



**UNSW**  
THE UNIVERSITY OF NEW SOUTH WALES

**EDUCATIONAL ASSESSMENT  
AUSTRALIA**

# THE SECRET LIFE OF A MATHS QUESTION

AN OVERVIEW OF THE ART AND SCIENCE OF  
PRODUCING EFFECTIVE TEST QUESTIONS IN  
MATHEMATICS AND NUMERACY

NICK CONNOLLY

MANSW Conference 2005



Copyright in this paper is owned by Educational Assessment Australia, NewSouthGlobal Pty Limited unless otherwise indicated. Every effort has been made to trace and acknowledge copyright for materials used. Educational Assessment Australia apologises for any accidental infringement and welcomes information to redress the situation.

Any views expressed in this paper are those of the individual author, except where the author expressly, and with authority, states them to be the views of Educational Assessment Australia.

## Content

<b>Content.....</b>	<b>2</b>
<b>The Secret Life of Maths Questions .....</b>	<b>3</b>
<b>The Technical Language of Item Writing .....</b>	<b>3</b>
<b>The Ideal Test Item.....</b>	<b>4</b>
<b>The Fundamental Question of Test Construction: .....</b>	<b>4</b>
<b>Two Basic Approaches .....</b>	<b>4</b>
A: “What specific skills/knowledge does student X have?” .....	4
B: “How good, in general, is student X at mathematics?” .....	4
The relation between A and B .....	4
<b>Examples of Approach A and B: .....</b>	<b>5</b>
Approach A or Approach B? .....	5
Approach A or Approach B? .....	5
<b>Multiple Choice .....</b>	<b>6</b>
<b>Structure of a Multiple Choice Item .....</b>	<b>6</b>
Distractor Reasoning Example 1 .....	6
Distractor Reasoning Example 2 .....	7
<b>Distractors Affect the Tested Content of an Item.....</b>	<b>7</b>
<b>Distractor Reasoning Highlights Faults in an Item .....</b>	<b>8</b>
<b>Distractors Affect the Difficulty of the Item.....</b>	<b>8</b>
<b>Developing Better Practice.....</b>	<b>9</b>

# The Secret Life of Maths Questions

Nick Connolly October 2005

This article is an overview of the art and science of producing effective test questions. I will explain some technical language and discuss key issues. I will also give the reader some tips and guidelines from my many years of item writing and test construction experience for state testing programmes and maths competitions.

## The Technical Language of Item Writing

Like most fields of work or study, designing tests has generated its own vocabulary. Whilst technical terms can be off putting they allow for clearer discussion of key concepts.

The first piece of jargon is the word “item”. In common parlance it was what most of us would refer to as a test question.

### Difficulty

“Difficulty” is a term with many related meanings. At a naïve, intuitive level its meaning is obvious, how ‘hard’ or ‘easy’ an item is. Putting that intuitive meaning into a more scientific context is difficult and the term ends up meaning two different things that we hope are related:

- The probability that a given student (or group of students) will get an item correct. There are many measure of this and finding reliable figures is part of the science of educational measurement and the related field of psychometrics.
- Features of an item that taxes a student’s ability. This second meaning includes such things as “cognitive load” i.e. the extent to which features of an item tax a students short-term memory.

### Discrimination

“Discrimination” is the primary aspect by which the effectiveness of an item can be judged. It is a measure of the extent to which an item changes in difficulty for students of greater ability. An item with low discrimination will be nearly as difficult (or easy) for the students of high ability as it will for the students of low ability. An item with very high discrimination will be trivial for the most able students but next to impossible for the least able students taking the test.

### Guess

“Guess” is the extent to which a student who **doesn’t** have the requisite skill, knowledge or ability to answer the item can still get the item correct or have a higher than expected chance of getting the item correct. In statistical terms it is the chance of a ‘false positive’ on an item.

### Slip

“Slip” is the opposite of guessing i.e., the extent to which a student who does have “Guess” is the extent to which a student who **does** have the requisite skill, knowledge or ability to answer the item can still get the item wrong or have a higher than expected chance of getting the item wrong. In statistical terms it is the chance of a ‘false negative’ on an item.

### Type

With “type” I’m referring to the style of response the item demands from the student. Major types of items include Multiple Choice and Constructed Response (where the student must write their own response) sometimes called “Free Response”. This article will look primarily at multiple choice items.

## The Ideal Test Item

The ideal test item has:

- An appropriate level of difficulty for the test group. Follow the Goldilocks and the Three Bears principle: Not too hard, not too easy but just right! Items that are too difficult for a group are off putting for the students and also provide you with little information about their ability.
- An appropriate level of discrimination for the kind of test. This assumes that the majority of items really are good tests of mathematics. A good maths item won't discriminate well in a History test.
- As small a guess factor as possible. The reasons for this are obvious but the degree to which this can be tolerated will vary depending on the nature of the test.
- As small a slip factor as possible. As for guessing the reasons for this are obvious.
- Has content free of external biases. The item avoids ethnic, gender and other stereotypes. The item is accessible to the cohort in general except in terms of mathematical content.
- And last but not least... It tests mathematics! The content of the item is mathematics (and/or numeracy)!

