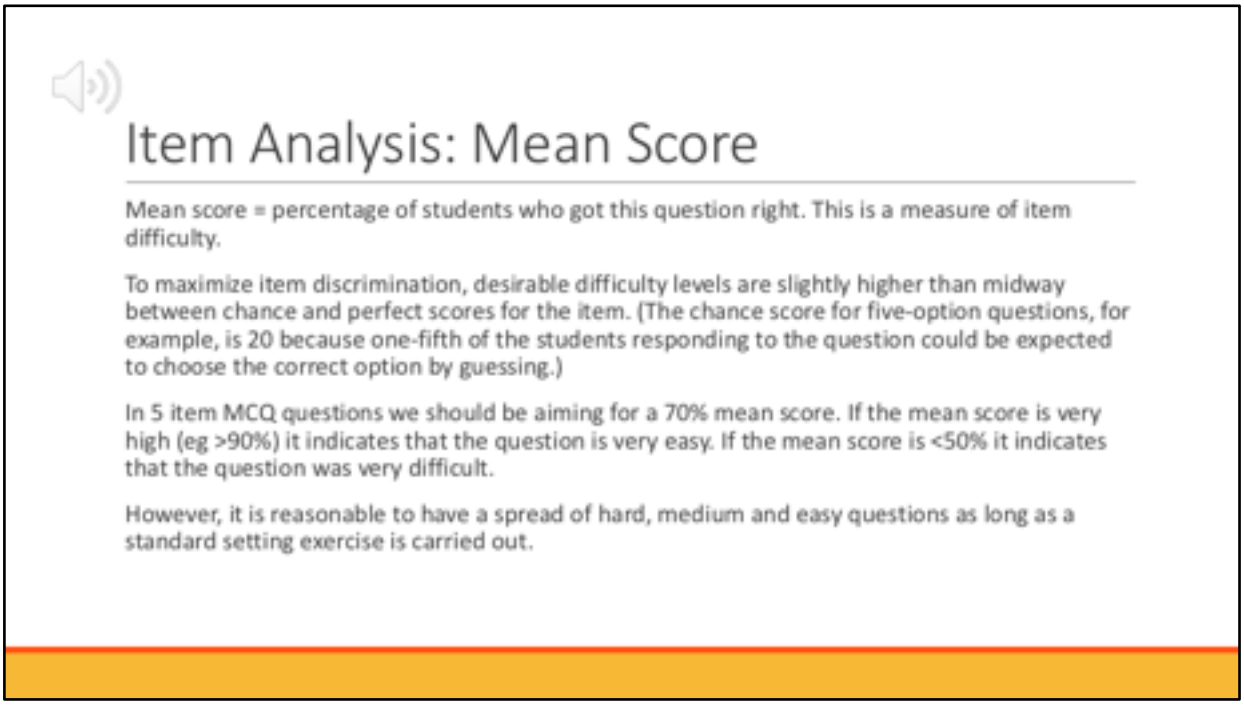
We'll start with the KR-20. This statistic measures inter-item consistency as an indicator for the reliability of the exam as a whole.

It should appear once on the item analysis usually at the top or beginning of the analysis.

This is an example from a 60 item MCQ paper used with 2nd year graduate entry students.

A higher value for the exam indicates a strong relationship between items on the test. This shows the reliability that the same students taking the test again, would achieve the same scores.

