What is a good maths assessment?

Sarah Maughan, Head of Assessment Policy, NFER, UK

Maths assessment in its current form poses a dilemma for test developers and maths educators alike: on the one hand its measurement properties are good, the marking is straightforward and the results reliable, on the other hand it is criticised for not assessing the essential skills or encouraging real understanding of the maths being assessed. Recent changes to the curriculum for both primary and secondary schools in England have been designed to give greater flexibility to teachers, and to encourage them to bring more creativity into the maths classroom. New approaches to testing and examining could both exemplify what this means in terms of changed outcomes and encourage new approaches to teaching and learning to be adopted.

The National Foundation for Educational Research has conducted a short research project to try to answer the question 'what is a good maths assessment?'. The project was made up of two stages: in the first current maths tests and exams were analysed in terms of the content included, the skills assessed, and the nature of the questions used. The second stage involved interviews with maths educators and experts in England to collect their views about how maths should be assessed. Conclusions have been drawn about the features of good maths assessment.

This presentation will describe the results from the research and present some suggestions for good maths items. It will also present some examples of items that ought to be avoided.

Introduction

The theme of 2009's IAEA conference is Assessment for a Creative World. This is a crucial theme that is stimulating debate around the world: what are the new skills that will be needed for the 21st Century? How should they be assessed? What are the implications of this for teachers and learners? However, the theme applies equally to existing curriculum subjects and assessments. Is the way that we assessed these core subjects in the 20th century still appropriate for the 21st century? Do new technologies provide us with opportunities to do things differently? Does the world in which our young people will find employment require a change in focus in what is taught and assessed in schools and colleges? National curriculum and assessment systems around the world are going through a period of rapid change, and England is one example of a country in which this change has been dramatic and extensive (see Whetton, 2009 for a brief history of the national curriculum in England). There has been change, or change is currently being implemented, at every phase of the curriculum: early years, primary and secondary, and changes to the assessment systems at all stages (see the Qualifications and Curriculum Development Agency website for details of the system and the changes: www.QCDA.gov.uk).

This change is also apparent in mathematics education. A new curriculum for lower secondary education was introduced in September 2009, proposals for a revised primary curriculum will see a move away from single subjects to learning areas, one of which will be mathematical understanding (Rose, 2009), new specifications are being introduction for the school leaving examinations (GCSEs) in which functional maths will become a core element (testsandexams.qcda.gov.uk/). This change within the system opens up an opportunity to 'do things differently'. So what should maths assessment be like? The National Foundation for Educational Research recently conducted a small research project in which we reviewed the content and skills assessed in current maths tests, and also interviewed a number of maths experts, including from maths teacher education and maths higher education backgrounds, as well as maths test developers, and asked them what they think are the features of 'good' maths assessment. This paper presents a brief overview of the findings of this research.

Maths tests include items assessing different areas of the curriculum and different skills and understanding, to a greater or lesser extent. The correct balance between items assessing mathematics content knowledge and mathematical skills is a topic for much debate. This research aims to provide evidence to inform this debate.

Review of Existing Maths Tests

For the review stage of the project we asked a number of maths assessment experts to classify the questions included in current tests and examinations against content and skills categories. This task, by its nature, is subjective and other reviewers may well have come up with slightly different categorisations, however, the results provide a guide of the types of items that are included. The review covered national curriculum tests at key stage 2 and 3, tests from a publisher (one primary and one secondary), school leaving examinations (GCSEs from each of the three boards in England) and basic skills assessments of numeracy (key skills tests), at the end of 2008 (using that year's tests where appropriate). The tests covered a range of age groups and a range of purposes, and as would be expected they included a variety of different item types and a different balance between content and skills assessed. It is not possible to give the results in detail here but a number of themes emerged which are described in brief below.

The tests were classified in terms of the content areas assessed. All of the tests included questions assessing aspects of number, such as counting, using number facts, understanding number and calculating. The tests also included items assessing shape, measures and data handling. At the higher levels the tests also included items assessing algebra, geometry and statistics.

In most cases the tests included a majority of questions assessing number. Some of these questions were limited to straightforward recall of number facts or simple calculations, however, many of the tests included more abstract questions, such as 'fill in the gap' number calculations. This was considered to be positive as some of these abstract number questions challenge bad habits that students may have learned from working on repetitive worksheets. For example, a question may prove difficult for a learner who holds the misconception that an 'equals sign' demands a single number on the right of it as an *answer*. Rather, students must realise that the sign indicates that the two sides of a calculation are equal and each side can be made up of numbers as well as expressions.

