MCQ Tests – pros and pitfalls

Multi choice tests take centre stage in the new competition model, in particular for specialist competitions. And they also play a prominent role in both internal and EPSO competitions to assess the candidates' knowledge of EU institutions and policies.

What are therefore the advantages and disadvantages of MCQs and what challenges arise when implementing MCQs?

Touching briefly on the advantages, two considerations are key. Firstly, MCQs, if computer-based, are easy to scale up and can therefore be set-up for several thousand candidates without incurring huge extra costs. All the more so as no markers are required and the scores are generated automatically.

Secondly, the results are objective in the sense that no marker has to read the reply and assess it (which inevitably introduces an element of subjectivity, not to mention bias). Once the candidate has replied to all questions in an MCQ, the result is immediately available and so are the results for all candidates who have completed the test.

Alas, the possible advantages of MCQs have been more than overshadowed by the experience with MCQs in recent competitions, which has not necessarily been positive - to put it mildly. And here we are not only talking about technical issues such as frozen screens or faulty connections. Such pitfalls can occur regardless of the type of online test, while others are more inherent to the nature of the test.

Some problems arise for instance from the fact that EPSO competitions are meanwhile subject to the 24-language regime, meaning that MCQs have to be translated. In particular questions on verbal reasoning are especially sensitive to nuances in translation. What is easy to understand in one language may turn out to be ambiguous or even confusing in another if not translated very carefully.

But there have also been problems with MCQs that had nothing to do with translation issues. In several competitions, a disproportionate number of questions had to be cancelled. While the reasons may differ between questions (we are not privy to the reasoning of the juries involved), there are basically two possible reasons for this: either the answer assumed to be correct is demonstrably wrong, or there is more than one correct answer. Last but not least, somewhere between these two extremes lies the possibility that, even before any translation takes place, both the question and the answer options have not been carefully formulated and are therefore likely to mislead the candidates.

Arguably, these aspects are at least as important as the technical implementation. After all, the result of an MCQ cannot be better than what was previously put into it. And, as we shall argue, this requires much more than providing a meaningful question and a choice of replies, one of which has to be correct.

The focus in what follows will therefore be on MCQ tests as such, their generic content and design.

So, in a nutshell, this is what the analysis is about. Firstly, we will introduce some terminology to prepare the ground for further discussion. We will then focus on criteria for assessing multiple-choice questions (MCQs) as a whole and as individual questions. Against this background, we will examine some recent open competitions to evaluate the design of the included MCQs.

The final section of the paper discusses criteria for good MCQ questions and draws some conclusions regarding current Commission practice.

Terminology

To begin with, let's look at the standard terminology that is used in the context of MCQs. Here is what Wikipedia has to say about the matter:

"Multiple choice items consist of a **stem** and several **alternative answers**. The stem is the opening – a problem to be solved, a question asked, or an incomplete statement to be completed. The options are the possible answers that the examinee can choose from, with the correct answer called the **key** and the incorrect answers called **distractors**."[1] Note that in the case of a problem or an incomplete statement, a question of the type 'What is the solution to the following problem?' or 'How can the statement be completed?' can always be asked.

In contrast to everyday language, an item therefore refers to an – explicit or implicit – question *and* the various possible answers that are proposed to the candidate.

For items used in MCQs on verbal and numerical reasoning, the stem can consist of multiple parts. The stem may contain extended or additional material, such as a text, a case study, a diagram, a table or a detailed description with multiple elements. Anything can be included as long as it is necessary to maximise the validity and authenticity of the item.

This is by and large also true for knowledge tests. Here is an example of the latter:

> How many Member States did the European Union have on January 1 2025? [STEM]
>
> a) 29 [option 1]
> b) 27 [option 2 - KEY]
> c) 26 [option 3]
> d) 28 [option 4]

Or another one:

> Which of the following is NOT an objective of the Green Deal? [STEM]
>
> a) no net emissions of greenhouse gases by 2050 [option 1]
> b) economic growth decoupled from resource use [option 2]
> c) no person and no place left behind [option 3]
> d) raise environmental awareness among EU citizens [option 4 - KEY]

MCQ *tests* consist of several items. When examining such a test, it is therefore useful to distinguish between assessing how the test items perform as a set, and how items perform individually or in relation to other items. The former is called test analysis whereas the latter is known as item analysis. In what follows, we will look at both dimensions of analysis, referring as much as possible to concrete examples.

