Quantitative research methods in educational planning

Series editor: Kenneth N.Ross

Module 5

Graeme Withers

Item writing for tests and examinations

UNESCO International Institute for Educational Planning

Quantitative research methods in educational planning

These modules were prepared by IIEP staff and consultants to be used in training workshops presented for the National Research Coordinators who are responsible for the educational policy research programme conducted by the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ).

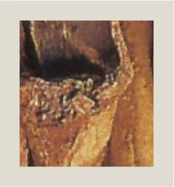
The publication is available from the following two Internet Websites: **http://www.sacmeq.org** and **http://www.unesco.org/iiep**.

International Institute for Educational Planning/UNESCO
7-9 rue Eugène-Delacroix, 75116 Paris, France
Tel: (33 1) 45 03 77 00
Fax: (33 1) 40 72 83 66
e-mail: information@iiep.unesco.org
IIEP web site: http://www.unesco.org/iiep

The designations employed and the presentation of material throughout the publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning its frontiers or boundaries.

Graphic design: Sabine Lebeau
Typesetting: Sabine Lebeau
Printed in IIEP's printshop

Content



Item writing – art or science? 1

Within the field of test development, the tasks and/or questions that are used to construct tests and examinations are referred to as ‘items’, and the range techniques involved in preparing those items are collectively referred to as ‘item writing’.

Is item writing an art or a science? The best item development techniques combine elements of both these intellectual activities. On the one hand, there is a fair amount of experimental method, which we might recognize as scientific, incorporated within the whole set of procedures for developing a good item, or the sets of such things we call ‘tests’. However, as this document will make clear, writing a good item is also a highly creative act. By the end of the process something new, powerful, and useful has emerged – a test instrument which has used words, symbols or other materials from a curriculum or a syllabus in a new way, often to serve a variety of educational purposes. In doing so, the item developer needs imagination and ingenuity as well as knowledge: form, structure and balance become important, as they are to a sculptor or a musician.

Why are these building-blocks of tests called ‘items’, anyway? Is it merely educational jargon? Why not call them ‘questions’? The choice of the word ‘item’, in preference to ‘question’, draws attention to two matters. One is that items are often not in question format – the test-taker is required to perform a specific task, or reveal specific knowledge, which is implied in the words given on the test-paper, rather than explicitly offered as a direct question.

The other matter concerns the independence of items in a test. Like the items on a shopping list, they are discrete, or they should be. If you can't get one right, that should not stop you from having a fair chance of obtaining success on all of the others.

The term 'item writing', used in the title of this document, draws attention to this essential independence – the separate skills, abilities or pieces of knowledge which make up human learning are considered individually in the test. This is the prime focus. However the discussion (and the test development process) begins and ends with consideration of a second focus. It considers what happens when these items achieve additional significance or importance by having been grouped or combined with other items to form a test instrument. We must continually remember that our building blocks are part of a larger whole.

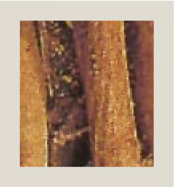
Several times in the paragraphs above, the word 'development' has been used. This is not mere jargon either. It draws attention to the fact that items do not spring to life ready-made in an item writer's brain. A thorough developmental process, often in a defined and specialised sequence, occurs. Like the student knowledge to be tested, it occurs gradually – sometimes with false starts, and often with much wastage on the way, as the test writers clarify their initial ideas, add to them, try them out, and finally decide what will serve their purposes best. Note also the use of the plural 'writers'. Item writing is best done using a team approach at various stages of the exercise.

The purposes, or objectives, of the final test are crucial to the process too. This is where the activity of item writing really starts. What exactly do we want to or need to find out, about student knowledge? Textbooks on educational assessment offer long lists of possible purposes for testing: they cover mastery of processes or knowledge; assessing more general achievement within an educational domain; aptitude for further study; diagnosis of specific learning difficulties, and so on. However, the real objectives for any test come directly from considering two things in conjunction. The first is the actual educational context in which the results of the testing will be used. The second is the knowledge and understanding the test-taker will be expected (or able) to bring into the test-room. The objectives which are determined will therefore be local and specific, and determining what they are is the starting point for determining what the test and its items ought to look like.