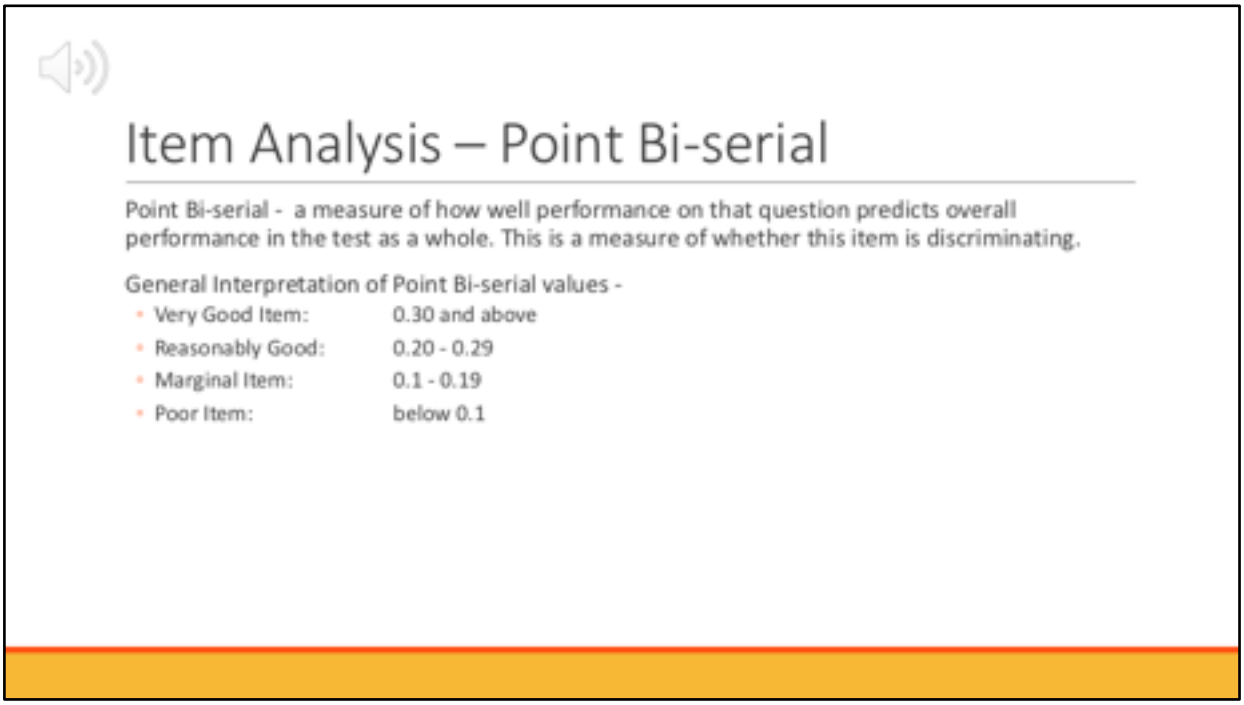
A value of at least 0.70 is desirable.



Next we move onto the mean score. This is the percentage of students who got this question right. This is a measure of item difficulty.

To maximize item discrimination, desirable difficulty levels are slightly higher than midway between chance and perfect scores for the item. (The chance score for five-option questions, for example, is 20 because one-fifth of the students responding to the question could be expected to choose the correct option by guessing.)

