ESRC National Centre for Research Methods

Qualitative Evaluation of LEMMA questions

Ayesha Ahmed

Report prepared for the LEMMA (Learning Environment for Multilevel Modelling and Applications) node, University of Bristol

October 2011

Aknowledgements

The LEMMA project is funded under the ESRC National Centre for Research Methods (NCRM). The LEMMA online course was established in April 2008 under Phase 1 of NCRM (grant number RES-576-25-5006) and further developed under Phase 2 (RES-576-25-0003). A third three-year project began 1 October 2011 (RES-576-25-0032).

The LEMMA online course

The LEMMA (Learning Environment in Multilevel Modelling and Applications) online training system has been developed by the Centre for Multilevel Modelling, with the first release available in April 2008. The course was developed under the LEMMA research project, a node of the ESRC-funded National Centre for Research Methods. The LEMMA course is housed in Moodle. After registration, the materials can be accessed free of charge. An overview of the materials and registration details can be found at http://www.bris.ac.uk/cmm/learning/course.html.

The primary target audience for the course is researchers from the social sciences and public health, at all career stages and from both academic and non-academic sectors. With this audience in mind, the materials use examples from a range of social science disciplines and public health.

An important aim of the course was to cater for all levels of learner, from those with minimal experience of traditional elementary statistical methods (possibly needing a refresher) to more advanced users, including those who wish to train others in multilevel modelling. The course was therefore designed so that learners can enter at different points according to their prior experience of statistical modelling.

Most modules consist of a 'concepts' and 'practice' component. The 'concepts' gives a detailed description of a statistical model, including the types of data and research questions it can be used to investigate and its interpretation, but without reference to any statistical software package. The 'practice' component then provides instructions for the analysis of a particular dataset using a range of software packages, currently MLwiN, Stata and R.

Although the course is unsupported, quizzes (with feedback) are provided to allow learners to assess their understanding of the material. As part of an evaluation of the course an independent consultant, Dr Ayesha Ahmed, reviewed the quizzes. As a result of the report, many of the quiz questions have been revised to take account of the recommendations.

About the report author

Ayesha Ahmed has a PhD in cognitive development, and has worked in the psychology of educational assessment since 1998. She has designed and delivered training for exam question writers and carried out research on improving the quality of exam questions. Her particular interest is in ways of controlling demands and reducing construct-irrelevant variance in examinations.

This report considers the quality of the questions that form part of the Learning Environment for Multilevel Methodology and Applications (LEMMA) online course. The questions and the feedback that comes with them are an important part of the learning resource and are overall of high quality. One of the things that makes them so good is the carefully written feedback that is given after each incorrect attempt at answering a question. In this report I will mainly consider the ways in which the questions are presented and the language used to communicate the tasks, and I will discuss how these factors may be affecting the learners' scores on the questions. The users of the LEMMA materials range from students to senior academics, and I will refer to them all as 'learners' throughout.

Most of the questions are very well presented in the learning environment, but there are a few in which the learners may have been misled, and it is these that I will focus on. I will illustrate with examples the main issues that arose. These were: questions that were particularly difficult in which the clarity of the wording may have been a factor; questions with complex item formats leading to deductive reasoning; questions for which the options sometimes gave clues to the answer; and marking issues. The purposes of this analysis are a) to improve the LEMMA questions themselves and b) to inform a wider audience about the kinds of problems that can arise in designing statistics assessment questions.

I will start with describing the method used for analysing test questions, and the theoretical basis for this approach. I will then go through some examples from LEMMA of questions that proved to be particularly difficult and explain what might have caused some of this difficulty. I will go on to outline an approach to question writing that is designed to minimise *construct irrelevant variance,* that is differences in scores that are caused by unintended task demands. The report will conclude with a summary of the main issues to consider when writing test questions, and in particular multiple choice questions.

Introduction

In any assessment, we must be concerned first with its purpose and what it is we want to assess – the construct. Only with these in mind can we turn to the writing of assessment tasks. When writing questions and mark schemes, the construct – what is being assessed – should be kept in mind at all times. This can be formally written down as an 'Importance Statement' for the subject being assessed (e.g. QCDA, 2011, p14). Doing so makes it available to all involved: question writers, markers, learners, teachers, etc. In the case of LEMMA the Course Contents states:

> "This site is a course that brings you from introductory-level statistics to a basic understanding of Multilevel Modelling."

It is clear from the above statement what the purpose of the course is and what the test questions are for in general. For each 'lesson' learners work through the concepts, and in most cases a practical exercise (in a range of statistical software packages – MLwiN, Stata and R), before answering the questions. The construct

that is being assessed within each section is therefore elaborated by the curriculum material.

Another critical issue is that questions and mark schemes are written alongside one another and should not be considered separately. The writing of the mark scheme informs the writing of the question and vice versa. In the case of multiple choice questions, deciding on the correct answer (key) and the other options (distractors) is a similar process to that of writing a mark scheme for a question requiring a short written answer or a calculation. I will address this issue in more detail in the final section.

The Question Answering Process

The process of question writing can be informed by understanding what happens in learners' minds when they meet a test question. This is particularly useful for thinking about what might go wrong when a learner meets the question you are writing. Testing is a process of communication and the language used to convey the task to the learners is critical (Rea-Dickins, Afitska, Yu, Erduran, Ingram & Olivero, 2009). The job of a test question is to convey the task to the learners – that is all – and if the language is not clear there is a danger of the question getting in the way of what we want to achieve. If this happens, the learners' answers may not be telling us what we need to know, as the learners are not actually carrying out the task as intended.