## The Fundamental Question of Test Construction:

What is the Test For?

- What do you want to find out?
- How reliable do your results need to be?
- What decisions will you make based on the results? What are the STAKES?
- What are you going to report?

Tests can serve multiple purposes but a test should still have a clear objective or objectives. Be aware that some objective can conflict with each other. Remember also that the students' attitude towards the test can affect their performance.

## Two Basic Approaches

### ***A: "What specific skills/knowledge does student X have?"***

The test needs a narrow focus with several items in a group focusing on very specific skills. Each item should have a high discrimination, at least within that group of items. Each item will be of similar difficulty compared with the whole cohort. The 'ideal' item would have only two values for a probability of success, 1 and 0. In other words the fact that a student got the item right should be very strong evidence that they do have that particular skill or piece of knowledge.

The pattern of answers in this kind of test is very meaningful. The overall score is less meaningful.

Diagnostic quality: gives specific information about skills a student has.

### ***B: "How good, in general, is student X at mathematics?"***

The test needs to be broad with a range of items incorporating many skills. Items should have a good discrimination but not too high. Items can be of a range of difficulties. Degree of success of an item should vary continuously and gradually for students of different ability.

The pattern of answers in this kind of test is less meaningful. The overall score is very meaningful.

Diagnostic quality: wrong answers can reveal lack of skills.

## ***The relation between A and B***

I'm presenting these two approaches as a dichotomy. The reality is that the two aren't unrelated. We would hope that somebody who has been accumulating sets of discrete skills will get better in their overall performance in mathematics. We also assume, and not without reason, that there is a progression of skills and knowledge in our teaching and our curricula. Also we can make some reasonable

assumptions about what skills a student does or does not have by observing any test item so long as it is well written.

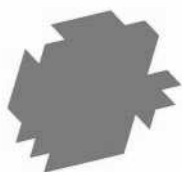
If we had a fundamental theory of mathematics learning in which we fully understood the ways in which students learn and progress in mathematics these two approaches would amount to the same thing. However as we lack such a theory (which would probably entail a full understanding of human intelligence) we can't assume these two approaches end up telling us quite the same thing.

## Examples of Approach A and B:

Consider the following two items in terms of the two approaches to maths tests:

### Approach A or Approach B?

Which one of these shapes will tessellate?



(A)



(B)



(C)



(D)

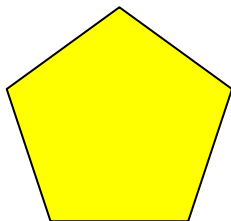
[Adapted from Australasian Schools Mathematics Assessment 03Y7Q34]

Judging by Approach A the item has many faults. It appears to be testing familiarity with tessellation but the problem is very difficult. It could be expected that many students who know what tessellation means would get the item incorrect. The multiple choice format includes a degree of guessing.

Judging by Approach B the item is better. Students more able, in general, at mathematics are more likely to get the item correct. A similar item in the 2003 Australasian Schools Mathematics Assessment had a percentage correct of 55%. The top third of students has a percentage correct of 71%.

### Approach A or Approach B?

What is the name of this shape?



Write your answer here: \_\_\_\_\_

[Item writing training item]

Judging by Approach A it is quite clear what this item aims to test. There are issues as to whether to accept spelling mistakes of 'pentagon' but few other issues.

Judging by Approach B the item has a number of problems. Either you know the answer or you don't. If you have never met the word "pentagon" there are no other mathematical skills you can use to find the answer. The item is a test of vocabulary rather than mathematical ability.

## Multiple Choice

Multiple choice items are a sensible choice for tests of general ability. They are easy to administer and mark but they also provide kinds of information that can be hard to extract from free-response items.

### Structure of a Multiple Choice Item

The strength of a multiple choice item lies in the options. The options are the responses a student can choose from. The correct option is called the 'key'; the other options are called 'distractors' or 'foils'. The options can affect nearly all the properties of an item; its difficulty, its discrimination and even what the item is testing.

It is not uncommon to see in some tests multiple choice items with little thought behind the distractors. However the distractors provide valuable information on students reasoning only if they have been well written.

### Distractor Reasoning Example 1

What is the reasoning for each option in this item?  
Which option would a Year 9 student pick?

This formula converts a temperature from degrees Fahrenheit ( $F$ ) to degrees Celsius ( $C$ ).

$$C = \frac{5}{9} (F - 32)$$

Which of these formulas converts Celsius to Fahrenheit?