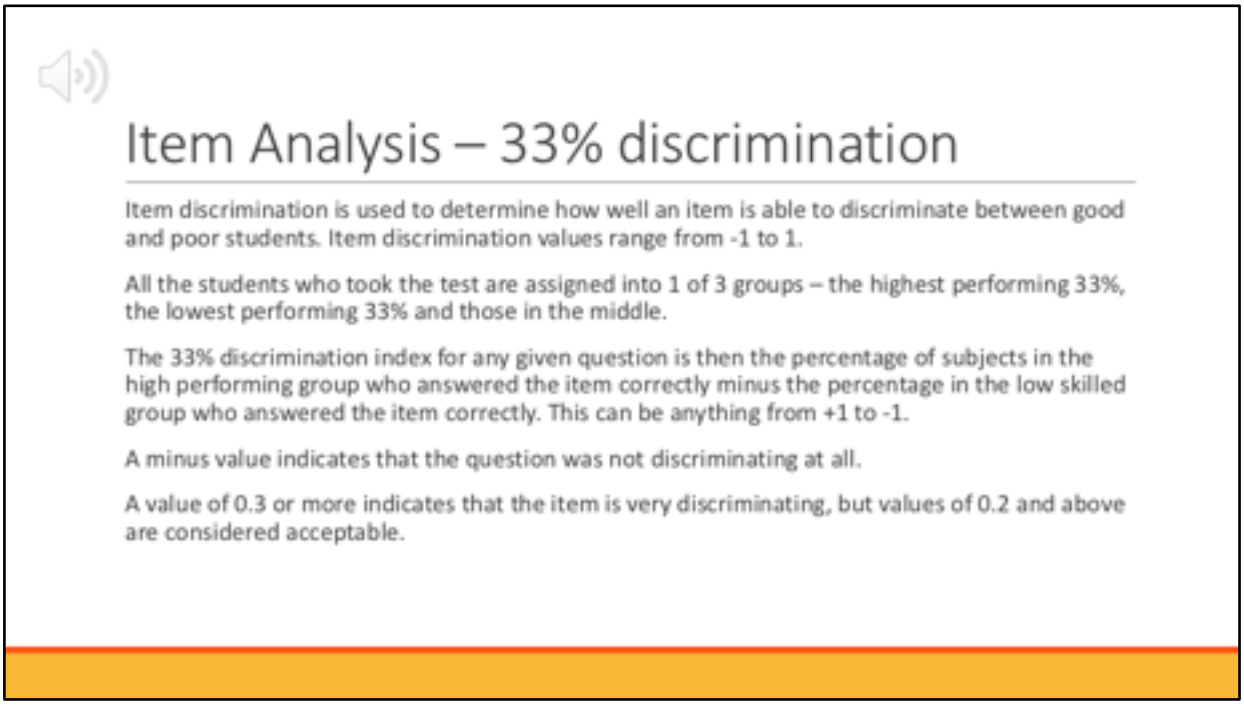
In 5 item MCQ questions we should be aiming for a 70% mean score. If the mean score is very high (eg >90%) it indicates that the question is very easy. If the mean score is <50% it indicates that the question was very difficult. However, it is reasonable to have a spread of hard, medium and easy questions as long as a standard setting exercise is carried out.



The point bi-serial is a measure of how well performance on that question predicts overall performance in the test as a whole. This is a measure of whether this item is discriminating.

General Interpretation of Point Bi-serial values:

- Very Good Item: 0.30 and above
- Reasonably Good: 0.20 - 0.29
- Marginal Item: 0.09 - 0.19
- Poor Item: below .009