Some of the shape questions asked students to visualise aspects of the shapes. This makes them more interesting questions than just, for example, recall of properties. A good example of a 3D shape question is one where students are asked to identify which of a set of nets would fold to make a pyramid. There were questions assessing pupils' understanding of such shape properties/concepts as the number of faces/sides/equal sides; rotation symmetry; line symmetry; area and perimeter.

In the questions targeting measures, there were few questions which actually ask students to measure anything physically (length, area and angle would all be possible in a written test). Questions for primary students focus on reading of partially numbered scales, conversion of units (minutes to seconds, litres to millilitres). Some questions were more complex and required students to bring together, for example, understanding of perimeter with knowledge of the properties of a square, and some calculation and logical thinking in an interesting way.

Data handling questions were fairly well represented (in general over 10% of the total marks in the tests reviewed). Various forms of charts and graphs were used to present information.

However, there were few questions that ask for any real interpretation/evaluation of the data, or ask students to extrapolate.

The questions were also reviewed in terms of assessing different skills. For the primary tests the questions assessing the using and applying strand of maths were considered, that is the questions assessing: solving problems, representing, enquiring, reasoning and communicating. Several questions were found which required students to solve problems, a small number required students to reason and communicate, but no questions required students to represent or enquire.

The tests for secondary students were reviewed in a similar way, using the mathematical processes and applications strand of the programmes of study. Again this analysis found only a relatively small number of questions assessing the higher level skills and almost no questions assessing extended reasoning.

A quote from one reviewer reflects the findings across the tests and examinations: '*However, the questions are limited in that children are not asked to do anything with their answers. They are not asked to interpret, discuss or analyse their answers in any way*.'

Other themes emerging from the review are given below. A number of these are picked up in the section on the interviews later in this paper.

Multi-step problems

It could be argued that questions with multiple stages may require the students to use higher order skills and so the reviewers were asked to consider questions with multiple marks to assess whether this was the case. A number of questions asked students to 'show their working' and marks could be achieved for this even if the final answer were incorrect. This is a valuable technique to encourage in students, although it may not demonstrate the use of higher order skills.

Other examples of multi-mark questions required more than one step but generally did not require the use of different strands of mathematics in an integrated sequence of reasoning. Only a very limited number of questions were found which were considered to require extended reasoning.

More than one correct answer

In general the papers had very few questions where there is more than one correct answer. As one of the reviewers suggested '*it is not surprising that questions with a single correct answer have been given preference because they are relatively straight-forward to mark consistently and reliably. However, they tend to have a slightly limited scope for assessing higher order thinking and pupils' skills in coping with situations where there are a variety of correct mathematical solutions or interpretations to the problem.*'

'Explain why'-questions were used in a number of the maths tests reviewed. These clearly have multiple correct answers and generally demand that the students demonstrate their understanding of a particular area of maths. Questions of this type usually start with a problem scenario that the student is asked to interpret or analyse in order to produce an explanation of some key feature. Sometimes students were also asked to make a decision (eg 'yes' or 'no'; or agreeing with one of two opinions) before they explain their thinking. Usually students answer by writing, but sometimes it is possible to accompany that with drawings or calculations.

Statistics questions

Very few questions across all papers explicitly mentioned a source of real data. In addition there were other questions that might use real data, but a majority of the statistics questions used invented data. Real data can often be 'messy', meaning that there may be many variables that have to be considered, or the numbers and graphs are demanding to work with. It is also time-consuming to find a suitable context that remains up-to-date at the time that the test is used. For these reasons it is clear that using invented data is more convenient for a test developer. However, providing examples of real data (together with the source) has important benefits such as adding interest, showing relevance of maths, assessing skills in a real setting and providing ideas for further research both for students and teachers.

Interviews with Maths Experts

The second part of the study used a simple interview schedule to collect the views of a number of 'experts', including maths teacher educators, maths test developers and university mathematicians, about examples and features of good maths questions. The interviewees were sourced through personal contacts and through a posting made to a widely used maths education discussion list in England. We particularly requested interviewees with both maths teaching or maths teacher education experience, and with an interest in assessment. Eleven interviews were conducted in total, one of the interviewees had also acted as a reviewer during the previous stage of the project.

'Good' Maths Questions
As part of the research design we decided to ask about 'good' and 'bad' questions. We purposely did not define in what way the questions should be 'good' or for what the tests were to be used, although we did say that we were investigating summative, paper and pencil tests. It was felt that this would allow interviewees to approach the questions from their own perspective. Two interviewees did raise the purpose of the tests in their responses, stating that this would affect the nature of the questions that should be included.