Much of the material in this paper concerns one particular branch of item writing which is perhaps the most difficult to master – the development of multiple-choice items and tests. However this emphasis should not obscure the fact that all items, from simple short-answer formats to extended response essays, need the same careful and methodical processes during their development if the final tests are to be reliable and valid measures of learning outcomes. The suggestions made as to process can (and should) be applied whether the user is a teacher in a school preparing a semester test or an administrator setting up a large national assessment program.

Where does basic test-writing expertise come from? Documents such as the present one can provide a survey of some of the intellectual processes required of a test writer, and the organisational structures which will assist the work. However the key knowledge a test writer needs derives directly from professional experience as teachers or educators – our knowledge of the courses we present and the learning habits and patterns of students as they make their way through those courses. This knowledge builds up gradually over years, and while we might not be able to articulate it very clearly to another person as we begin to write the test, we will use it constantly as we review what it is we want to test and how we might best do it. We will sometimes make into ‘testing points’ the sorts of problems that we know from classroom or lecture room experience students find in mastering the material. We don’t do this from a negative position. A 'golden rule' presented later in this document firmly asserts: 'Avoid tricks and trivia'. We are not out to trap or trick as many of the student test-takers as we can. We do it because we need to be able to confirm that students (or a majority of them) have indeed mastered the material we have taught them.

There is no substitute for this detailed knowledge of the teacher's craft and the learning which goes on in classrooms. No matter how expert test writers might be in the art of test writing itself, they will produce poorer tests than students deserve if they are not equipped in this way. If we do not have this knowledge ourselves we should involve people who do within the test development program as often and as decisively as possible. While the processes of testing will use many insights we have learned over the years as to the best ways of measuring human capacities, the measurement aspect should not be allowed to predominate. What must predominate are the specifically educational aspects of the testing. Involving practitioner teachers will help ensure this.



2

Test specifications or blueprints

Why specifications? In most professional endeavours, it is economical of people's time, and necessary from the point of view of sound practice, to have a thoroughly planned view of the whole exercise before one starts to cobble together bits and pieces (in this case, the items), hoping they will do to meet a professional objective or serve a professional purpose. The term sometimes used is 'blueprint', by analogy with an architect's need to design the whole structure in detail before the builders start work.

What do the specifications or blueprints consist of? What needs to be considered? The objectives mentioned in the previous section provide the foundation statement – once this statement is clarified and defined, the other steps may then follow. Other modules in this series will offer more theory and detail as to why and how specifications should be prepared, but they are particularly relevant to the item-writer's endeavours and so a summary is given here, too.

Figure 1 offers a summary of the whole sequence of such steps in the specification process.

Figure 2 presents an example of an initial specification for a recent piece of UNESCO-sponsored test development. The matrix (step 8) for this specimen appears later, as ***Figure 3a***.

Figure 1

WHAT A TEST SPECIFICATION SHOULD INCLUDE

1. The test title.

2. A statement of the fundamental purpose for the test (e.g. testing prior achievement; developed ability; aptitude for further study).

3. The curriculum (or part thereof), or some statement of the learning experiences of the test-takers, which is to be covered by the instrument.

4. A brief description of the clientele or test population (age: educational level or background: assumed knowledge or skill level: any varieties or special groups within this commonality).

5. The range of appropriate assessment types to be used (both in terms of the formats within the proposed test and also any other assessment practices which might run alongside the test itself, such as oral assessments, interviews, practical work, etc.).

6. The intended uses that will be eventually made of the test scores (these need to be prepared in discussions with eventual users of what the test reveals).

7. The time and other relevant conditions available for testing.

8. A detailed matrix which shows how the test will be developed.

Figure 2

A SPECIMEN TEST SPECIFICATION

1. THE PACIFIC ISLANDS LITERACY LEVELS

2. A study of the achieved levels of literacy development by primary school students.

3. The test will encompass writing and reading comprehension skills in English and relevant vernacular languages, together with basic numeracy.

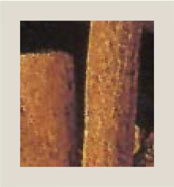
4. The test is intended for students in Class 4 in ten countries of the Pacific region. A national sample based on school types and geographical regions will be tested in each country. (It should be noted that the age of such students will vary from country to country, as will the amount of English and vernacular instruction students have received).

5. The test will elicit samples of writing in each language, based on one task statement per language, together with short-answer comprehension responses based on two short story-passages, again one per language. Twenty-five numeracy items, testing basic computation skills only, will be given to all test-takers.