A psychological model of the question answering process can show us where these dangers may lie, and below is a simplified version of the model (see Pollitt & Ahmed, 1999; Pollitt, Ahmed, Baird, Tognolini and Davidson, 2008).

Phase 0 : Learning-	what we are measuring
Phase 1: Reading -	the construction of a mental representation of the task
Phase 2: Thinking -	activation of concepts related to the question words, secondary activation of other concepts, making connections, getting an idea of an answer
Phase 3: Writing -	producing a visible response to complete the task

The first phase is the learning and it is this that we are trying to measure. We call it Phase 0 because in summative assessment it happens before the test while the rest of the phases happen during the test. In the case of LEMMA the test questions are very much embedded in the learning process with questions at the end of each part of each module. Also, feedback is given, so the assessment is formative and learning occurs during the testing process.

Phase 1 is Reading the Question and this is the phase at which misunderstandings often occur as students build a mental model of the task they must complete. Such misunderstandings of the question can cause students to be

completing a task that is not the one intended by the question writer, and is therefore a threat to valid assessment. If the students' minds are not "*doing the things we want them to show us they can do*" (Pollitt & Ahmed,1999) then we are not measuring them on the intended construct: we are not measuring their learning on that topic. Also, some students may misunderstand while others comprehend, or some may misunderstand in different ways, and this can result in different students carrying out different tasks. In this situation we cannot measure any of them fairly.

In normal reading we all make occasional errors, but in a test situation these are more likely to occur as students are working at a higher level of arousal or stress. In these circumstances, expectations are very important and students will often see what they anticipate (Crisp, Sweiry, Ahmed & Pollitt, 2008). It is crucial then that question writers use language that is as clear as possible and as natural as possible. Even though the LEMMA quizzes are not high stakes formal tests, it is still important that the demands in the questions reflect the learning as closely as possible and are not confounded by language demands. We should not be making any reading demands on the learners unless reading is part of the construct that we wish to assess. Reading demands in the LEMMA test questions would be considered *construct-irrelevant demands,* that is demands that are not part of the area of learning being assessed. I use the term 'demands' to mean challenges that are put into a question. The term 'difficulty' I reserve for what the data tell us about what happened when the question was answered. A question can be made more or less demanding, but only when we have the data do we know how difficult it actually was (Pollitt, Ahmed & Crisp, 2007).

Phase 2 is Thinking and this is an unconscious and automatic process in which *activation* spreads in the brain as the student's mind searches for relevant concepts, matches these to those in the question and generates an idea of an answer. In the final phase, Writing, this idea must be translated into (usually) a written answer. In the LEMMA questions this is sometimes a calculation but is most often multiple choice. In reality the phases of Reading, and Activation are not discrete, but are happening in conjunction with one another, with the Writing phase also often co-occurring with the earlier phases.

For multiple choice questions the Reading phase includes reading the options. In some multiple choice questions the answer will be generated and *then* matched to an option. For other types the answer is generated by choosing amongst the options. For many the process will be a mixture of those two.

Question writers must consider carefully the wording of the question to avoid reading errors, and they must also consider what concepts will be activated in the learners' minds as they read the questions, and whether these are the intended concepts.

Qualitative analysis of LEMMA questions

For any question writing there are certain pitfalls that we know cause construct irrelevant demands and hence threaten validity (Pollitt & Ahmed, 2000). Many, though by no means all, are to do with the language of the question. My approach to the qualitative analysis of test questions is based on my experience carrying out experimental studies and interview studies on aspects of test questions that affect their demands and difficulty (see for example Ahmed & Pollitt, 2007; Ahmed & Pollitt, 2010; Pollitt et al, 2008; Crisp et al, 2008) and, in the case of LEMMA, is an explorative approach informed and led by the data from the test. I use the model of the question answering process to work through the learners' cognitive processes, looking for anything in the question that might affect its difficulty in a way that is not relevant to the construct being assessed. This is guided by the data showing percentage correct on each question. I focused on questions with the lowest percentage correct with the aim of trying to understand why these were so difficult and whether it was for construct relevant or irrelevant reasons. I also looked closely at those questions for which the percentage correct was a lot lower than expected by the raters. I would ordinarily also look closely at questions with the highest percentage correct but in the case of LEMMA many of these were multiple choice questions with just two options, which was contributing to their high facility values. I have also included in the examples some questions for which the score was not particularly high or low but there were some issues in the question presentation which I felt might be affecting the scores in a construct irrelevant way.

In this section I will go through the issues that arose in a qualitative analysis of the LEMMA questions, and illustrate these with examples. In some cases I have inserted the actual question to help illustrate the point, but in other cases the question was too big and contained large graphics, so it will be necessary to refer back to LEMMA to see the example being discussed. Before I go any further I would like to point out again that the LEMMA questions were very well written in general and there were few instances in which I felt that learners were scoring low because of the question rather than due to a lack of understanding of the concept. One of the things that makes the LEMMA questions so good is the feedback that learners are provided with immediately after answering. Also, learners are encouraged to give their own feedback on the course materials, and this has been used by the writers of LEMMA to revise the wording of questions and of answer feedback where deemed appropriate.

Examples

Up to the end of Module 5, LEMMA contains a total of 78 questions, some of which are further divided into sub-questions making a total of 152 items. Twenty-five of these are used as examples in this report, in order to illustrate the issues arising. The examples are divided into four sections according to the

issues identified. These are: Difficult questions, Complex item format, Clues in options, and Marking.

Difficult Questions

The questions in this section were all particularly difficult. I have indicated possible reasons for this difficulty for each example.

PR04ii proved to be a very difficult question (16% correct) despite being a two option multiple-choice question.

