Science or Reading?:

How students think when answering TIMSS questions

Alastair Pollitt & Ayesha Ahmed

Research & Evaluation Division

University of Cambridge Local Examinations Syndicate

A paper presented to the International Association for Educational Assessment,

in Rio de Janeiro, Brazil, May 2001.

Contact details:

Alastair Pollitt & Ayesha Ahmed, RED, UCLES, 1 Hills Road, Cambridge, CB1 2EU

pollitt.a@ucles.org.uk tel: +44 1223 553847, fax: +44 1223 552700

ahmed.a@ucles.org.uk tel: +44 1223 553836, fax: +44 1223 552700

Science or Reading?:

How students think when answering TIMSS questions

Alastair Pollitt & Ayesha Ahmed

Research & Evaluation Division

University of Cambridge Local Examinations Syndicate

A paper presented to the International Association for Educational Assessment,

in Rio de Janeiro, Brazil, May 2001.

Abstract

Several recent publications have commented on detailed differences in performance on particular questions from the Third International Mathematics and Science Study. These comments often take for granted the validity of the questions being analysed. Conclusions about countries' average scores cannot be trusted unless all the questions in a test really do measure 'science' or 'maths'. Careful study of some of the questions suggests that they may tell us more about how groups of students read than about their Maths or Science ability.

We have applied our model of the question answering process (Pollitt & Ahmed, IAEA 1999) to some of the TIMSS items to see if many questions suffer potential invalidities. If there is evidence that the students' minds were not doing what the question writer intended them to do then we can conclude that the question was not valid.

Strategies for anticipating validity problems can be developed; these will help to strengthen the utility of future surveys.

Validity

From a cognitive psychologist's point of view, construct validity must concern what is happening inside students' heads while they are trying to answer the task. When we ask them to apply certain skills to certain subject content we hope that their minds will really be doing whatever these particular skills involve and really thinking about the subject content we intended. If they are, then they will be working on the construct we intended to measure; if their minds are doing anything else, then there is a clear risk of invalidity.

This can be summed up in our psychological definition of construct validity:

A question is valid if and only if the students' minds are doing the things we want them to show us they can do.

In order to discover whether a particular question is valid we need to try to understand the processes that will be occurring in students' minds when they are answering that question. If these are the processes we want to measure then the question is likely to be valid. We have developed a general model of the question answering process (Pollitt & Ahmed, 1999) which can be applied to specific questions to give us an idea of what may be occurring in students' minds when they are answering them. This can be extended to allow us to predict the sorts of right and wrong answers students may give to a question.

The model of question answering

The model of the question answering process consists of six phases from learning the subject through to writing the answer. It can be used to form an understanding of the psychological processes that are occurring when a student is answering a particular exam question, and so to anticipate the sorts of answers students will give to a question.

0 Learning the subject
1 Reading the question
2 Searching the memory
3 Matching question to memory
4 Generating an answer
5 Writing the answer

In Phase 0 students learn the subject before the exam; it is this learning that we are trying to measure. Then in Phase 1 the student reads the question forming a mental representation of the task. In Phases 2, 3 and 4 students search their mental representation of the whole subject, a rather fuzzily defined subset of their memory, looking for relevant concepts to match those in their representation of the question, and use these to generate an answer. In Phase 5 the student's mental representation of the answer is turned, in most cases, into a string of words.

The order of the phases is, to some extent, logically necessary. However, Phases 2-4 are likely to occur rapidly, automatically and pre-consciously. In some questions these phases will occur simultaneously or there will be repeated cycles of Searching, Matching and Generating.

The model is a generalisation and will be different for different question types. In several familiar question types, such as cloze or multiple choice, there may be no discrete phase of generating the answer. In cloze students often find it impossible to say, by introspection, how they arrived at their answer: it 'just popped into' their heads (Taylor, 1991). Faced with a multiple choice answer set, students will very often identify an answer during the Searching and Matching phases and the Generating phase will not occur; the Writing phase, in this case, is simply ticking a box. The Generating phase will vary for other question types: for mathematical problems or other problem solving tasks there will be an explicit phase of generating an answer, while for essay answers the Generating phase is intimately bound up with the Writing phase.

Because the model can be applied so generally it is an effective tool for understanding the processes occurring in students' minds as they are answering particular questions or tasks. This in turn allows us to make some conclusions about the validity of questions, based on the processes they are causing in students' minds. This understanding of what makes a question valid can enable us to improve the validity of questions and marking schemes, giving us more control over the assessment process.