Two examples of questions cited by those interviewed as being 'good' questions are provided on the next page. The examples are not given in full and have not been developed in the way they would be for a real test, rather they are included just as they were described in the interview. The reason why they have been considered to be good has also been included. After asking the interviewees to describe a particular example of a good maths question they were then asked to generalise from that to come up with defining features of good maths questions. These features are grouped and discussed below.

❖ Features of Questions

A number of the features related to the nature of the questions, in particular it was felt that the questions should be open ended (6 mentions), and related to this it was also suggested that the questions ought to allow scope for creativity (1 mention), in reality these two may be facets of the same feature, in both the student is expected to come up with their own approaches to answering the question. The review of current maths tests described earlier concluded that few maths questions had more than one correct answer, never mind being open ended. A similar feature is that the questions should be multi-part, providing students with scope to develop their thinking (2 mentions). Again, in our review few multi-mark questions requiring multiple steps or development of ideas were included. Finally under this section,

five interviewees suggested that links should be made across different areas of maths in a good question, rather than compartmentalising different strands of mathematics.

❖ Context

A particular theme that came up a number of times in the interviews related to the context of the question. A number of interviewees stated that not all questions need to be given in context, and some without context are useful for assessing procedural knowledge or technical aspects of mathematics. This may be useful for ensuring coverage across the maths curriculum. However, where a context is used it should be interesting or unusual (5 mentions), it should use real life situations rather than artificial (3 mentions) and it should be relevant to the students' experience (1 mention). Our review of current maths tests found that there were few examples of real life contexts at least in the area of statistics.

❖ Design of the Question

A number of features related to the assessment design of the question. For example it was felt that a good maths question would allow teachers to prompt the student during their answer (and for this to be taken into account at the marking stage (2 mentions)). Similarly, it was felt that teacher judgement should be included in the assessment process (1 mention). On a number of occasions it was suggested that good maths assessment should be more like assessment in English, history or art, although one interviewee commented that assessment in maths is different to these subjects in that it is possible not to know how to start, or to get stuck with a certain part of, a maths questions and this is not so much the case with the other subjects mentioned. By allowing the teacher to prompt the student and incorporating comments about this or teacher judgement in the assessment process, this difficulty with open-ended maths assessment is minimised.

It was also felt that no time limit, or a long time limit, should be given for the students to respond so they are able to develop their thinking (this relates to the feature of open ended questions (1 mention)). A number of related features include: that the question should promote good practice (3 mentions), that it should stimulate ideas for the classroom (2 mentions) and that the students should learn something from responding to the question (3 mentions). Finally, it was suggested that access to resources such as the internet or books should be allowed to research the answers to good maths questions (1 mention). This would support a number of the other features described above such as more open ended questions, allowing creativity and different ways of responding, and without a time limit.

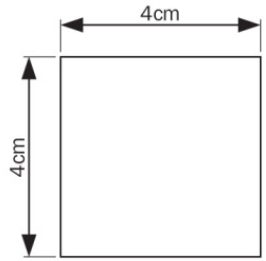
❖ General Features

Two of the given features were very general, including that a good question should assess what it is intended to assess (that it is valid) and that it should be 'interesting'. Obviously these are both true, although it is difficult to relate these features to particular maths questions.

'Bad' Maths Questions

A number of examples were given by the experts of what they considered to be 'bad' maths question and two of these are reproduced below. As with the good questions above these are not always provided in full, they are provided as they were given in the interviews. The explanation of why the example was given is also included.

Q1.



The square above is an **equitable shape**, that is it's area is the same as it's perimeter.

a) Draw **two** more equitable shapes

b) When does a shape have the same area and perimeter?

This example was provided by two separate interviewees. It was provided as a good example of an open ended question that could be answered in different ways by different students. This could allow students to be creative in their answers and allow them to each work at their own level.

Q8. Ann and John have some marbles. Ann gives 5 to John. John now has twice as many as Ann. If John gives ***x*** marbles to Ann, she has 3 times as many as John.

a) How many marbles do they each have?

b) Construct another example like this with 3 people.

c) In what way is your example like the original?

Parts b and c of this example were provided to demonstrate how mathematical thinking can be explored beyond the answers to a reasonably straightforward question.

Q1. Solve the equation $2x^2 - x - 2 = 0$

This kind of question was provided on more than one occasion. Clearly it reflects a type of question that is currently seen in maths tests. It was suggested by this group of interviewees that this kind of question merely requires the student to follow rules or procedures rather than to understand what is actually involved. One interviewee said that '*this is not maths'*.