> 4 (ii) The Pearson correlation coefficient (*r*) is calculated for a pair of variables, X and Y.
> **Decide whether the following statement about *r* is true or false**
>
> A correlation of zero implies that there is no relationship between X and Y
>
> True
>
> False

The low percentage of correct answers could reflect a genuine difficulty with the question as the Pearson correlation coefficient is widely misunderstood as a measure of *any* kind of association rather than a *linear* association. However, it is also possible that the wording of the question has increased its difficulty.

The correct answer contains a double negative: it is 'false' that there is 'no relationship'. The use of negatives should be avoided wherever possible as negatives are hard to read and easily missed. The word 'not' is particularly dangerous as a question usually makes sense without it, but not the sense intended. If negatives have to be used they should always be highlighted. In multiple choice it is sometimes hard to avoid using double negatives: one in the stem and one in an option as in the example above. This can be very confusing and should be avoided if possible.

Is the use of the word 'implies' fair in the above example? It could be argued that although a zero correlation means there could be a relationship (just not a linear one) it is likely that there is no relationship at all so perhaps it does 'imply' this. The result could be a bit of a '*trick question'* in which those who know the answer may have chosen the wrong option due to the wording. We certainly don't want to 'trick' anyone into giving a wrong answer. However, the fact that this is part of the pre-requisite test and not in an actual module test, and the fact that feedback is given, alleviate the effects of this problem.

The feedback means that those who got it wrong, for whatever reason, have now understood the point of the question:

> The statement is **false**. A correlation of zero implies that there is no LINEAR relationship between X and Y. It is possible that a strong curvilinear relationship can have a correlation coefficient of zero. You should ALWAYS

examine a scatterplot of X vs. Y before calculating the correlation coefficient to check that their relationship can be assumed linear.

Q1.4.2 Learners are performing poorly on this question considering it is a two option multiple choice (43% correct).

> **Is the following statement true or false?**
> If every member of a population is included in our sample then we do not need to employ statistical generalisation techniques.
>
> false
>
> true

This appears to be a case of a demanding concept, which is clearly explained in the feedback:

> The statement is **false**. Even if every member of a population is included in our sample the idea of generalisation is still relevant. This is because a population can be thought of as a set of realisations from some underlying process, or *superpopulation*, that could extend through time and possibly space. For example, if our population is all teachers in the UK, then the superpopulation would be all potential UK teachers. This underlying process has driven the observations, but the statistics we compute from the observed data refer to a particular point in time and are subject to random fluctuations. We are interested in the underlying process that has generated the data we observe, and use the 'sample' data- i.e. our population data- to make inferences about this process.

Again though, note the use of a double negative in the question, i.e. it is "false" that we "do not need to". The demands of interpreting the question may be distracting learners from thinking carefully about the important concepts in the question.

PR07 This is an example of a question with a high level of reading demand. It is the options in particular that are challenging to read, and may have resulted in the question being difficult for the wrong reasons (21% correct). However, confidence intervals are a very commonly misunderstood concept, and this will have contributed to the low percentage of correct answers.

> 7) What is meant by 95% confidence with regard to a *95% confidence interval for the mean*?
> **Choose *one* of the following descriptions**
>
> There is a 95% chance that the interval calculated using data from a given sample contains the population mean.

There is a 95% chance of selecting a sample such that the 95% confidence interval calculated from that sample contains the population mean.

There is a 95% chance that the interval calculated using data from a given sample contains the sample mean.

There is a 95% chance of selecting a sample such that the 95% confidence interval calculated from that sample contains the sample mean.

When there is dense information in the options, readers tend to move quickly to reading the options without giving enough attention to the stem, and this can cause reading errors. Their focus of attention is on the differences between the options rather than where it should be: on the relationship between the options and the stem.

The order of the options in this question contributes to the reading demands. Options 1 and 3 go together naturally, and so do 2 and 4. Options should always be given in a natural, logical order unless there is a specific, *construct relevant* reason not to do so.

Most of the questions in section 1.4 of LEMMA were quite easy, and much of it is testing the same knowledge. A notable exception is **Q1.4.6iii** which was predicted by one of the raters to have a 5% facility. The data show a facility of 19%, which is still fairly low, in particular when compared to the two questions with similar content that follow Q1.4.6.iii.

How would you approach the following research question?
Is there a relationship between whether pupils are required to wear uniform and the incidence of graffiti and vandalism to school facilities?

pupils (level 1) and schools (level 2)

pupils (level 1) and teachers (level 2)

teachers (level 1) and schools (level 2)

pupils (only)

schools (only)

The mention of 'pupils' in the question stem is likely to have caused learners to make a connection between wearing uniform and individual pupils, and then to think of this as a pupil level variable when it is in fact a school level variable. This is an important distinction and we surely want to make sure learners have understood how to identify the level of a variable, but it could be argued that the wording of the question caused the wrong concepts to be activated in the minds of even those learners who knew the answer, thus measuring them unfairly.

It would be interesting to see how different wording would have affected the results. For example instead of 'whether pupils are required to wear uniform' we could have 'the wearing of uniform' or even 'the wearing of school uniform'.

Would the latter reduce the demands too much? This issue turns on identifying those demands that we want to assess and eliminating those that are *construct irrelevant*. Do we want to know whether the learners can identify the correct approach when the research question is posed in a misleading way, or do we want to present the research question as clearly as possible and simply test their ability to identify the correct approach from this?

Q2.1.2iii proved very difficult with only 11% correct. It is a calculation question whereas many of the others are multiple choice so the difference in format could be causing some of the extra difficulty. The use of the numbers 201, 251 and 301 may be a distraction as this could cause confusion over whether to use n-1 or n in the calculation. If the learners were not using calculators then using these numbers makes sense, but if they were then it introduces an irrelevant distraction.