6. The student scores will be published in five levels for literacy and four for numeracy. Criterion descriptors of the levels will be published. School results will be aggregated. Administrators and policy-makers may then compare English and vernacular performance in each country, and determine overall levels of performance according to gender, geographical and other variables.

7. The total test will be administered on school sites by supervisors external to the school, and will take 45 minutes. Papers will be hand-scored by external assessors.

8. [see Figures 3-6]



3 Developing the detailed matrix

Once the early steps of the sequence of specification have been taken, the test-writer is in a position to be more specific about what the test will actually look like. It is probably easiest to do this in the form of a matrix, with cells to be filled in progressively as work on the test proceeds. The size and complexity of these matrices can vary enormously – for a school end-of-semester examination in a particular course or subject they might get very complex indeed. Many different learning areas will need to be covered and many different item-types used.

Who designs them? In a school, the teacher who is setting the test is the designer, with help and a critical perspective being given where appropriate by a senior teacher or subject leader. If the test is to be given to more than one class taken by a number of teachers, each teacher should participate in the design until agreement is reached that the test would be fair or valid for each class.

At a systems level, such a matrix is designed by a panel or committee consisting of policy experts, curriculum experts and those who will eventually develop the test.

The basic matrix will look something like ***Figure 3***.

Figure 3

	Broad areas or objectives
Detailed content	

If we now ‘translate’ ***Figure 3*** to represent the **objectives** of the UNESCO study used as an example in ***Figure 2***, the matrix would look like: ***Figure 3a***.

Figure 3a

	English	Vernacular	Numeracy
Writing skills			
Reading ability			
Computation			

The next addition to the matrix is the **score or value weighting**. In the specification which led to ***Figure 3a***, a separate score or level rating was to be given for each of the three broad areas of English, the vernacular and numeracy. Hence ***Figure 4*** shows a total of 100% in each column, and becomes:

Figure 4

	English	Vernacular	Numeracy
Writing skills	50%	50%	nil
Reading ability	50%	50%	nil
Computation	nil	nil	100%

The word 'nil' in two of the Computation cells indicates that no reading or writing of words in any language was to be involved in the computational process – everything had to be done using numbers or symbols. In a more sophisticated test for a higher grade level, this decision may well have been changed.

The **time weighting** for the test now becomes easier to define. The specification for whole test (***Figure*** **2**) shows that 45 minutes is available. ***Figure 5*** shows the matrix with appropriate time values inserted.

Figure 5

	English	Vernacular	Numeracy
Writing skills	8 mins 50%	8 mins 50%	nil
Reading ability	8 mins 50%	8 mins 50%	nil
Computation	nil	nil	13 mins 100%

One more element remains to be added to the matrix used in this example – the formats which have been chosen for the test items themselves.

In the UNESCO study being used as an example in this section, only a small space of time was available for each of the cells of the matrix. Hence the solution to the format problem could not involve vast amounts of reading or writing by the test takers, or large numbers of computational exercises to test their basic numeracy. For policy reasons, equal amounts of time had to be given to testing in English and testing in the student's vernacular language, and this suggested selection of formats which were similarly parallel.

The time allocations were suggestive. Thirteen minutes for computations suggested 25 exercises at about two a minute. Eight minutes for reading comprehension suggested 8 short-answer questions (one a minute), and the free writing exercise suggested a sentence every two minutes. So long as the test-takers were given a few minutes to read through the paper before the test actually began, the stimulus material for the reading and writing exercises could be fully informative but not particularly extensive. More reading comprehension questions could perhaps have been asked if a multiple-choice format had been chosen, but this format was decided against – additional information about a student's writing skills would be elicited if they prepared their own sentences as responses to the comprehension questions.

Accordingly, the fully-developed version of the matrix looked like: ***Figure 6***.

Figure 6

	English	Vernacular	Numeracy
Writing skills	8 mins 50% 5 sentences free response	8 mins 50% 5 sentences free response	nil
Reading ability	8 mins 50% 8 questions short answer	8 mins 50% 8 questions short answer	nil
Computation	nil	nil	13 mins 100% 25 questions fill-in-blank

What the literacy test designers in the worked example in this section were doing was trying to achieve an economical and elegant solution to their particular problem of test design, as it had been presented in the early steps of their specification. Had they started at 'the other end' of the process, and tried writing items before they had a view of the whole, much of their time and effort may well have been wasted.