Q3. 4x6=

This question is identical to one provided in the section on 'good' questions. There it was included to show that all questions can be good depending on the context, and that for a test for a particular purpose, this kind of question might enable broad curriculum coverage, or might act as a check that basic number bonds had been learnt. Here it was provided as an example of a question that is merely about recall of facts, not understanding.

When asked what makes a bad maths question, a number of the interviewees responded '*the opposite of the features of a good maths question'*. However, a number of features were cited and these are grouped and discussed below.

- Features of the Questions

The most popular feature of a good maths question is that it would allow an open ended response. Similarly a feature of a bad maths question is that it would allow a limited number of actions in response (3 mentions). Similar features included that the question would include too much scaffolding (3 mentions) which is related to the first point in that it could mean that the students are not free to choose their own approach to responding to the question. Related to this is that a question is too fragmented (3 mentions): the structure and prompts provided for the students remove a flow to the question. One interviewee particularly noted how criterion referencing in test design can mean that questions target very specific areas of maths only and therefore means it is split into separate topics. Finally, it was commented by five interviewees that a bad maths question is one which involves memory or the replication of processes that have been learnt by rote as a response to a given prompt.

- Context

Another feature that came up when discussing good questions and also came up when discussing bad questions was context. In a bad question the context could be obviously fabricated (4 mentions), unfamiliar (1 mention) or without a purpose in the question (1 mention).

- Language

A feature of bad maths questions that was not mentioned under the features of good maths questions was the use of language. Three separate mentions were made of use of language in bad maths questions, suggesting it could contain unnecessary academic language, unclear language or too great a language demand on the students in either the question or the response. This is an interesting feature because many of the comments related to good maths

questions would suggest that they should be longer and more open, and set in interesting and relevant contexts. However, one interviewee particularly pointed out that there does not have to be too great a language demand if open ended maths questions are well-designed.

❖ General Comments

Two comments arose which did not relate to the questions themselves. The first highlighted the incorrect use of statistics in questions and was mentioned by two interviewees. This related to statistics questions that were set in unrealistic contexts, or where the statistics were used for inappropriate reasons. The second related to the design of the mark scheme and is similar to the comments about the features of the questions mentioned above, that is a mark scheme which is too structured (1 mention) will lead to a question with a limited number of possible actions and which may be too fragmented.

Discussion

Both the stages of the research were interesting to conduct, and were conducted largely independently of each other. There is a definite mismatch between the current tests and the features of good maths questions highlighted by our experts. There are a number of features of good maths questions that are not included in all of the current maths tests that were reviewed, and may be explicitly excluded in the test design and administration instructions. These features include:

- open ended questions: none of the tests reviewed included any questions that could be classed as open ended;
- no time limit: all of the tests reviewed include strict time limits;
- teacher judgements: none of the tests reviewed include teacher judgements as part of the assessment process;
- teacher prompts: teachers are specifically barred from prompting students during the administration of the tests which were reviewed;
- access to resources: students are not permitted access to resources during any of the tests which were reviewed (other than maths instruments such as rulers or calculators).

This mismatch may relate back to the issue of purpose, which was raised a number of times during the research. Tests to be used in a high stakes context at the end of a course of study, with the results being used for accountability purposes as many are in the English context, may need to include different items to those which are used to support teaching and learning.

A number of the features of good maths questions related to the contexts in which the questions are set. In terms of the review of the current tests this feature was not specifically investigated, although a number of questions in current tests are set in context and care is taken in the development process to ensure that these are appropriate for the target students. One of the reviewers commented that the contexts used were '*appropriate but ordinary*'. The limitation in the use of real life data in the statistics questions was highlighted.

A further feature highlighted as part of good maths questions was the integration of content from different parts of the maths curriculum. Although again not reviewed specifically, there was the suggestion that there is little content linking across different areas of the mathematics curriculum in the current tests. No examples were found in any tests in which maths from across the curriculum needed to be integrated in a meaningful way.

It was suggested that good maths questions would mean that the students learnt new maths as part of the testing process, similarly it was suggested that the questions would exemplify good practice or generate ideas for the classroom. As the questions in existing tests tend to be short and structured, within largely familiar contexts, it is reasonable to assume that there is limited scope for new learning to occur. However, the rhetoric surrounding a number of the tests would suggest that the developers do aim to exemplify good practice and to have an impact on what goes on in the maths classroom.