Let's start with some theory though.

The theory behind MCQs

The idea behind, and purpose of, MCQ tests is to measure a construct (= a not directly observable psychological property), such as knowledge of EU institutions and policies, or knowledge of

[1] https://en.wikipedia.org/wiki/Multiple_choice

a language, or the analytical abilities of a candidate, the hypothesis being that a person who possesses the property to a greater extent is more likely to choose the correct answer and will score higher. The observed score X is therefore considered to be an estimate for the unobservable property T.

Why is it an estimate and not the real thing as it were? The answer is threefold. Firstly, most MCQ tests will only consist of a sample of items and therefore never comprise all available items in the population.[2] Secondly, even the population from which the item is chosen at random cannot be said to comprise everything that there is to know about a subject such as the EU Institutions, even assuming that such a field of knowledge can be clearly defined.[3] Thirdly, and for various reasons, one being pure good or bad luck, even a knowledgeable candidate may choose occasionally a wrong answer, and an ignorant candidate may sometimes choose more correct answers than what could be expected. Hence, the observed score X (i.e. the number or percentage of correct answers), the estimate, equals the true score T plus some error E, X=T+E.

Importantly, a candidate who throws a four-sided dice (if there was one) to determine the option s/he will tick in a test comprising items of the type exemplified above, will on average get 25% of the questions right. Hence a score in percentage terms of 25% would suggest that the candidate has no knowledge whatsoever. That is why the threshold for passing a MCQ test comprising items of this kind must be higher than 25%.

Reliability

Beyond these rather straightforward considerations, psychometricians have identified other questions and issues. While it would go far beyond the scope of this analysis to scratch even on the surface of many of those, there is one concept, namely reliability, that is of obvious relevance and that can be easily illustrated.[4]

Accordingly, a MCQ test can be said to be *reliable* in terms of what it is supposed to measure if the candidates with similar scores got mostly the same answers right. In other words, correct answers are correlated across candidates. By way of illustration, consider a situation in which two groups of candidates got each 50 out of 100 questions right, but candidate group 1 answered mostly questions 1 to 50 correctly, while candidate group 2 answered mostly questions 51 to 100 correctly. Can it be said that this test measures the same property for both candidate groups despite equal scores and that, as a corollary, a difference of, say 10 points between two test takers always means the same? Probably not. Both groups were equally good, if judged by the obtained score, but on different matters. If, on the other hand, both candidate groups answered questions 1 to 50 correctly or at least a much bigger proportion, then it can be argued that they possess to the same extent the same (not directly observable) property.

This point is important because, arguably, the purpose of the MCQ on EU knowledge or in a specific field for instance is to identify suitable candidates (= candidates possessing the required knowledge), not just reduce the number of candidates. The latter could be achieved with a lottery or a quiz; the former requires a reliable test.

---

[2] The reason is that tests may not take place on the same day or that candidates have to be retested. In both situations, equal treatment demands that candidates are given a set of questions not yet known to them.

[3] In statistics, population refers to the entire group of individuals, objects, or events that a study or analysis is interested in. It's the complete set from which a sample is drawn for examination.

[4] See for instance Paniagua and Swygert (2016), Stemler and Naples (2021) or Matlock-Hetzel (1997).

Alas, in particular MCQs in specialist competitions are unlikely to lead to much better results than a lottery for reasons very similar to those just outlined.

Here is the argument in detail.