Q2.3.3ii This question mentions three explanatory variables: Subject Area, Term Time Job and Year of Study. The table does not include Year of Study, which will cause some learners confusion as they re-read to see if they have misunderstood the question. This question was a lot harder than predicted, with only 10% correct.

One issue is that the order of the options is not logical – options 2 and 4 should appear next to each other. The critical difference learners have to identify in this task is whether or not you can tell if Hours of Study is homoscedastic or heteroscedastic with respect to *one* of the variables, as one of them is clear and the other one not so obvious. This is a subtle distinction and a more logical ordering of the options would make it simpler to work out – reducing construct irrelevant demands without reducing intended demands.

Q3.3.1iii This is a two-option multiple choice question with learners scoring only 47% correct as compared to 97% and 90% on the previous two questions in this section.

> Is the following statement true or false?
>
> R^2 is the square of the correlation between Y and X
>
> **Please click one answer**
>
> **true**
>
> **false**

The key to it is that the question is referring to multiple rather than simple regression. However, the question itself does not mention multiple regression and is therefore misleading. It can of course be argued that this is a section specifically on multiple regression so the learners should realise that this is what

the question refers to. However, it may be unfairly misleading not to make it as clear as possible by stating it in the question.

Q3.5.1 was another difficult question (19%). It looks like the demand is to decide between heteroscedastic and homoscedastic. The other options can be eliminated fairly easily, some through understanding, but some through simple logic. Once option 1 has been eliminated, options 5 and 6 are eliminated. As we are told only to select one option we might conclude that only one is correct. Therefore option 1 can be eliminated, as if that is correct then so are 5 and 6. Similar logic applies to options 4 and 7. That leaves options 2, 3 and 8. As this is not a test of logical reasoning, it probably would have been better to have fewer options without interdependence amongst the options. In some cases the logical elimination of options will of course make a question easier. In others though it could cause difficulty by focusing the learner's attention on the logic rather than on the concepts being assessed. The point is that in either case it is introducing construct irrelevant variance.

Q3.5.4 was extremely difficult, with only 6% correct. Again the option combinations are complex. This question covers similar concepts to Q3.5.1 and I feel here that the instruction is not really clear. This same issue may also be contributing to the difficulty of Q3.5.1. The question asks learners to 'Please select only one interpretation', with the emphasis on the 'one'. The main point however is that learners need to select the *best* interpretation. One or more of the interpretations may be correct, so the judgement is not which is correct but which is the best or the most accurate or the most detailed interpretation. It is possible that some learners are selecting the first correct interpretation that they read, assuming that if they have found a correct one they can eliminate the others, which is very often the case in multiple choice questions. There is nothing wrong with the idea of asking learners to pick the best of the options in this way. It just needs to be very clear in the question stem that this is what is required. The feedback does indeed make it clear that some answers are 'partly correct':

> Partly correct
>
> Yes, the residuals are heteroskedastic. However, we can also see from the graph that there are some very large residuals (values of around 4) which suggests the presence of outliers.
>
> The fully correct answer is therefore "there appear to be several outliers, and the residuals are heteroskedastic".

Q5.1.2 This question was difficult for a three option multiple choice question (20.7% correct).

> Suppose that for dataset A of Question 1:
> <Graph>

we estimate the school residual u_1 for school 1.

What would happen to the estimate of u_1 if more students were sampled from each school, but the distribution of exam scores within all four schools remained the same? (Assume that the proportion of the total sample in each school also remains unchanged.)

It is possible that it functioned as a bit of a 'trick' question as the better learners may not think of shrinkage being the important issue here, even if they have understood all of the concepts. They may be focused on an idea that the estimate would not change with sample size and not think about the accuracy of the estimate changing due to shrinkage. This may be considered to be a valid demand if the concepts are clearly outlined in the lesson, but it would be interesting to know the effect of some highlighting with the question to point learners in the right direction. It would also be interesting to know the reasoning of learners who choose a wrong answer, and it may be that this issue does not lend itself well to multiple choice testing but would address the key points better as a short answer question.

Q5.5.5 is another very difficult question given that it is a three option multiple choice question (23%). The question stem starts with a long sentence with some quite awkward wording, which may be contributing to the difficulty.

The likelihood ratio test statistic for a test of the null hypothesis that the between-student variance is the same for social classes 2, 3 and 4 (but different for social class 1) is 1 on 2 degrees of freedom.

The test statistic is the difference between the -2loglikelihood values for the models fitted in Question 4, and the model fitted in Question 1, i.e., 276545 - 276544 = 1.

Based on this test, which of the following conclusions do you reach?

The between-student variance is the same for social classes 2-4, but different for class 1

The between-student variance is the same for all social classes

The between-student variance is different for each social class

Reading demands in these questions are construct irrelevant, and there may be a way of wording this question stem that would make it clearer. The general advice on this is to make wording as close to natural language as possible, and for someone other than the question writer to read for possible ambiguities once the question has been drafted.

Sometimes there is technical language in a question that will increase the reading demands. If the understanding of the technical terms is part of what is being assessed then these reading demands are valid. The reading demands of the carrier language though should always be as low as possible.

RQ1-5 These research questions are well written. Interestingly, RQ1 proved the most difficult, and this may be the effect of a new style of question that becomes familiar by RQ5.

Complex item formats

Q2.2.2iii The instructions for this question are confusing.