A tool developed from the model of question answering is what we call the Outcome Space Generator, and is intended to help us to improve questions and marking schemes. At the same time it will enable us to extend our model to include multiple choice questions, both to anticipate likely wrong answers and to select appropriate distractors during question setting.

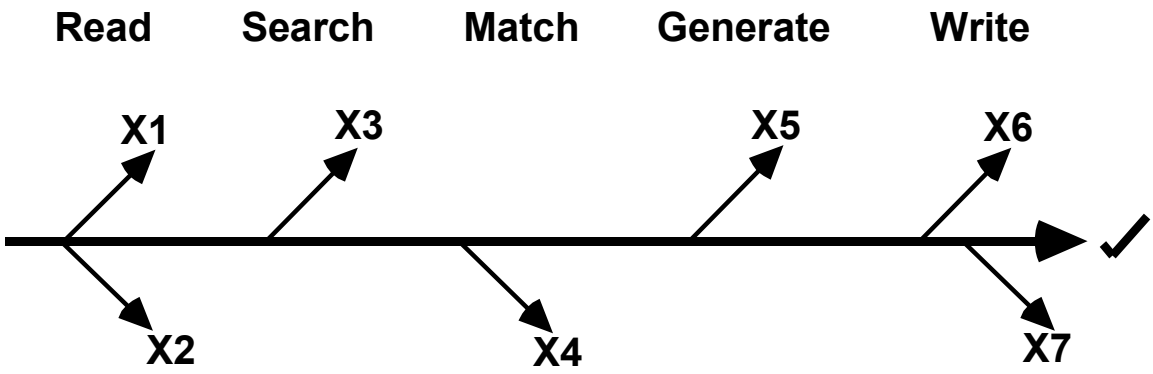
The notion of outcome space was introduced by Marton and Saljo (1976) in their research on differences in learning. They asked students to read passages of text within suggested time limits and then to answer some specific questions and to explain what the passage was about. They found that the responses varied greatly but that each student had comprehended parts of the text in one of only a few fundamentally different ways. Marton and Saljo referred to these types of answers as the 'levels of outcome' that make up the 'outcome space' for a particular question. Hence, outcome space can be thought of as the range of qualitatively different types of answer to a question or task.

The Outcome Space Generator

The Outcome Space Generator (OSG) is a systematic procedure for question writers to anticipate possible types of answer, right and wrong, to their questions. It was developed for use with free response and structured questions as a device for generating marking schemes by allowing examiners to evaluate all of the range of responses they are likely to see when the questions are used. In questions of these types responses may vary in *nature*, as well as in how *adequate* they are, and these two dimensions may vary quite independently. We have found that it is possible to extend the usefulness of the OSG to multiple choice questions, to try to understand why students made particular choices in questions such as those used in TIMSS. Because we have little evidence of the students' thought processes from looking at their answers which are simply a choice of one of four or five options, we need a way of systematically analysing the questions to form theories of why certain answers were given. This in turn allows us to re-examine the question and consider its validity.

The Outcome Space Generator can also be used at the question creation stage to decide on suitable distractors for a multiple choice question. By following in detail the likely thought processes involved in forming a mental representation of the task, including understanding the words in the question, the writer can predict errors. Some will be 'valid', when the student shows misunderstanding of the concepts supposedly learned in Phase 0, but others will be 'invalid' and related more to misunderstanding the question than the subject matter. The former set of errors suggest valid distractors, the latter need to be put right.

Applying the Outcome Space Generator



In this diagram the X's represent the kinds of wrong answer that might arise. Starting with Phase 1 of the model, Reading, the question writer can think through the processes that occur in students' minds while they are reading a question. Certain concepts are activated in their minds, some relevant and some irrelevant. A correct reading of any sentence activates many concepts, not just the ones intended by the writer. For example it has been shown experimentally that hearing a sentence about paying in a cheque will activate all sorts of financial concepts including 'bank' which will in turn, perhaps surprisingly, activate the concept of 'river'. If an irrelevant concept is activated and a student incorrectly identifies it as a relevant one this could result in wrong answers. If question writers can anticipate the irrelevant concepts that may be activated in students' minds when reading their question, they may then be able to predict the sorts of wrong answers these concepts could lead to.