As the test development proceeds, many other aspects of test design will come into play, which ***Figures 3-6*** merely hint at but which are detailed later in this document. One which might be mentioned immediately is the need for **variety**. Test-taking can be a desperately dull experience for the candidate – item writers should aim for some degree of variety in both the stimulus material and in the formats for test response. Even in the 45 minute test used as an example above, there was considerable variety – two writing tasks, two comprehension passages to read (with a different response mode from the writing exercise), and a totally different format for the quantitative exercise.

Even now, the specification procedure is not quite complete. For example, if we take the bottom right-hand cell from ***Figure 6*** dealing with computation, we might be a little more specific about what is to go on when students complete this part of the test.

Decades ago, Benjamin Bloom and others constructed what was called *The Taxonomy of Educational Objectives*. This list divided educational experiences into two main domains, cognitive and affective. Within the cognitive area, he pointed to matters such as acquiring factual knowledge, being able to comprehend, developing the ability to analyse, to synthesise and to prepare evaluations as being a useful (and important) way to structure what we teach and hence what we test. We can use these elements to structure our specification too, to ensure that we obtain a good coverage of what went on (or should have gone on) during the learning process.

Hence our bottom right-hand cell might be expanded to look like ***Figure 6a***. Remember that in this example we are dealing with very young children, so the topics will not be very sophisticated, and the objectives also will be the simpler ones in Bloom's list. We will not be testing the higher-order ones he mentions such as Analysis, Synthesis or Evaluation. The totals could be expressed as the number of items, or the number of marks given for that topic or objective within the total sub-test – or both, as we did in ***Figure 6*** itself, when we were looking at the full test.

Figures 6 and ***6a*** also raise two other important issues with regard to test specification. The first is: 'Just how long should a test be?' There are, of course, no hard and fast rules, but common practice suggests that 45 minutes is a **maximum** for any sort of formal testing in middle primary school. This maximum might rise to an hour for upper primary and junior secondary, two hours for mid-secondary, and three hours for the most senior school students. The test-writer might comfortably achieve a satisfactory coverage of the topics and objectives to be tested in less time, and should aim to do so where possible. Two tests a few days apart will work better that one large, over-long one.

The second issue relates to the fact that it is very easy to write items to test simple knowledge, and too often that is all that test-writers do. They forget application, analysis, synthesis and so on. Putting these categories clearly into a test specification reminds us that they are there to be tested, and they **need** to be tested. They may also require different mark weightings: one mark for a simple factual recall item is fine, but an application or detailed analysis of some learning may deserve more, as in ***Figure 6b***. The number of items per cell will depend on the available time, and your estimate of a good balance to cover all the learning to be tested. Remember too the test writer's responsibilities to create a good impression on teachers, and have a positive effect on learning. If the instrument does no more than test factual recall, often that is all that will be taught – teachers are great ones for scanning past papers to see what is expected, in order to give their students the best chance. If they find nothing but lower-order skills, then learning and teaching in the whole education system may become the poorer for it.

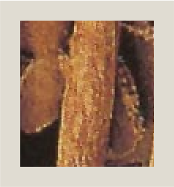
A final review of the specifications (including the finished matrix) then needs to take place. The basic question during this review relates to the decisions which have been taken from the point of view of **coverage of the curriculum**. The design panel needs a positive answer to the following – "how adequate is the proposed design as a sample of the areas of learning or knowledge under review?" Even now, the item-writer might not be quite ready to start work on the actual items!

Figure 6a

Objective or behaviour Classroom topics or content	Knowledge	Comprehension	Application	TOTALS
1. Addition				
2. Subtraction				
3. Multiplication				
4. Division				
TOTALS				

Figure 6b

Objective Content	Knowledge	Comprehension	Application	TOTALS
1. Addition	2 items @ 1 mark	2 items @ 1 mark	2 items @ 2 marks	6 items 8 marks
2. Subtraction	2 items @ 1 mark	2 items @ 1 mark	2 items @ 2 marks	6 items 8 marks
3. Multiplication	2 items @ 1 mark	2 items @ 1 mark	2 items @ 2 marks	6 items 8 marks
4. Division	3 items @ 1 mark	2 items @ 1 mark	2 items @ 2 marks	7 items 9 marks
TOTALS	9 items	8 items	8 items	25 items 33 marks



4 Setting parameters for layout and instructions

A start should now be made on developing a view of the layout of the eventual test paper and sketching, before they get forgotten, the instructions which the candidates and supervisors will need to smooth the test process.