> Here is the same table as in 2(i) and (ii) below showing summary statistics for the income of a sample of employees at public and private firms.
>
> <table>

1. **Which of these summary statistics tells us something about the unexplained variability of Income (before adding Type of firm as an explanatory variable)?**
2. **Which tells us something about the unexplained variability (after adding Type of firm)?**
3. **Which tells us something about the explained variability (after adding Type of firm)?**

To answer these questions, match the correct description to the summary statistics fiven next to A, B and C below.

> **A: 8,100**
>
> **B: 10,100**
>
> **C: Employees in Private firms earn on average £13,500 more than employees in Public firms**

The sentence 'Which of these summary statistics…' is the first question learners are asked but they then have to read the sentence a few lines below: 'To answer these questions, match the correct description to the summary statistics ….' in order to see how to answer the question. They then have to look at the three options A, B and C and match these to the three questions, taking into account the information on 'before' and 'after' given in brackets. The whole exercise is one of complex deciphering, which is not the intended demand. There is also a typing error in the question (highlighted). A simpler way to present this would be to use the three questions in the matching boxes without first stating them as questions 1-3. Alternatively, the matching boxes could be left out completely and options A-C put under each of the three questions presented separately.

Q2.3.1 is another matching question.

> A researcher is interested in whether increasing the number of policemen in an area causes a decrease in the crime rate for that area.

> As a preliminary part of tackling this question she draws a graph to look at the relationship between the number of policemen in an area (measured at the end of 2004) and the number of recorded incidents in that area (during the whole of 2005).
>
> She divides the areas into those with high levels of deprivation (which she plots in red) and those with lower levels of deprivation (which she plots in black).
>
> Below are shown different possibilities for what the plotted graph might look like.
>
> **Match the description of the effects to the graph.**

The format of this matching question is a little different from most of the others: the learners have three descriptions and six graphs so they have to conclude that each description could be used more than once. It may be that some learners thought they just had to pick the three graphs that matched the three descriptions.

It is also quite difficult (20% correct) to get the three available marks as learners have to get all six correct to get the marks – one mistake causes you to lose all three marks.

A minor but important point is the use of the word 'policeman' in the question. It should really be 'police officers'. This is the sort of thing that will cause some learners' attention to be distracted from the business of answering the question, causing the activation of irrelevant concepts, and so should be avoided.

Q2.1.4iii also proved quite difficult, with 24% correct. Could the A and B in the table be confusing as the options are then labelled A, B and C? Choosing among the options is quite complex in this question: the learners have to pick up on certain cues. This may mean the question is difficult for the right reasons. For example, the scale going up to 200 in option A makes this an attractive choice, which directly tests whether the learners are focusing on height or area. The rationale for option B as a distractor however is not immediately clear.

Q3.3.5 Three graphs, G1, G2 and G3 are presented and learners have to match equations E1, E2 and E3 to these. The graphs appear in order but the list of graphs to match appear as Graph 2 then Graph 1 then Graph 3 and the equations in the drop down boxes appear as E2 then E3 and then E1. The consequence of this will be to confuse some learners, possibly causing them to mismatch as the order is not natural, but certainly causing them to spend time untangling the order, none of which is construct relevant.

Q5.1.3 This question is presented in a highly complex format. Learners are asked to choose amongst options A,B and C as follows:

> Suppose we calculated a 95% confidence interval for the residual for school 1 in dataset A of Question 1
> Dataset A
>
> <graph>
> If more students were sampled in this school, would the interval:
>
> **A** - stay the same
> **B** - become wider
> **C** - become narrower?

They are then given another instruction and nine more options:

> **Please answer by picking the appropriate letter and a reason for your choice**
>
> **A** Increasing the number of sampled students will **not affect** the standard error of the estimated school residual
>
> **A** Increasing the number of sampled students will **increase** the standard error of the estimated school residual
>
> **A** Increasing the number of sampled students will **decrease** the standard error of the estimated school residual
>
> **B** Increasing the number of sampled students will **not affect** the standard error of the estimated school residual
>
> **B** Increasing the number of sampled students will **increase** the standard error of the estimated school residual
>
> **B** Increasing the number of sampled students will **decrease** the standard error of the estimated school residual
>
> **C** Increasing the number of sampled students will **not affect** the standard error of the estimated school residual
>
> **C** Increasing the number of sampled students will **increase** the standard error of the estimated school residual
>
> **C** Increasing the number of sampled students will **decrease** the standard error of the estimated school residual

Such a complex format is likely to cause all kinds of unpredictability in the sorts of mental models the learners will build of the task. There is a great deal of reading and as it is repetitive it is not easy to read. When language is unnatural it becomes difficult for learners to spot what the question is actually asking them to do. A solution to this would be to present the question over two screens so that in the first screen the learners make a choice amongst A, B and C and in the second screen they choose the appropriate reason out of the three possible

reasons. Having said this, they did perform well on this question (75%). Might they have done even better with different presentation?

Clues in options

PR12 This question uses of 'all of them' and 'none of them' amongst its options.

> 12) In a regression analysis of respondents' education (*Y*) on fathers' education (*X*), the test statistic associated with the slope (calculated as the estimated slope divided by its standard error) is 17.05. The critical value for a two-sided test at the 5% level is 1.96.
>
> **Which of the following statements are true?**
>
> A: The null hypothesis is that there is no relationship between X and Y in the sample.
> B: The p-value for the test will be less than 0.05.
> C: There is strong evidence of a relationship between X and Y in the US adult population.
> D: There is evidence of a strong relationship between X and Y in the US adult population.
>
> **Please choose one answer below showing the correct combination of true statements.**
>
> A, B and C
>
> all of them
>
> B and C
>
> none of them
>
> A and D

Using 'all of them' or 'all of the above' is problematic and is rarely recommended. In this case, one recognised false statement eliminates this option. The danger is that answering the question becomes an exercise in logical elimination rather than understanding of statistics. Similarly, the option 'none of them' can be eliminated as soon as one statement is recognised as true. However, 'None of the above' can sometimes be useful, especially in calculation questions if examiners want to avoid the problem of learners getting a second chance if their calculation goes wrong.