It is important to remember that it is not just reading errors that will provoke irrelevant concepts. As each individual student reads a question their different minds will generate idiosyncratic interpretations consisting of the activation of a slightly different set of concepts in each case.

Both relevant and irrelevant concepts are activated automatically; the student will not be aware of them since they occur *before* the reading of the text reaches consciousness. Thus these concepts, although selected unconsciously, will affect the subsequent interpretation of the text. Evans (1989) calls this 'pre-attentive bias'. In most cases irrelevant concepts are

activated only weakly or are ignored as attention is focused on the relevant ones. However, when students are reading exam questions they are under stress and are therefore more easily distracted by irrelevant ideas. Attentional resources are finite, and under stress some attentional capacity is used in coping with aspects of stress such as anxiety and the need to monitor time. There is therefore less attentional capacity available to focus on the task, making misreadings and misinterpretations more likely (Kiwan, Ahmed & Pollitt, 2000).

When a question is set in a real-world context, as a few of the Science questions in TIMSS are, the ways in which students will understand the question are even more varied. There is usually more text in contextualised questions, some of which is not relevant to the task, so misinterpretations are more likely. The level of familiarity of the student with the context is also a factor in each student's understanding of the task. The issue of producing valid contextualised questions is too complex to address here but is discussed in Ahmed & Pollitt (2000).

Most errors are likely to originate in misinterpretations during Reading. After Phase 1 many concepts have been activated in the students' minds, forming a mental representation of the task. In Phases 2, 3 and 4, Searching, Matching and Generating, students search their mental representation of the subject knowledge, actively but not consciously, looking for matches to the mental representation of the task. From these matches an answer is generated. Any of these processes can go wrong and will result in various types of wrong answer. Just as in Reading, these phases will involve activating still more concepts with still more scope for going wrong. Errors in Phase 5 may arise because students often struggle to express their understanding accurately in words. Summarising a mental representation as a linear text has been shown to be a very difficult task for students under the age of 18 (Brown and Day, 1983) and their findings seem to indicate that it is a developmental ability.

When analysing multiple choice questions the OSG can be used to identify likely reasons why students chose a particular distractor instead of the right answer. We have applied the OSG to some of the multiple choice items used in TIMSS 1995 Science for Population 2 (grades 7 and 8).

Examples from TIMSS 1995 Science

We will consider three questions and analyse the response processes in each to try to understand, or predict, the most likely responses. From this, considerations of validity will arise. First, we will analyse one question is considerable detail.

Question K16

Which is made with the help of bacteria?

A. Yogurt

B. Cream

C. Soap

D. Cooking oil

Consider how the typical English-speaking student will process this question. As they read each word in the sentence many different concepts are activated in their minds, and these in turn activate associated concepts in a gradually diminishing cascade of activation. A student trying to answer this question must somehow select out of all of this activation a pattern of activation whose meaning satisfies their need to answer the question.

Let us look in more detail at this process. 'Words are cues to build a familiar mental model' (Johnson-Laird, 1981). As students read the question they are building a mental model of the task out of the concepts provoked by the words. Each word will contribute some concepts to the overall network with varying degrees of activation, and out of this pattern of activation they construct their mental model of the task. In this case the task is simply to choose one of four options.

The first few words provoke their default meanings most strongly, influenced only by the overall context of taking a science test; this means that **Which** will lead the student to expect a question (since questions will be a highly salient concept in every student's *test* schema). Other interpretations, such as the relative clause construction and even the consumer magazine *Which?* will also be made active, but less strongly. As **is** and **made** are processed the initial default interpretation may change; for this question the relative clause interpretation looks more promising at this stage. **made** provokes many concepts: it may be a de-lexicalised verb (as in 'made more important') or it may mean some kind of fabrication/construction etc. **with** will modify the interpretation of **made**, suppressing the de-lexicalised senses and reinforcing the fabrication meaning. **made with** will activate meanings like 'made from' or 'made out of'

the help will provoke many thought sequences most of which will rapidly be suppressed, but some associations will persist, probably involving people intentionally supporting an activity.

of bacteria then becomes part of the phrase 'with the help of bacteria'. All of the concepts associated with 'bacteria' are activated and compete to be built into the mental model that is being assembled; these will include – with high salience – those concerning disease and germs. The concept of 'help' is rather incongruous in this context, and some dissonance will result. Confident students will handle this easily, being familiar with the idea that bacteria are sometimes "helpful"; less confident ones, who are less sure of their knowledge, may look for 'help' elsewhere.