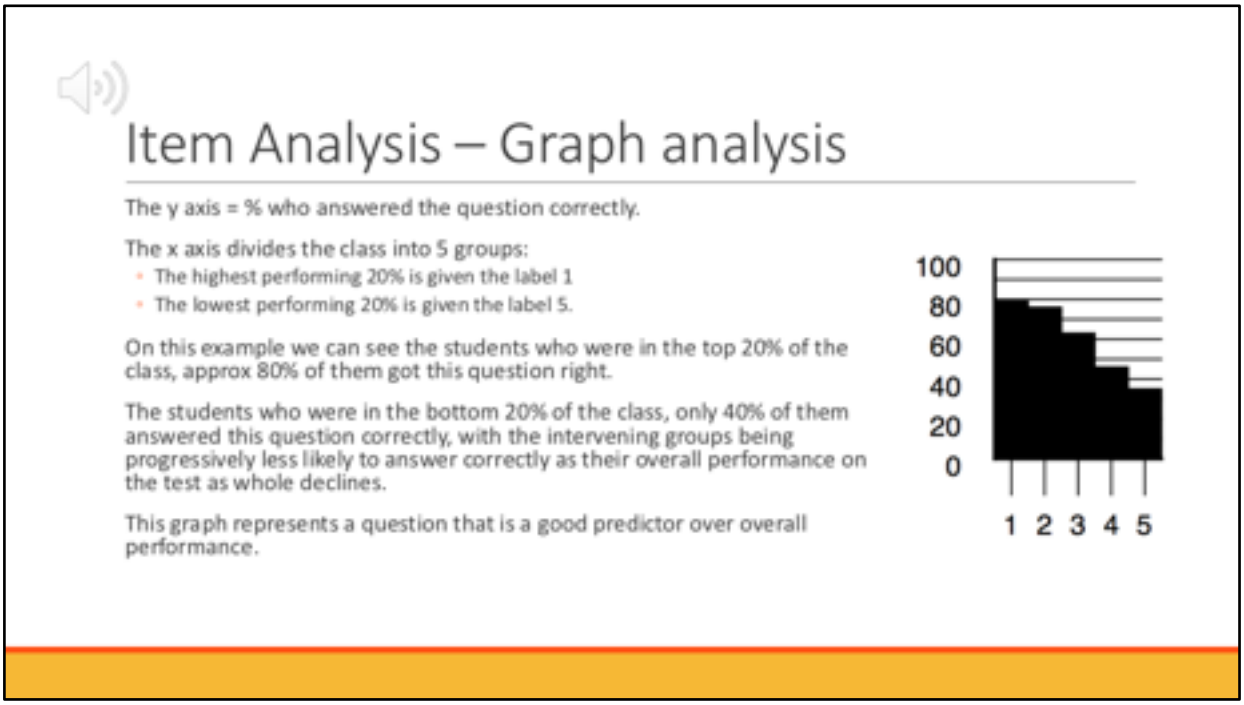
Item discrimination is used to determine how well an item is able to discriminate between good and poor students. Item discrimination values range from -1 to 1.

To calculate the 33% item discrimination index all the students who take the test are assigned into 1 of 3 groups – the highest performing 33%, the lowest performing 33% and those in the middle.

The 33% discrimination index for any given question is then the percentage of subjects in the high performing group who answered the item correctly minus the percentage in the low skilled group who answered the item correctly. This can be anything from +1 to -1.

A minus value indicates that the question was not discriminating at all.

A value of 0.3 or more indicates that the item is very discriminating, but values of 0.2 and above are considered acceptable.



The item analysis for each question usually includes a graph of how well this question was answered by students compared to their overall performance in the exam.

The y axis = % who answered the question correctly
The x axis divides the class into 5 groups based on their performance in the exam overall -

- The highest performing 20% is given the label 1.
- The lowest performing 20% is given the label 5.

On this example we can see that of students who were in the top 20% of the class, approx 80% of them got this question right.

Regarding students who were in the bottom 20% of the class, only 40% of them answered this question correctly, with the intervening groups being progressively less likely to answer correctly as their overall performance on the test as whole declines.

This graph represents a question that is a good predictor of overall performance.

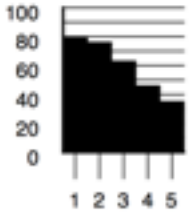
Item Analysis – worked example 1

Question 1

		Thirds	Right	Wrong	R-W	Void
Mean Score:	0.597	% All	59.7	40.3	19.4	0.0
33% Item Discrimination:	0.467	Upper	80.0	20.0	60.0	0.0
Point Biserial:	0.349	Lower	33.3	66.7	-33.3	0.0
Correct Answer:	D	Facility	59.7			

% Answer Frequency

Multi	Blank	D	E	A	C
0.0	0.0	59.7	34.7	4.2	1.4



I'm going to go through the item analysis for 4 different questions in the next few slides.

This item analysis shows a good question.

- It was relatively difficult, with just under 60% of students answering it correctly.
- The 33% item discrimination was 0.467 indicating that the question was a good discriminator between students who took the test.
- The point bi-serial was 0.349, indicating that the question was a good predictor of how well each student would perform on the test overall.
- And the graph shows an acceptable distribution of marks.