It should be noted that the limitations of the software used to present the questions led to some items having more options and more wordy options. For example, it was not possible to ask questions where more than one of the options was correct. The software also randomised the order of options for some questions, but this has been rectified for future modules. This brings up an

important general point for question writing in e-assessment: the particular characteristics of the delivery system should afford the best possible presentation of the questions. We must be aware of the dangers of question presentation format being dictated by software limitations.

In PR12 we have a situation where more than one option could be true. Option A in this question must be read very carefully to spot that it says 'sample' and not 'population' and is therefore false. As this is a Prerequisite Test question with feedback, that is fine: it reminds the learners that they need to pay close attention to the critical distinctions in the options, and read them carefully. A question can become a test of careful reading rather than of statistics.

Q2.5.2 The correct option being twice as long as the other three in this question gives learners a clue that this is the correct answer.

> **What does a standard error measure?**
>
> The dispersion of values in a sample
>
> The dispersion of values in the population
>
> The dispersion of values of a sample statistic across all possible samples of a given size taken from the same population
>
> The error in the standardised sample statistic (Z-scores)

This is a common problem for multiple choice question writing, and can be difficult to avoid, as the key option needs to be accurate and is therefore often longer than the distractors.

Q3.2.2 In this question the options are qualitatively quite different:

> Same as before, ie: -4
>
> I'd have to fit the new model to find out!
>
> 2
>
> 4
>
> -2

In this situation learners pick up on clues about which options may be correct or incorrect and the measurement of that kind of reasoning creeps in and causes variance (**Q3.2.1** is similar in this respect).

Q5.2.1 does not appear to test knowledge of 'between-area variance' as it is a simple case of matching the graph that looks largest with the word 'largest' and the one that looks smallest with the word 'smallest'. The demands being assessed here are more trivial than those intended, and the facility was high at 93%.

Q4.1.2 This question uses the phrase 'do you think'.

> We are interested in assessing to what extent the difference between boys' and girls' educational achievement varies across secondary schools for 16 year old students...
>
> **Which of the following designs do you think would be most effective?**
>
> We randomly sample 5 schools and take achievement scores for 100 boys and 100 girls aged 16 in each school
>
> We randomly sample 30 schools and within each school take a random sample of 10 boys and 10 girls aged 16 and take these children's achievement scores
>
> We sample 1000 schools and take 1 boy and 1 girl aged 16 from each school

There is some danger in asking for an opinion in a question that then feeds back answers as 'correct' or 'incorrect'.

Marking

Another question worth looking at is **Q1.4.5**. This is a matching question in which learners have to match 5 different descriptions to 7 questions. The score is 5 marks, and they have to get all 7 right to get the 5 marks. Just one error in answering the 7 questions will result in zero marks, and this is contributing to the difficulty of the question (40%).

The effect of awarding either 5 marks or zero for this question is that it is not discriminating as well as it could. If we consider this to be one demand that we want learners to meet, and for this reason mark it as either all 5 marks or zero, then the effect is to weight this question as five times more important in the test that one for which one mark is given.

Q4.1.1 This question consists of six yes/no option choices. There are five marks available, and one incorrect response out of the six will give a score of zero for the learners. First, it is not clear why there should be five and not six marks available for the six demands as the answers are not interdependent. Secondly, a score of zero for one error on this type of question may be discouraging for learners. However, as this is not a formal test, the feedback is more important than the score, and the question can be attempted again following the feedback which is comprehensive and well written:

> Citizens and areas can both be regarded as a random sample from a wider group (of citizens and areas respectively) and therefore could be treated as a random classification.
>
> Voting intention, gender and ethnicity each have only a limited number of categories and can not be regarded as a sample from a wider population of

voting intentions, genders, or ethnicities. Of course, it would be possible to include more categories than we have done here- for example Voting intention could be expanded to include the BNP, Respect, and so on. Ethnicity could be more finely divided, so that for example instead of having one category Black, we had three consisting of Black Caribbean, Black African and Black Other. It could even be argued that psycho-socially and biologically there are more than two genders. However, even if we expand the categories of these variables in this way, there will still be a limited, if larger, number of them. More importantly, the target of inference for each of these variables is the specific categories we are using. We are, for example, interested in what we can say about how the voting intentions of black and white people differ.

It is not sensible to treat Age as a level either. This is partly because it is not a classification, but a continuous variable.

Summary of findings

To summarise, the main issues arising in a qualitative analysis of the LEMMA questions were:

- question wording, that is overall reading demands,
- the use of negatives and double negatives,
- the clarity of instructions,
- complex item formats leading to deductive reasoning and
- options that might clue learners to the answer.

The above issues are all common problems in multiple choice question writing (see Ahmed & Pollitt, 2001). Multiple choice questions are the most difficult question type to write as the task being conveyed to the students is so highly constrained.

The issue of reading demands getting in the way of the question communicating the task effectively is a very common one in question writing in general (see Pollitt & Ahmed, 2000). Reading demands can be reduced by focusing the learner's attention using the highlighting of key points.

Another issue that arises often in the qualitative analysis of test questions is the use of a real world context for the task. Real-world contexts can be very powerful as they can cause activation of many concepts in the learners' minds, most of which will be irrelevant to the construct, and so must be used with great care (Ahmed & Pollitt, 2007). In LEMMA there are a few contexts and usually more than one question relating to each of these. The contexts are all relevant to the concepts in the questions, and provide a setting for the datasets in which appropriate questions can be asked. This is a good way to use context and it is effective in this case.