Questions of economy may well arise first, if resources are scarce. Here are a few:

- how much paper will be available for printing papers for the total number of candidates to be tested? Or will that not be an issue?
- will this mean a four- or eight-page test paper, or will a longer one be possible?
- will students answer on the test book (which means it is not re-usable) or on a separate answer sheet?
- will the test be hand-scored or machine-scored, by an optical mark-reader (OMR)?
- is colour printing of any illustrations on the paper feasible?

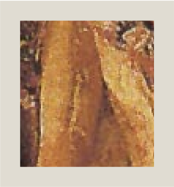
Final answers to these questions will not necessarily be decided at this stage, but preliminary answers certainly should be. Obviously with layout much will depend on the actual items used. However, the broad parameters ought to be ready for the item-writers to have in mind once they begin work.

Economy matters in a different way, too – economy of words and space. When the test writer receives the detailed specification and the layout design, it ought to contain all the detail which will be needed during the item writing process itself. Before this process can start, consideration should be given to including the following within the intended layout:

- an introductory section, where student information will be gathered. In a school, this may be no more than the candidate's name and class group. However, in a larger test program, school name, sex, and various other pieces of personal information may take up valuable space.
- a section of the paper where the criteria for assessment are specified for the students to read before the test begins.
- various procedural hints for the candidates on how they might do their best on the test.
- layout explanations, if the paper is in sections, or if students are expected to spend varying amounts of time on particular parts of the test.

In large test programs, particularly standardised ones, some of these might also need to be converted into instructions for supervisors to speak. This will help to ensure that all candidates (wherever they are sitting the test) have access to the same information. Printing the information is often effective, but having a supervisor saying it aloud guarantees that everyone has been given the same chance.

In both test-paper instructions and supervisor's script, aim for the **minimum number of words** to convey the essential messages.



5 Item types or formats

In preparing ***Figure 6*** above, the literacy item-writing team made a selection from a wide range of options so far as appropriate item-types or item-formats were concerned. Later in this section (particularly in ***Figures 10*** and ***11***) some view of this range is specified.

However, first it would be useful to look at what all items might have in common: see ***Figure 7***. This will help to distinguish between the various options when they are presented later.
Figure 8 offers some examples of stimulus material used in tests, while ***Figure 9*** shows how the parts of an item are exemplified in one format in particular – the multiple-choice item.

Figure 7

THE PARTS OF AN ITEM

1. **All items use stimulus material of one sort or another.**
   - Sometimes it is no more that a sentence or a set of symbols which directs the student what to do.
   - Sometimes it is a passage, or a diagram, or an illustration, relating to a whole set of items, which informs the student.
   - Some items have both directive and informative material to stimulate candidates' thinking.

…

…

2. **All items have at least one 'right' answer or response, in the sense that it would earn full credit from an assessor if offered by a candidate.**

   - Sometimes the choice is 'closed', as in multiple-choice items. The right answers are actual and printed on the paper as an option.
   - Sometimes, as in essay tests, these responses are potential – not realised until someone reads the stimulus and makes the response. (Sometimes, of course, in reality nobody does get full credit – but all items should be written in such a way that somebody might.)
   - Some items are 'open'. They have more than one 'right' answer, such as two essays which each score full marks even though they are different. (Multiple-choice items never do.)

3. **All items will have inadequate answers, which might earn partial credit from an assessor, or wrong ones which would earn no credit at all.**

   - In most multiple-choice tests, inadequate or wrong answers get no credit at all.
   - In most essay tests, very few answers get no credit at all unless a blank page has been submitted.

ANSWERS OR RESPONSES – RIGHT OR WRONG, ACTUAL OR POTENTIAL – ARE ALWAYS CONSIDERED TO BE A PART OF THE ITEM.

Figure 8

SOME SAMPLES OF STIMULUS MATERIAL

informative

- a passage in a multiple-choice test to which a set of items refers
- a map in a Geography test which candidates are expected to refer when answering some short-answer questions
- a photograph in an Art test about which candidates are expected to write an essay
- a diagram in a maths test which forms the basis for the solution of a problem

directive

- the leading sentence or ‘stem’ of a multiple-choice item, such as: “In the passage, who ate the cake?”
- a short-answer item such as: “Use the scale of the map to calculate the distance from the tower to the bridge”.
- a specific essay topic, such as: “Compare and contrast the painting by Picasso in the photograph with two other paintings you have studied by the same artist”.
- an extended response item such as: “Write a critical review of one novel you have studied this semester”.
- a statement of a problem to be solved, such as: “What is the area, in square metres, of the shaded part of the diagram above? Show all your calculations”.