? will at least confirm that the student has been asked a question, important confirmation for the mental task model they have constructed.

It is very unlikely, though, that they will try to answer it before reading the options. The question began with **which**, and they will therefore be expecting some possibilities to choose from. In addition, although some students may have a prototypical example of "helpful bacteria", perhaps because it was in their textbook, most will not have an immediate answer available.

Students, then, will read the four options: **yogurt**, **cream**, **soap**, **cooking oil**. Many trains of thought will be started, and the ones that persist most will be the ones that follow lines already activated by the reading of the question. For many students, of course, the task will end with **yogurt** as they "know" that yogurt is the correct answer (though they will almost

certainly continue to process the other options just to make sure). For all students, the concepts and associations activated by 'yogurt' might include nice, dairies, healthy, fitness, fruit etc. **cream** will activate a similar set of concepts, including perhaps strawberries, good, fattening, bad, smooth, goes off, ice-cream, cows, cats, off-white or meringues. **soap** might suggest cleaning, smelly, dirt, hygiene, germs, clean behind your ears, slippery, bath, or television. **cooking oil** will be more complex, since it is compound. The phrase 'cooking oil' might activate chips, fry up, olives and sunflower, fat, diet, grease, etc; 'cooking' will provoke frying, grilling, mother, heat, killing bacteria, kitchen, television chefs and, of course, food; 'oil' will activate concepts like water, petrol, greasy, sticky, chemistry, yucky, brown, wealth, oil wells.

And all of these will, in turn, activate still more. If it has not already "known" the answer the student's mind will now seek a pattern that makes sense, linking one of the options to the question. Brains are good at pattern matching (Hofstadter et al, 1995) and many connections will be recognised; a few of them may reach consciousness and one is going to be chosen. But which one?

Consider the alternatives, reduced to content words only:

A .. made .. help .. bacteria .. yogurt

B .. made .. help .. bacteria .. cream

C .. made .. help .. bacteria .. soap

D .. made .. help .. bacteria .. cooking oil

Those who don't "know" the answer must construct a story around these words – it is well known in linguistics that readings of a text are dominated by the heavily lexical content words rather than the little grammatical ones. The most coherent story of the four is probably C, since 'soap' is used (made) in order to help us resist germs (bacteria). The three concepts 'soap', 'bacteria' and 'help' are closely associated in the overarching schema of hygiene, familiar to every 13 year old. We would therefore predict that, of the three wrong answers, 'soap' will be the most popular.

In fact, the overall percentages were as follows:

A	B	C	D
33.1%	13.4%	29.9%	19.7%

'Soap' was indeed the most popular, almost as popular as 'yogurt'. We might infer that many students knew that yogurt is made by the action of bacteria on milk, but that a large proportion of those who did not were tempted into choosing soap by the associations activated.

There were curious differences in the relative popularity of the three wrong answers in different countries. In most, as shown above, 'soap' was the most popular. But in two, Germany and Hong Kong, 'cream' came first; 26.5% of German students chose 'cream' when 29.6% chose 'yogurt'. In eight countries the most popular wrong answer was 'Cooking oil': Australia, Belgium (Flemish), Canada, England, Ireland, Scotland, South Africa and the United States. It is remarkable that this list contains all the English language countries except New Zealand.

We have no very convincing explanation for these inter-national patterns. Perhaps there is an active association between 'cooking' to help kill off 'bacteria' which is somehow more salient in English speaking countries, or perhaps they think of bacteria's role in making petroleum oil? And we have no idea at all to explain the Germans' difficulties – perhaps they just didn't know how yogurt is made? If the question was intended just to assess knowledge, then the German confusion may not be valid.

Question N2

In this second example we will pass more rapidly over much of the detail of mental model building that students are likely to undergo. We will consider the interactions between the question stem, the options, the science learning and the background cultural knowledge of the students.

Which of these meals would give you most of the nutrients that you need?

A. Meat, milk, and a piece of chocolate

B. Bread, vegetables, and fish

C. Vegetables, fruit, and water

D. Meat, fish, and bread

Reading

As the student starts to read, **'Which of these meals'** sets up in their mind expectations that the rest of the task will concern a set of choices from a menu such as you might make when a waiter asks if you are ready to order your meal.