If question writers are aware of the above issues they can try to avoid construct irrelevant demands and ensure that all learners are attempting the intended task. In the next section I will explain a systematic method for question writing

which may prove useful, and then summarise the general advice for the writing of multiple choice questions.

Advice and recommendations

The Question Writing Process

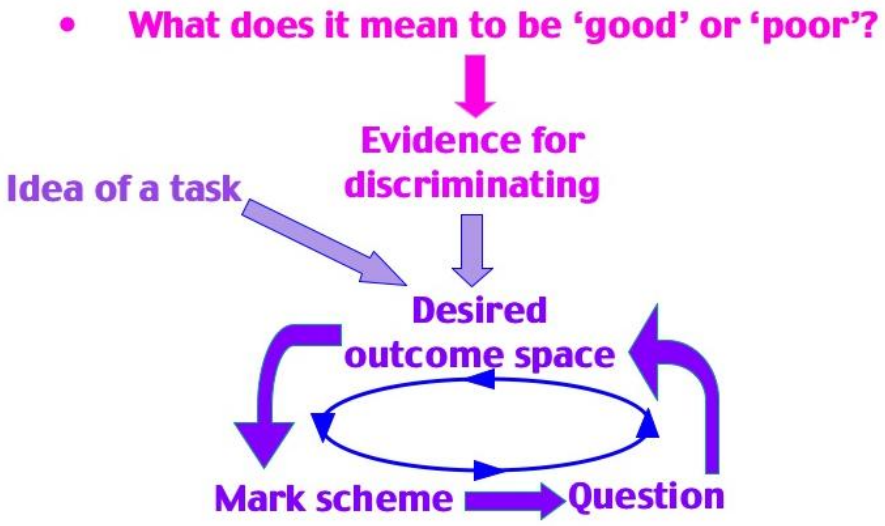
How do we go about writing questions that communicate the task clearly and provoke the right concepts to be activated in the students' minds? That is, how do we write questions that contribute to valid assessment, by measuring the intended construct?

Central to any assessment is the concept of *evidence* (Mislevy, Steinberg & Almond, 2003). In the case of LEMMA, what evidence do we need from these tests? What evidence will tell us and the learners what they have learned and what they need to work on? What do we want the learners to show us they can do? The evidence that we want learners to produce is sometimes called the *Desired Outcome Space* (Pollitt et al, 2008): it is the collection of answers we would *like* to see in order to be able to gain the information we need from the test.

With an idea of a specific task in mind, the next step is to consider how to *evaluate* the evidence we will get - and this means thinking about the mark scheme. What are the answers we will give credit for and what will we not give credit for? Is there a rule or principle to distinguish these sorts of answers? Are there example answers that would be helpful to markers?

Following this the question writer must consider how to word the actual question in order to elicit the sort of evidence that is needed. There is then a process of iteration between question and mark scheme until all are satisfied that the wording of both is as it should be.

The following diagram illustrates the question writing process. It begins with the notion of what it is to be good (and poor) at the learning being assessed (i.e. the construct). Once this is established we consider what evidence is needed and this gives us the Desired Outcome Space, which, alongside the idea for a task, we use to inform the writing of the mark scheme and the question. We call this method Outcome Space Control and Assessment (OSCA) (Pollitt et al, 2008).



Writing multiple choice questions

When writing multiple choice questions the process is similar but the options take the place of the mark scheme. When considering what distractors to use in the question it is important to consider what errors learners might make in terms of the construct. We do not want the distractors to test for reading errors or to give 'trick' wrong answers. We are interested in whether the learners can identify the correct answer amongst other incorrect answers that are plausible in terms of the *conceptual understanding being assessed*. And in order for the test to be valuable for learning, we want the distractors to show us and the learners where they have gone wrong in their learning, and where they might need help.

Thinking about the Desired Outcome Space – the sorts of answers we would want the learners to produce in order to measure their learning if the question were **not** multiple choice - can allow us to produce appropriate distractors. We call this method the *Outcome Space Generator* (Pollitt et al, 2008). It involves thinking like a student and trying to predict students' answers. That is, the question writer works through a question and considers first the Reading phase – what will a student's mental model of this task be like? They then consider the Thinking phase – what concepts will be activated by the question? Are these the concepts that will allow us to measure the construct? What incorrect concepts will be activated? If they are incorrect for construct irrelevant reasons, we can use these to help us to write distractors.

Much of the advice for question writers based on students' cognitive processes applies to all question types. However there are some issues specific to certain question types and I will outline those specific to multiple choice as these form the bulk of the LEMMA questions. Haladyna (1997, 1999) has written extensively on multiple choice item writing and some of his advice can be seen in the context of our models described above.

Writing good multiple choice questions depends greatly on choosing appropriate distractors. As outlined above, this process can be improved by considering the types of errors we would expect learners to make if the question was not multiple choice, and using the errors that indicate a lack of the construct.

When considering the student reading the question it is obviously important but not always easy to ensure that the grammar of the stem matches that of all of the options. It is also advisable that the key idea in the question is contained in the stem. The function of the stem, above all, is to help learners to build a mental model of the problem; there should be no reading difficulty in this process. The stem should tell them what the point of the question is: usually, they should be able to generate the answer before looking at the options. Most authors prefer the stem to be a complete question, but an incomplete sentence can be perfectly acceptable. The general rule is to include as much information in the stem and as little in the options as possible.