MCQs in specialist competitions

In specialist competitions, the MCQ on EU knowledge is replaced by a field-related MCQ (FRMCQ) where candidates are given a set of items from the field of the competition. To illustrate the point, consider the recently finished competition for administrators (AD 7) in the field of transport (EPSO/AD/410/23) where candidates were given 30 questions. This seems to be a substantial number. However, according to the Notice of Competition (NoC), candidates are expected to have acquired professional experience in at least one out of 11 areas. These areas reach from "Connectivity, automation, digitalisation, artificial intelligence in transport and intelligent transport systems" to "Labour law and social protection in transport".

Now, although it is not specified in the NoC, it is only reasonable to assume that the MCQ should cover these 11 areas to more or less the same extent. For otherwise, it would give candidates with knowledge in some areas an undue advantage compared to candidates with knowledge in others.

If this is the case, on average fewer than three questions can be asked per area. As a knowledge test for one of the 11 areas, such a small number is obviously neither here nor there. And even if a candidate has knowledge in several areas and has some ideas in others, the small number of questions per area essentially means that the final result is determined to a significant extent by chance, as candidates must still guess the correct answers to many questions about which they know virtually nothing. A corollary of this observation is that, referring to the above argument, it cannot be said that the resulting reserve list comprises experts *in the field*. Rather, the final reserve list comprises a somewhat random selection of candidates who know a little about a handful of topics and are otherwise simply good at guessing.

Unfortunately, there is no way of knowing the scale of this problem because nobody is investigating it. And even if an issue was discovered, it would only be when the mass had already been sung, as the next specialist competition is bound to be different. One thing is clear, however: selecting future staff for the institutions on such a weak basis makes a mockery of any responsible and forward-looking personnel policy.

A similar problem may also occur at the level of individual results. Accordingly, the score of a candidate is considered to be a reliable measure of the underlying property if, for instance, candidates with a high score also answered a higher percentage of the more difficult questions correctly (and got right not only more of all questions or even only of the easier ones).

For this to be the case, items must have sufficient discriminatory power (see below). That is to say, items must be construed such that they help to distinguish conclusively between better and worse candidates.

These observations are important because they highlight once again why an MCQ differs from a quiz, which merely determines a winner who is awarded a prize. The latter only involves providing the correct answer to select a winner. In an MCQ, however, the correct answer is a means of judging whether a candidate possesses the desired quality.

And now to the Elephants in the room

As the above considerations should have made amply clear, designing a reliable and meaningful MCQ is far from straightforward. And this concerns both the development of individual questions *and* the selection of questions for a comprehensive test. But things do not finish with the development of the questions. This is only the first step, even though it poses a number of challenges in its own right – we will come to that in a moment. MCQs need to be quality-controlled both before they are rolled-out and after they have been executed. If issues are found with any of the questions, then the problematic ones need to be corrected or eliminated.

Let's start with the development of questions.

It goes without saying that a question should be formulated in such a way as to allow for an unambiguous answer. The above example concerning the number of Member States is an obvious case, albeit rather simplistic. The correct reply can be easily verified, of course. One reliable source is sufficient as proof.

The second example is trickier. Firstly, because it is about a negative ("Which of the following is NOT an objective of the Green Deal?"). In other words, something is not the case or does not exist. The problem with this type of question is that you have to check all the relevant documents to make sure that the alleged objective isn't mentioned anywhere, even by mistake. And that can be a considerable undertaking.

Secondly, all the other options – the distractors – must be valid in their own right. They must be – in the above example – objectives of the Green Deal and this must be demonstrable and demonstrated. Therefore, the development of an item where the question is negative is even more demanding than where the question is positive, suggesting that such questions should be avoided.

Thirdly, the correct answer is a phrase open to interpretation, as is the question. However, interpretation may sometimes not be straightforward or univocal, which creates ambiguity and doubt, thereby reducing the meaningfulness of the test.

These considerations point to a more general difficulty in creating items for MCQs. Whether a question (stem) makes sense, whether it is difficult or easy to answer does not only depend on the question itself and the correct answer – the key – but also on the distractors – the alternatives offered to the candidate.