The phrase **'most of the nutrients'** is ambiguous. It could mean all of most of the nutrients that you need (i.e. counting nutrients) or most of all of the nutrients that you need (i.e. measuring quantity). It is most unlikely that a student will choose one or other of these alternatives; rather both will be carried forward in a rather fuzzy representation of the task, in the subconscious hope that reading the options will help to resolve the ambiguity.

To add to the fuzziness, the word **'nutrients'** can be defined differently in different contexts or countries. In England there is a debate as to whether there should be five or seven nutrients in a healthy human diet: everyone agrees on the three 'food groups' (protein, carbohydrate and fat), together with vitamins and minerals, but some would like to see fibre and water included as well. It is also questionable in this context whether animal and vegetable protein should be considered separately or together. The word 'nutrient' is also used quite differently in the context of 'plant nutrients' to mean nitrogen, phosphorus and potassium, the key elements of fertilizers.

When the student reaches the options, they do not really sound like 'meals' but more like lists of food *types*. On this basis a balanced meal is not what students will focus on; instead the most salient concepts in their minds are likely to concern healthy types of food.

Searching and Matching

Everyone has their own unique knowledge, based on their own learning and experiences, of food and health. In addition, it is an area strongly influenced by *affect*, since everyone knows what they like and dislike about food, and even by *morals*, since most students are well aware of the debates surrounding animal rights and vegetarianism. It's even possible that some vegetarian students, having read the question as literally about giving **'you'** most of the nutrients that **'you'** need, might feel that A, D and even B are unacceptable answers. The – mostly unconscious -processes of searching the mind for the most relevant concepts amongst all those activated and matching these with one of the options will be constrained by students' everyday knowledge about food and health as well as by what is given in the question.

The overall percentages choosing each option were as follows:

A	**B**	C	D
9.5%	**37.4%**	38.5%	12.9%

The correct answer was B, 'Bread, vegetables, and fish', but slightly more students overall chose option C, 'Vegetables, fruit, and water'.

B seems the most balanced of the four options, and certainly can be seen to include all five or seven 'nutrients'. However the question asks about nutrients that you *need*, and not about a balanced diet, which may have misled students who were focusing on a 'healthy' diet rather than a 'balanced' one. Many students think of 'fruit and vegetables' as being essential for a 'healthy' diet, and many will have been told how important it is to eat more fruit and vegetables. Moreover C could include all seven nutrients, depending on the particular vegetables and fruit chosen; olives, for instance, would provide plenty of vegetable fat. B certainly includes both animal and vegetable protein, but vegetarians might reasonably argue that no one actually 'needs' animal protein.

We summarised the previous section with the statement that "the most salient concepts in their minds are likely to concern healthy types of food"; the popularity of C supports this view.

Cultural effects

We would expect there to be strong cultural effects here if students from different countries are taught differently about healthy food. Advertising or government campaigns to tell people to eat more healthy food probably differ from one country to another, with some emphasising that fish is healthy while others emphasise fruit and vegetables. It is likely that the students answering this question may not be focusing on the science, that is the nutrients contained in these foods, but instead on their everyday knowledge of what is portrayed to them as healthy. Is there support for this in the data? The table below shows the ratio of the number of students choosing the right answer B to the number choosing C. We might characterise this ratio as bread and fish versus fruit and water.

Ratio of B to C choices

Norway	4.58	Cyprus	1.32	Switzerland	0.64
Singapore	3.79	England	1.23	Philippines	0.64
Thailand	3.66	Netherlands	1.11	Hungary	0.59
Canada	2.40	**AVERAGE**	**0.97**	New Zealand	0.59
Korea	1.96	Greece	0.95	Spain	0.57
BelgiumFl	1.83	Bulgaria	0.92	Slovak Rep.	0.55
Hong Kong	1.82	Germany	0.92	Romania	0.53
Iceland	1.73	Japan	0.83	Latvia	0.50
United States	1.71	Slovenia	0.79	Scotland	0.47
Iran	1.55	Austria	0.75	Russian Fed.	0.44
Czech Rep.	1.50	France	0.74	Lithuania	0.42
Australia	1.47	Ireland	0.74	South Africa	0.37
Sweden	1.44	BelgiumFr	0.71	Colombia	0.20
Denmark	1.42	Portugal	0.69		

Overall, Norway, Thailand and Canada performed quite averagely in this TIMSS survey. Yet in Norway, students were four and a half times as likely to choose B as C; does this reflect a bias for fish over fruit in their national consciousness of healthy eating? In contrast, students in Colombia were five times as likely to select C as B. Such a wide disparity, not all explainable in terms of overall performance level, does seem to indicate that substantial cultural influences were active.