Good questions to our interviewees would appear, in general, to be those which assess the students' understanding of mathematical concepts rather than whether they had learnt particular facts or procedures. As mentioned earlier it is likely that this particular perspective may be down to the perceived purpose of the tests. At the end of a programme of study, and as part of a summative assessment, there may well be value in assessing whether key facts or procedures have been mastered before moving on to the next level. The correct balance between these aspects of maths assessment, and the assessment of mathematical understanding and higher order skills is likely to vary with the purpose of the particular test.

Those questions aiming to assess using and applying mathematics and mathematical processes and applications may be considered to be most similar to the types of questions generally being suggested as good questions in this research. A number of questions were reviewed which contain multi-step problems, which was suggested as a feature of good maths questions. The second strand of using and applying maths is 'reasoning and communicating'. Again this could reflect some of the suggested examples of good maths questions in which students were required to come up with their own approaches and explain their reasoning.

'Enquiring' is the fourth strand of using and applying maths. Its description is included here as it demonstrates clearly the similarity to the features of good maths questions suggested by the experts in the interviews '*Suggest a line of enquiry and the strategy needed to follow it; collect, organise and interpret selected information to find answers'*. Unfortunately there are no questions classified as assessing this strand in either of the reviewed tests at the primary level. Perhaps this is the most stark omission in the current tests for primary students, in light of what the experts consider to be good questions.

When considering mathematical processes and applications, the strand most similar to the key stage 2 using and applying strand for secondary pupils, a number of questions were categorised as assessing this. The majority of the questions relate to 'analysing using appropriate mathematical procedures' and to 'analysing using mathematical reasoning'. A small number of items test interpreting and evaluating, however, no questions test communicating and evaluating in any of the tests that were reviewed.

The analysis of the GCSE papers provided the most surprising results. It may have been reasonable to expect that these papers, targeted at older students in an academic context, may have included questions assessing higher order skills, especially in the higher tier papers. The analysis found, however, that the majority of the questions assessed computational skills with only a relatively small number of questions assessing understanding. Over all of the 2008 GCSE papers reviewed only one question was considered by our reviewer to be really assessing extended reasoning.

Conclusions

Perhaps the most interesting finding of the research is the mismatch between the questions in the current tests and the features of what our experts considered to be good maths questions.

In essence the research appears to highlight the fact that a paradigm shift is needed in maths testing before it will be viewed positively by many of the people that we spoke to.

The current tests appear to be limited in the higher level skills that are being assessed, even taking into account the constraints of a formal testing environment. The classification of questions by skill and against the application strands of the maths curriculum would suggest that more questions assessing these skills could be included in current tests.

However, there seems to be a majority view amongst this small group that maths testing ought to change to become more like assessment in English, art or history. This would involve the inclusion of more open ended questions, in which the students could come up with their own approaches to answering the questions. Teachers would be available to provide prompts if required, and this would then be taken into account in the marking process, in which teachers may also be involved. Questions would be set in real world contexts, using real (or realistic) and relevant data. Technology may be used to present this data to the students.

However, similar approaches to this have been attempted in maths testing, not least in the coursework element of GCSE. Recent changes in examination policy have marked a move away from these approaches, with a reduction in coursework in GCSE. However, the controlled assessments being introduced to replace coursework may allow questions of this type to continue to be included.

Marking in maths is currently relatively reliable compared to subjects such as English, history or art, as there is frequently only one correct answer. If there is a move to more open ended questions then there would also be a decrease in the reliability of the marking. Is this something that maths teachers or users of the results would wish to contemplate? Perhaps more significantly, a shift such as this in the assessment model would mean a significant shift in the required teaching and learning practices in the maths classroom. Is this something that the current maths teaching force would be able to achieve? Finally, a move to the kind of questions suggested in this report is likely to make the maths tests more difficult. What impact would this have on the very public debate in England about standards?

Much of the discussion about the features of good maths questions comes back again to the decision about the purpose of the tests. Where the tests are to be used in a high stakes context, to hold schools and teachers to account, and to measure standards over time, perhaps some of the features of existing tests and examinations are a pragmatic option. However, perhaps to really see a marked change in students' understanding of maths and in overall standards then some change is required. In the short term at least there may need to be a middle ground in which some of the features described above are used to develop new question types for current maths tests, and the examples of good questions that were highlighted could be used as a model to increase the examples of such questions in current maths testing.

References

Rose Report (2009). Independent review of the primary curriculum: final report. Department for Children Schools and Families, available online at http://www.dcsf.gov.uk/primarycurriculumreview/ [23 July, 2009]

Whetton, C. (2009). 'A brief history of a testing time: national curriculum assessment in England 1989–2008', *Educational Research*, **51**, 2, 137–159.