Consider again the first example above with somewhat different numbers in the distractors:

> How many Member States did the European Union have on January 1 2025? [STEM]
>
> a) 193 [option 1]
> b) 27 [option 2 - KEY]
> c) 38 [option 3]
> d) 4 [option 4]

Even someone with limited knowledge of European affairs can easily rule out options 1 and 4. Option 1 (the number of UN member states) is far too high, and option 4 (the members of the EFTA) is far too low, as even a cursory glance at a map of Europe reveals. This would leave the candidate with options 2 and 3. Obviously, a candidate who is still not sure about the right answer has now a much better chance of ticking the right box, namely around 50% as compared to 25%.

The observation helps to emphasise an important point. Whether a multiple-choice question is easy or difficult, does not only depend on the question and the correct answer but, crucially so, also on the distractors. If these are so implausible as to rule them out immediately, then a candidate can identify the correct reply by eliminating the other options. Setting a threshold at 50% is therefor rather meaningless if candidates can easily rule out some answers. Concomitantly, if the distractors are formulated in such a way that they are too similar, that could create an unwanted level of difficulty.

Consider the following example:

> How many standing committees does the European Parliament have during its current term? [STEM]
>
> a) 26 [option 1]
> b) 24 [option 2 - KEY][5]
> c) 25 [option 3]
> d) 23 [option 4]

Someone with a rough idea of the number of committees, but who is unsure whether the two special committees count towards that number, could easily confuse options 1 and 2. They don't, but this detail may easily be overlooked. Options 3 and 4 are so close to the correct answer as to prompt some candidates to select them.

Whether this example is a well-designed multiple-choice question is therefore at least not obvious. And the issue of whether it is an easy question or a difficult one is also not so straightforward to answer. An expert in the field of European affairs may argue that it is a difficult but meaningful question given the perhaps not widely known number of committees and the distinction between standing and special committees. But this argument ignores not only the role of the distractors, it also rests on a perception of the question that may not be shared by candidates.

At this point at the latest, we therefore need to take a closer look at the statistical analysis of MCQs, i.e. the question of how candidates themselves answered the questions and how often they selected certain options.

For instance, the choices of candidates could be distributed as follows:

a)	26 [option 1]	35%
b)	24 [option 2 - KEY]	30%
c)	25 [option 3]	20%
d)	23 [option 4]	15%

Accordingly, more candidates have chosen the (incorrect) option 1 rather than the correct option 2. If this is the case, then the item should be redesigned.

Similarly, if the choices of candidates are distributed as follows

a)	26 [option 1]	35%
b)	24 [option 2 - KEY]	40%
c)	25 [option 3]	20%

[5] https://www.europarl.europa.eu/committees/en/about/list-of-committees

d) 23 [option 4] 5%

then perhaps option d) is not a good distractor because most candidates can easily recognise it as evidently wrong and will choose one of other options. As a consequence, the item is easier than it seems at first sight.

Now, look at the following distribution:

a) 26 [option 1] 5%
b) 24 [option 2 - KEY] 85%
c) 25 [option 3] 5%
d) 23 [option 4] 5%

If the distribution of choices is like this[6], then it suggests that the question is perhaps too easy. And if all questions of a multiple-choice test are like this one, then too many candidates will pass the test, and the test does not fulfil its purpose. Likewise, if the choices are equally distributed among options (25% each), then this suggests that the question is perhaps too difficult as even good or knowledgeable candidates do not achieve a better outcome than those who just guess (who would choose the right answer in 1 out of 4 cases).

This brings us back to a topic that was briefly mentioned above. How well can a question distinguish between good and bad candidates?

Look again at the following distribution. But now we will assume for the sake of the argument that these figures are for the top-scoring 50% of candidates.

a) 26 [option 1] 5%
b) 24 [option 2 - KEY] 85%
c) 25 [option 3] 5%
d) 23 [option 4] 5%

For the lowest scoring 50% of candidates, the figures are assumed to be as follows:

a) 26 [option 1] 25%
b) 24 [option 2 - KEY] 30%
c) 25 [option 3] 25%
d) 23 [option 4] 20%

Accordingly, a much higher percentage of the good candidates has answered that question correctly compared to the percentage of not-so-good candidates who choose the right answer. This means though that the question can help to identify good candidates.