When the data show evidence that very different thinking processes are the norm in different countries we are led to ask if it was wise to include this question in a comparative study of achievement.

Question N1

In this next question we shall address an issue of validity more directly.

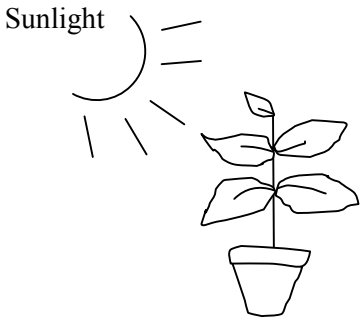
Reading

The question starts by telling students:

> *A girl had an idea that plants need minerals from the soil for healthy growth. She placed a plant in the Sun, as shown in the diagram below.*

The form of the English language text causes students to focus on the word 'Sun' rather than the intended focus 'minerals'. 'Sun' is capitalised and so becomes the dominant concept in the second sentence. It is far more salient than the subordinate clause 'plants need minerals from the soil'. Some students will misread or misinterpret the task thinking that the girl is looking at the effects of sunlight on the plant.

The second sentence also directs students' attention to the diagram of a plant in Sunlight. Thus, just as the 'Sun' is made salient in the text it is immediately reinforced in the diagram.



Sand, minerals and water

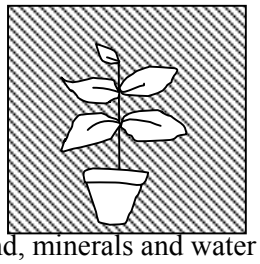
The word ‘minerals’ is buried in the middle of the phrase ‘Sand, minerals and water’ and there is no drawing of minerals to make it salient as there is of the sun.

The question continues:

> *In order to check her idea she also needed to use another plant. Which of the following should she use?*

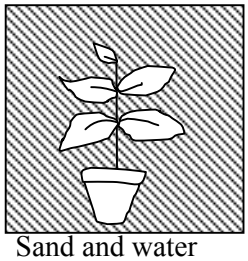
Note that there is an inconsistency here in the use of verb tenses which may have disturbed some students. The options are a series of five pictures, two showing the plant in a dark cupboard instead of Sunlight and three that differ only in the phrase under the drawing:

A. Dark cupboard



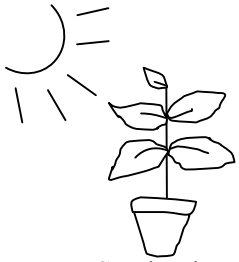
Sand, minerals and water

B. Dark cupboard



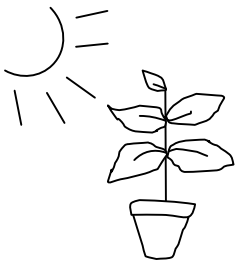
Sand and water

C. Sunlight



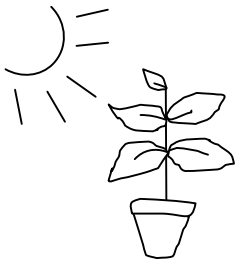
Sand only

D. Sunlight



Sand and water

E. Sunlight



Sand and minerals

Searching and Matching

The question turns on the phrase ‘to check her idea’; as a science question the focus is meant to be on knowing how to ‘check’ an idea, not on what her idea is. They need to know her idea was about minerals not about sunlight before they can be validly assessed

by this question. Because 'Sunlight' is far more salient than 'minerals', many of them may think that her idea is to check the effect of sunlight and therefore that she should now place the plant in the dark, keeping the other factors constant. This will lead these students to choose the incorrect answer 'A'. They are searching their minds for what to do to check the effects of sunlight on the plant, and find a match with the diagrams that show it in a dark cupboard. Notice that students selecting A have shown that they know how to make a scientific check, since they changed only sunlight and kept everything else constant. Unfortunately they got the wrong 'idea'.