The other side of this coin is to include no *unnecessary* information in the stem. Again, as learners are building a mental model of the task we do not want them to be distracted by irrelevant information. The language used in the question, to convey the task, is powerful and the focus of the learners' attention when they are reading can be manipulated by using highlighting to indicate key points.

The number of options is another important issue when writing multiple choice questions. There is no 'right' number of options. Recent research (Rogers & Harley, 1999) has shown that if a question has three options it is not worth the question writer struggling to find a fourth option: it is often obviously wrong so nobody chooses it and the internal consistency of the test is not affected. In such a situation even two options can be acceptable, though in terms of gaining an overall picture of test performance it is unwise to mix two-option items with another format. LEMMA has a real mix of number of options and of question formats. This is fine as part of a learning resource if we are not concerned with measuring each learner across all of the tests.

A sensible approach to selecting the number of options is therefore to write options that come naturally through thinking about the outcome space, and not to try to write any more. In some cases there will be an obviously natural number of options. For example if the answer is a day of the week, say Tuesday, then we would expect all of the other plausible days of the week to be amongst the options. If they are all plausible then there is no point in leaving one or two out, and the question will have seven options. Equally if a question has three possible answers then there is no point adding a fourth that will inevitably be hardly ever chosen.

There will sometimes be a logical or natural order in which to present the options, and this should always be used. Going back to the days of the week example, there is no reason to jumble them up: this just adds unpredictability to the reading of the question.

Another important issue linked to this is to consider whether there are any unintended clues to the key option that are given by aspects of the question. For example, the key is sometimes longer and more complex than the distractors as

question writers try to ensure that it is as accurate as possible. The aim is that the options should be qualitatively parallel as far as possible. It is also important to point out that distractors must be plausible but not correct, unless the question clearly states that students should select the best answer rather than the correct answer.

As with all item types, the use of negatives should be avoided whenever possible. When writing multiple choice questions it is easy to slip into using double negatives by using one in the stem and another in one of the options. Negative words are difficult to read especially if reading under conditions of stress during a test. The word 'not' is easy to miss, and usually the question will make perfect (but wrong) sense without it. Where negatives cannot be avoided, they should be highlighted. This should include words such as 'few', 'never', 'sometimes', 'hardly' as well as 'not'.

Complex item types are another possible cause of construct irrelevant variance. We are aiming to assess how well a learner understands and can use their subject knowledge, in this case multi-level modelling, so we do not want the learners to be using deductive logic to answer the questions. For example, questions in which there is an overlap in meaning, or some other interaction amongst the options, are vulnerable to this. The danger is that a complex format can be measuring intelligence rather than learning of the topic being assessed.

When writing any test question, it is of paramount importance to ensure that the task is communicated as clearly as possible. The LEMMA questions are in general very well written and include valuable feedback. If the question writers are aware of the issues outlined above, the LEMMA questions should give a clear picture of the learners' understanding of statistics and multi-level modelling.

References

Ahmed, A. & Pollitt, A. (2001) Science or reading? IAEA, Rio de Janeiro. Available from http://www.camexam.co.uk

Ahmed, A. & Pollitt, A. (2007) Improving the quality of contextualised questions: an empirical investigation of focus. *Assessment in Education: Principles, policy and practice* 14, 201-233.

Ahmed, A. & Pollitt, A. (2010) The Support Model for Interactive Assessment. *Assessment in Education: principles, policy and practice* 17 (2) 133-167.

Ahmed, A. & Pollitt, A. (in press, 2011) Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: principles, policy and practice. Special Issue on Marking.*

Crisp, V. Sweiry, E., Ahmed A & Pollitt, A. (2008) Tales of the expected: the influence of students' expectations on question validity and implications for writing exam questions. Educational Research. 50:1, 95 - 115

Haladyna, T. M. (1997) Writing test items to evaluate higher order thinking. Needhan Heights, MA: Allyn and Bacon.

Haladyna, T. M. (1999) Developing and validating multiple choice test items. Routledge.

Mislevy, R.J., Steinberg, L.S. & Almond, R.G. (2003) On the Structure of Educational Assessments. Measurement: interdisciplinary research and perspectives, 1(1), 3–62.

Pollitt, A & Ahmed, A (1999) *A new model of the question answering process*. IAEA, Bled. Available from: http://www.camexam.co.uk.

Pollitt, A & Ahmed, A (2000) *Comprehension Failures in Educational Assessment*. ECER, Edinburgh. http://www.cambridgeassessment.org.uk/research/confproceedingsetc/ ECER2000APAA

Pollitt, A, Ahmed, A, Baird, J-A, Tognolini, J and Davidson, M (2008) *Improving the quality of GCSE Assessment*. London: QCA. Available from: http://www.camexam.co.uk.

Pollitt, A., Ahmed, A. and Crisp, V. (2007) The demands of examination syllabuses and question papers. In Newton, P., Baird, J., Goldstein, H. Patrick, H. and Tymms, P. (Eds.) (2007) Techniques for monitoring the comparability of examination standards. London: Qualifications and Curriculum Authority.

QCDA (2011) Mathematics. The National Curriculum for England. http://curriculum.qcda.gov.uk/uploads/Mathematics%201999%20programme%20of%20study_tcm8-12059.pdf

Rea-Dickins, P, Afitska, O, Yu, G, Erduran, S, Ingram, NR & Olivero, F. Investigating the language factor in school examinations: exploratory studies, SPINE working papers no.2, Report for Study 5.1, for ESRC, DFID, 2009. ISBN: 9781906675912.

Rogers, W.T. & Harley, D. (1999) An empirical comparison of three- and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. Educational and psychological measurement. 59 234-47.