If the numbers were like this for the top-scoring candidates, then the question does not discriminate much:

a) 26 [option 1] 25%
b) 24 [option 2 - KEY] 35%
c) 25 [option 3] 20%
d) 23 [option 4] 20%

[6] The percentage of test takers who answered the item correctly is usually referred to as item difficulty. The larger the percentage getting an item right, the easier the item (Matlock-Hetzel 1997).

In other words, the reply to the question contributes almost to the same extent to the total score of a candidate, no matter whether this is a strong or a weak candidate. Therefore, adding a question of this kind to a multiple-choice test will not significantly impact the final outcome.

A statistic describing the discriminatory power of a test item is the so-called discrimination index. The discrimination index, *D*, is the number of people in the upper group who answered the item correctly minus the number of people in the lower group who answered the item correctly, divided by the number of people in the largest of the two groups.

Taking the above example and converting for the sake of simplicity the percentages into numbers of candidates, this is what we get for the first situation:

$$D = \frac{85 - 30}{100} = 0.55$$

With different numbers for the top-scoring candidates, this is what we get:

$$D = \frac{35 - 30}{100} = 0.05$$

Evidently, *D* is now much smaller, suggesting that the item has less discriminatory power.

How do we know?

The above examples and figures may be all good and well, but at the end of the day, proof of the pudding is in the eating, both before and after. This means that a MCQ must be thoroughly tested before it is given to candidates. It is not enough to ask experts in the field whether a question is easy or difficult, this must be tried with mock candidates. Based on the data from such a trial, the correct replies and the distractors must then be analysed with a view to determine whether they lead to the desired level of difficulty and whether none of the distractors is either misleading (too many candidates have selected it) or redundant (quasi nobody selected it).

To the best of our knowledge, the Commission does not carry out such analyses. Allegedly, the argument is that questions may leak, which would benefit candidates who manage to obtain the information. Ultimately, however, this is a weak excuse. Given that successful completion of a selection process leads in the vast majority of cases to recruitment – with resulting financial commitments running into millions – the Commission should spare no effort in thoroughly vetting MCQs in advance.

The same applies to the ex-post review of MCQs, especially since, by definition, there is no risk of secrets being revealed. It is therefore all the more important to determine whether field-specific MCQs meet the expectations set for them or whether they are little more than a quiz. The counter argument in this is case is that field specific MCQs are unlikely to be repeated any time soon. While this argument has some merit, together with the absence of any ex-ante analysis, it implies though that there are no quality checks whatsoever. Once again, given what is at stake, this lack of any quality control is staggering.

All this applies even more so to MCQs on EU policies and institutions, which are held with fair regularity. The large number of cancelled questions highlights the need for thorough preliminary checks, whereas the fact that such MCQs are conducted regularly is a compelling argument for an ex-post evaluation of the results. This is all the more true given that it is not even known whether the pool of questions from which each test is compiled contains only questions of

broadly the same level of difficulty. If this is not the case, then candidates are not treated equally across all selection procedures. Some may pass a competition with much less EU knowledge than others.

Of course, if the Commission's only objective was to reduce the number of candidates and end up with a manageable number on the reserve list, however they achieved it, that would be perfectly OK. However, the Commission cannot talk about making the European Public Service fit for the future and attracting talent from across Europe while implementing methods that fall short of any quality criteria.

References

Matlock-Hetzel, Susan. 1997. “Basic Concepts in Item and Test Analysis.” *Educational Research Association* (January 1997): 1–7.

Paniagua, Miguel A., and Kimberly A. Swygert. 2016. *Constructing Written Test Questions For the Basic and Clinical Sciences*.

Stemler, Steven E., and Adam Naples. 2021. “Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line.” *Practical Assessment, Research, and Evaluation* 26(1): 1–16. doi:10.7275/V2GD-4441.