These students have formed a clear understanding of the task, but not quite the right one. Other students will form a less precise understanding, by focusing on 'her idea' not as placing the plant in the sun but more generally as 'doing a controlled experiment'. Their minds then activate memories of experiments they have done in school or read about in textbooks, which may lead to one of the option drawings seeming a good match with what is in their minds; these students may choose as an answer any condition in which just one factor is changed. Thus 'A', 'D' or 'E' could be chosen by students who still understand the scientific principle being assessed.

Finally, particularly for students who have little idea of what the question is about, there is high visual salience to A and B, the picture of the plant in a dark cupboard. Even if they have not thought that the key issue is 'Sunlight' from reading the question, the strong visual contrast of the shading with the 'Sunlight' is likely to lead them to this interpretation when they look at the options. All of these factors mean that A is a distractor many students are likely to choose.

These sorts of misunderstandings may seem unlikely, but under the conditions of cognitive stress in an assessment context they really do happen (Ahmed & Pollitt, 2000).

For the whole international sample the results were as follows:

A	B	C	**D**	E
25.9%	7.5%	6.9%	**40.6%**	14.7%

As expected A was the most popular wrong answer, almost as popular as B, C and E combined. B and C, the two answers that show a failure of scientific understanding, were the least popular.

If we look at the relative popularity of A compared to the other wrong answers for each country some intriguing patterns emerge:

Choice of A, as a percentage of all wrong answers

Norway	0.17
Sweden	0.26
Lithuania	0.27
Greece	0.32
France	0.32
South Africa	0.37
Russian Fed.	0.38
Romania	0.38
Switzerland	0.38
Hungary	0.40
Bulgaria	0.40
Portugal	0.41
BelgiumFr	0.42
Germany	0.42

Spain	0.45
Cyprus	0.45
Slovak Rep.	0.46
AVERAGE	**0.47**
Australia	0.47
Austria	0.48
Slovenia	0.49
Netherlands	0.49
Iran	0.49
Czech Rep.	0.50
Latvia	0.51
Hong Kong	0.51
Iceland	0.52
United States	0.52

BelgiumFl	0.52
New Zealand	0.52
Canada	0.53
Philippines	0.53
Denmark	0.54
Ireland	0.55
Scotland	0.57
Korea	0.58
England	0.61
Colombia	0.63
Singapore	0.64
Thailand	0.66
Japan	0.71

Most of the countries which scored high on the test overall were more attracted to A than the other wrong answers.

In Germany, where most nouns will be capitalised as a matter of course, A was less popular than usual.

All of the English-speaking countries were more likely than average to pick A if they did not know the right answer (our analysis has only used the American English version of the questions and so we cannot make much comment on possible effects of other language versions).

Students in Norway and Sweden were surprisingly reluctant to choose A, and mostly chose E instead. In Norway, in fact, A was the *least* popular of all the five choices; in every other country A was more popular than B and C. We have no explanation for this.

It seems that many students were getting the wrong answer because they focused on the wrong 'idea' although they knew the principle of experimental control. To the extent that this is true the question is invalid as a test of scientific understanding, and functions more as a test of reading comprehension. This problem with the question was predictable on psychological grounds, as discussed above; by generating a predicted outcome space in advance it might have been avoided. We would suggest that options A and E – scientifically defensible but prompted by linguistic misunderstanding – might have been replaced by others which more clearly violated the scientific principle of experimental control in the way that B and C do.

Conclusions

Before this exercise we had only applied the Outcome Space Generator to free response questions. It can, however, also be applied to the three phases of Reading, Searching and Matching in multiple choice questions as we have shown.

Our Outcome Space Generator can be used as a systematic way of producing distractors for multiple choice questions. There are various well known books giving advice on how to write test questions, and they seem to agree on two principal ways to generate distractors. Where the emphasis is on diagnostic use of the tests, options are offered that reflect the results of particular anticipated errors or misunderstandings. In our terms the test constructors are interested *only* in errors that have already occurred in Phase 0 (Learning), before the student starts the test, and it is therefore of supreme importance for the dependability of their interpretations that there are *no* errors occurring in Phases 1-4. Identifying errors that might have occurred in Phase 0, and using them as options, may change the way in which the question is read and the task understood, inadvertently leading to errors in Phases 1-4 from students who otherwise would have answered correctly. The OSG is a systematic way of identifying as many as possible of the errors that might be made in those phases, so that questions can be focused more clearly on assessing Phase 0.