- (A)  $F = \frac{9}{5} C + 32$
- (B)  $F = \frac{5}{9} C + 32$
- (C)  $F = \frac{5}{9} (C + 32)$
- (D)  $F = \frac{9}{5} (C + 32)$

[Australasian Schools Mathematics Assessment 04 Year 9 Q13]

Each option serves a purpose, primarily to trap errors. In this case only 15% of Year 9 students chose the key (A). The strongest option was (D) which is correct apart from the superfluous brackets. The least popular option was (B). It is worth considering what that tells us about a student's skills and reasoning in algebra. Note that the percentage correct for this item is less than the 25% you would expect if the students just guessed. Multiple Choice items are not necessarily "Multiple Guess" items. Well constructed items in a well constructed test will minimise guessing. A multiple choice test with several items such that the least able students stand a better chance of getting the item correct by attempting it than they would by guessing will effectively penalise guessing. This doesn't require complicated scoring systems (such as penalising wrong answers) just a good spread of well written items.

## Distractor Reasoning Example 2

What would be the best options for this number puzzle?  
A zookeeper had some peanuts for the monkeys.



He gave each monkey 4 peanuts and had 2 peanuts left over.  
To give each monkey 6 peanuts, he would need another 22 peanuts.  
How many monkeys are in the zoo?

[Australasian Schools Mathematics Assessment 04 Year 6 QF3]

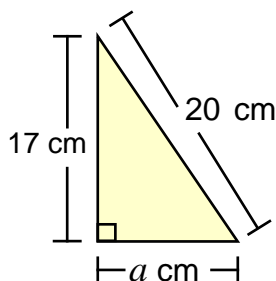
The top 6 responses for this item were:

12	26.34%		5	10.93%
11	22.10%		6	10.10%
4	21.96%		3	8.57%

As an exercise try and work out the reasoning behind each response.

## Distractors Affect the Tested Content of an Item

What content is tested in this item? Are the distractors suitable? What would be a better set of distractors?



What is the value of  $a$  to the nearest whole number?

- (A) 10      (B) 10.5  
(C) 10.6    (D) 11

[Item writing training item]

This is an artificial example to demonstrate the role of distractors.

The correct answer can be found using Pythagoras ( $20^2 - 17^2 = 111$ ,  $\sqrt{111} = 10.535653752\dots$ )

Each distractor is based on incorrectly rounding the square root of a hundred and eleven. Consequently the apparent content of the item (Pythagoras) is barely tested and the main focus of the item is rounding and knowing what 'to the nearest whole number' means. Note that a student who only knows what 'to the nearest whole number' means can eliminate two of the options straight away.

## Distractor Reasoning Highlights Faults in an Item

What is the reasoning behind each of these, numerically identical, options?  
How should the item be rewritten?

What is the value of  $2^2$ ?

- (A) 4
- (B) 4
- (C) 4

[Item writing training item]

A possible set of distractors for testing whether students are familiar with the square notation would be to give options based on possible interpretations. In the example these are:

- $2+2=4$
- $2\times 2=4$
- 2 squared =4 (the key)

Clearly nobody would pick  $2^2$  as the example to use to test that skill! However less trivial items can suffer from incorrect but plausible methods that lead to the correct answer. This is not always obvious until you consider what plausible strategies a student might take to answer the item.

Even on free response items it is worth considering what would be the most appropriate set of distractors.

## Distractors Affect the Difficulty of the Item

About 60% of Year 8 students could answer this question in a maths competition.

What set of options would have made the item harder? How could we change only the options of this item to test a related but different skill?

The picture shows the angle between two sections of a roof.



Which of these is the best estimate of the value of  $a$ ?

- (A) 90
- (B) 100
- (C) 120
- (D) 140

[Australasian Schools Mathematics Assessment 05 Y8 Q23]

The difficulty of estimation questions clearly depends to some extent on the options.

This set of options would have been much more challenging as they demand greater precision:

- (A) 130
- (B) 135
- (C) 140
- (D) 145

This set of options changes the item from estimation to recognition of an obtuse angle:

- (A) 40
- (B) 90
- (C) 140
- (D) 240

For anybody who is interested the picture is of Trial Bay Gaol near Arrakoon NSW.

## Developing Better Practice

Developing good test items helps clarify your own thinking about the curriculum and assessment. Taking care to write well structured items will improve the quality of the data you obtain from tests. The following 4 points sum up the major ways the quality of test items you use can be maintained or improved:

1. Be clear about what kind of data you want from your tests
2. Choose and write items suitable for the kind of test you are writing
3. Maintain a bank of items you have developed
4. Look at how items have performed over time and use that data to refine the items you have.

Finally remember that if you aren't enjoying writing an item you can be quite sure your students won't be enjoying answering it.