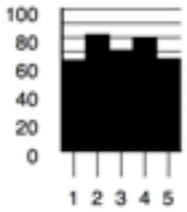
Item Analysis – Worked example 2

Question 18

		Thirds	Right	Wrong	R-W	Void
Mean Score:	0.722	% All	72.2	27.8	44.4	0.0
33% Item Discrimination:	0.008	Upper	71.7	28.3	43.3	0.0
Point Biserial:	-0.011	Lower	70.8	29.2	41.7	0.0
Correct Answer:	B	Facility	72.2			

% Answer Frequency

Multi	Blank	B	A	D	E
0.0	0.0	72.2	13.9	11.1	2.8



This item analysis shows a poor question.

- 72% of students answered it correctly.
- The 33% item discrimination was 0.008 indicating that the question was a poor discriminator between students who took the test.
- The point bi-serial was negative at -0.011, indicating that the question was an inverse predictor of how well each student would perform on the test overall.
- And the graph shows a fairly flat distribution of marks.

So this is a poor question overall and I would not put it in an item bank to re-use in future years. All is not lost though, perhaps there was insufficient or confusing information in the stem so you could look at re-working the question and trying it again in an edited version in subsequent exams.

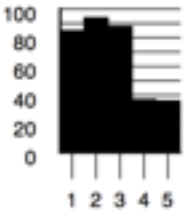
Item Analysis – Worked example 3

Question 33

		Thirds	Right	Wrong	R-W	Void
Mean Score:	0.681	% All	68.1	31.9	36.1	0.0
33% Item Discrimination:	0.542	Upper	87.5	12.5	75.0	0.0
Point Biserial:	0.490	Lower	33.3	66.7	-33.3	0.0
Correct Answer:	C	Facility	68.1			

% Answer Frequency

Multi	Blank	C	E	D	A
0.0	0.0	68.1	27.8	2.8	1.4



This item analysis shows a good question.

- It was neither too easy nor too difficult, with 68% of students answering it correctly.
- The 33% item discrimination was 0.542 indicating that the question was a very good discriminator between students who took the test.
- The point bi-serial was 0.490, indicating that the question was a very good predictor of how well each student would perform on the test overall.
- Just to repeat here, the values that we are looking for in the point bi-serial and 33% item discrimination tests are 0.3 or above for a good predictor or good discriminator, and 0.2 or above for an acceptable predictor or acceptable discriminator.
- And looking at the shape of the graph here, we see an acceptable distribution of marks with students in the top 3 bands quite likely to get it right, with those in the bottom 2 bands unlikely to get it right.

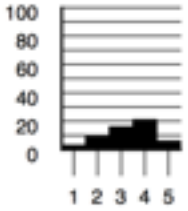
Item Analysis – Worked example 4

Question 7

		Thirds	Right	Wrong	R-W	Void
Mean Score:	0.110	% All	11.0	89.0	-78.0	0.0
33% Item Discrimination:	-0.074	Upper	3.7	96.3	-92.6	0.0
Point Biserial:	-0.085	Lower	11.1	88.9	-77.8	0.0
Correct Answer:	B	Facility	11.0			