The other recommended approach is empirical. Questions are offered to a sample of students without options, in free response format, and the three or four most popular wrong answers are chosen as distractors. This approach is designed with a particular intention – as many as possible of the students should find their first response present in the list of options, so that as few as possible are given an unfair second go by being 'told' to think again. The OSG attempts to simulate such a pretest through a better understanding of how students will think when answering the questions. Although it was intended for improving the wording and presentation of the questions themselves, we have found that it can also be used to suggest the most likely kinds of error. This is obviously of great value when pretesting is not possible for pragmatic reasons or when, as with two of our three examples in this paper, free response questioning is simply impossible in principle.

Our examples have concentrated on linguistic aspects of test questions, but it is important to remember that there are other significant features that can threaten the validity of questions. Reading comprehension is only part of the process of forming a representation of a task, and the representation may be affected by several other factors to do with question design, layout and marking. Further details can be found in Ahmed & Pollitt (1999).

Although our examples show significant differences between nations' performances that seem linked to culture and language, it is not our intention here to highlight these. Issues of translation and of cultural differences are familiar to the TIMSS participants. We have concentrated on a description of how students' minds process written questions in general, on what is common to all students rather than what is different between distinguishable groups. In TIMSS 1995, Population 2 Science, we found very few questions that gave us significant concern about validity in this respect.

Generating an outcome space in this systematic way would normally be a process occurring before, not after, questions are used. It involves imagining all of the possible ways that a student might misinterpret a question and all of the concepts and associations that are likely to be activated even when it is properly understood. Much of the work is linguistic, looking for ambiguity, polysemy or syntactic complexity, much of it is psychological, anticipating the sorts of associations that may be provoked and which are

likely to persist, and much is subject based, using knowledge of the kinds of confusions and errors that students commonly make in the topic area. Expertise in each of these domains should be involved in the scrutiny of questions before they are used.

Marton and Saljo's (1976) conceptualisation of Outcome Space is a powerful way of analysing the validity and invalidity of test questions. Rather than merely considering answers on one dimension – representing how adequate they are – it forces us to consider them on at least one extra dimension – representing qualitative differences between kinds of responses. These can only be anticipated through the application of a systematic procedure for predicting students' thought processes, based on sound theory and extensive empirical investigation. This is what the Outcome Space Generator allows us to do: following the various paths that students' minds are likely to trace we can predict their conclusions, confusions and blind alleys. In a multiple choice context we can use this to create distractors or predict how students will select from them. In any assessment context the discipline of generating the outcome space systematically will help to catch faults in the questions before they can disturb thinking patterns and cause invalidity.

Unless this level of scrutiny is part of the question setting process, we are liable to lose control of the students' mental processes, and so to undermine the validity of the assessment. Without valid questions, interpretations of national results will never be valid.

References

Ahmed & Pollitt (1999) Curriculum Demands and Question Difficulty. IAEA, Bled, Slovenia, May 1999.

Ahmed, A & Pollitt, A, (2000) Comprehension failures in educational assessment. European Conference on Educational Research, Edinburgh University, September 2000.

Brown, A.L. & Day, J.D. (1983) Macrorules for summarizing texts: the development of expertise. *Journal of Verbal Learning and Verbal Behaviour*, 22, 1-14.

Evans, J. St. B.T. (1989) *Bias in Human Reasoning: Causes and Consequences*. Hove: Lawrence Erlbaum.

Hofstadter, D, and the Fluid Analogies Research Group (1995) *Fluid Concepts and Creative Analogies*. HarperCollins.

Johnson-Laird, P.N. (1981) Mental Models of Meaning. In Joshi, A.K., Webber, B.L. & Sag, I.A. (Eds.) *Elements of Discourse Understanding*. Cambridge: Cambridge University Press.

Kiwan, D. Ahmed, A. & Pollitt, A. (2000) The Effects of Time-Induced Stress on Making Inferences in Text Comprehension. European Conference on Educational Research, Edinburgh, September 2000.

Marton, F. and Saljo, R. (1976) On qualitative differences in learning: 1 – Outcome and process. *British Journal of Educational Psychology*, 46, (p4-11).

Pollitt, A. & Ahmed, A. (1999) A New Model of the Question Answering Process. IAEA, Bled, Slovenia, May 1999.

Taylor, L. B. (1991) *Some Aspects of the Comparability Problem for Communicative Proficiency Tests.* Unpublished MPhil Dissertation, University of Cambridge.