% Answer Frequency

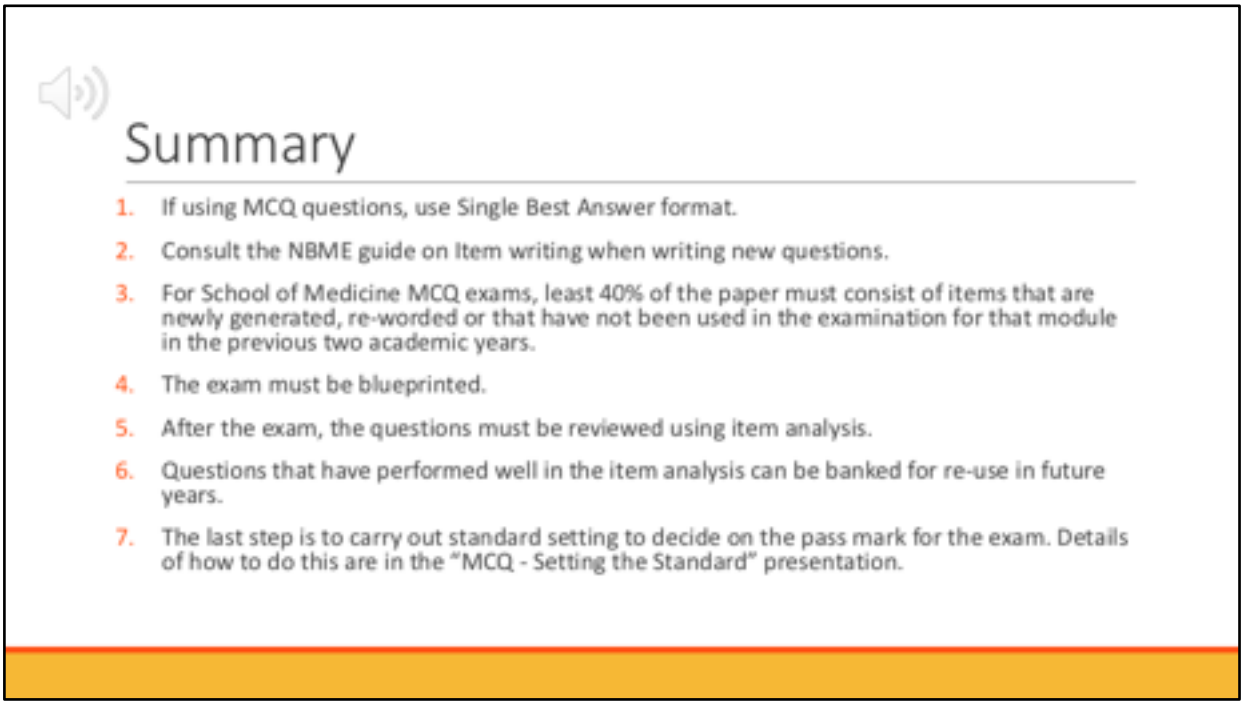
Multi	Blank	A	B	C	D	E
0.0	0.0	53.7	11.0	4.9	29.3	1.2



Finally this item analysis shows a very poor question.

- Only 11% of students answered it correctly.
- The 33% item discrimination was -0.074 indicating that the question was a poor discriminator between students who took the test.
- The point bi-serial was also negative at -0.085, indicating that the question was an inverse predictor of how well each student would perform on the test overall.
- And the graph shows an inverse distribution of marks.

So this is a very poor question overall. I would query whether the in fact the right answer was not, as indicated, answer B, but may instead have been answer A, which 53% selected. I would remove this question from the spreadsheet when calculating the marks of the exam. I would review the question to see if there were any errors. If obvious errors can be spotted and fixed, you might be able to use this question in a later exam and if it performs better the next time, then it could be considered for an item bank.



1. If using MCQ questions, use Single Best Answer format.
2. Consult the NBME guide on item writing when writing new questions.
3. For School of Medicine MCQ exams, least 40% of the paper must consist of items that are newly generated, re-worded or that have not been used in the examination for that module in the previous two academic years.
4. The exam must be blueprinted.
5. After the exam, the questions must be reviewed using item analysis.
6. Questions that have performed well in the item analysis can be banked for re-use in future years.
7. The last step is to carry out standard setting to decide on the pass mark for the exam. Details of how to do this are in the "MCQ- Setting the Standard" presentation.

References

1. Constructing Written Test Questions for Basic and Clinical Sciences. NBME, 4th Edition 2016: https://www.nbme.org/sites/default/files/2020-01/IWW_Gold_Book.pdf

2. Test Blueprinting II – Creating a Test Blueprint: NBME 2019: https://www.nbme.org/sites/default/files/2020-01/Test-Blueprinting-Lesson-2.pdf