STIMULATE MENTAL IS INTENDED TO STIMULATE MENTAL PROCESSING DURING THE TEST, EITHER GENERALLY OR DIRECTLY.

Figure 9

THE PARTS OF A MULTIPLE-CHOICE ITEM	
INSTRUCTION to candidates	Read the following passage and answer the question which follows.
INFORMATIVE stimulus material	*Bob, Carol, Ted and Alice had just begun a friendly game of poker, but already Ted had much the biggest pile of winnings. Carol had won a small sum, and Alice had lost more than Bob.*
DIRECTIVE stimulus material	
* **the NUMBER and STEM**	**1** At this stage of the game, who had lost the most money?
* **OPTIONS for answering**	**A** Bob **B** Carol **C** Ted **E** There is not enough information to say.
KEYED RESPONSE or right answer	**D** Alice
DISTRACTORS or wrong answers	**A**, **B**, **C** and **E**

Test items can be usefully classified into three main categories:

1. Selected response items

A response is selected by the test-taker either from a given list of possible choices, or from the stimulus material itself. The choice is 'closed': that is to say, only one option will receive credit. The type includes the following formats:

a. True-false

According to the map, Bombay is the capital of India – true or false?

b. Matching items

Find the word in the passage which means HOLLOW.

Match these words with their antonyms:

INEBRIATED : SOMNOLENT : LUGUBRIOUS

CHEERFUL. . . . .

SOBER . . . .

WAKEFUL . . . .

CALM . . . .

It is usual in items such as the above to include more options than the given number of matches required.

c. Classification items

Fill in the blanks:

FRANCE *PARIS* *EUROPE*

. . . . *KENYA* *AFRICA*

. . . . *BOGOTA* . . . .

SRI LANKA . . . .

The terms of the classification can either be given (country/capital/continent) or left out, as in this example, to be inferred by the candidate from the samples given for the categories. Simpler, two-column classifications are of course possible.

d. Multiple-choice items

See ***Figure 9*** for an example. It should be noted that multiple-choice items can come in two styles. One might be called stimulus-related, and the item in ***Figure 9*** is an example of this kind. The other might be called 'stimulus-free' in that options are printed, but the candidate is expected to draw on prior knowledge, gained during classwork, to select the keyed answer.

They can also vary according to structure. The example in ***Figure 9*** has five options from which to choose. Other items may have only four options: these are simpler to write and often easier to answer. Three or six option items should be avoided: the former increase the chances of successful guessing, and the latter are simply too complex to handle in the circumstances of a test. Whichever structure (four or five options) is ultimately chosen, all items in the multiple-choice section or test should conform to that structure.

In all testing using selected response items, it is important to ensure the independence or discreteness of each item. No late item in a series should depend, for understanding or for choosing the correct response, on the test-taker having got a previous item right.

2. Constructed response items

In items of this kind, the candidates construct or prepare their own responses on the spot, based on knowledge developed during a course, or by rewriting bits of given stimulus material in a new way. Many such items are 'open', in that equal credit during scoring might be given to a number of totally different responses. Others, such as some of the examples below, will have only one right answer.

a. Short-answer items

Write a good title for the story you have just read.

What did the three cousins do in the story?

Why do you think Tina was angry?

b. Fill-in-the-blank and sentence completion items

16 x 4 = ... or . . . .

Saigon was re-named Ho Chi-minh City, because . . . .

The word . . . . accurately describes Jo's behaviour during the argument.

c. Cloze items

One subset of completion items might be especially mentioned. These are so-called '**cloze items**', where candidates' word knowledge is tested by asking them to insert appropriate words in regularly spaced gaps in a passage of prose (e.g. where every fifth word has been deleted, and a space left). The format began in reading research as a test of readability of prose, but has been adapted for educational testing purposes. Here is a sample of a pure cloze task:

One subset of completion . . . . might be especially mentioned. . . . . are so-called 'cloze items', ... candidates' word knowledge is . . . . by asking them to . . . . appropriate words in regularly . . . . gaps in a passage . . . . .prose (e.g. where every . . . . word has been deleted, ... a space left).

Some testers use a slightly different form, called 'modified cloze' where the omissions are not regular, but selected to test particularly important words or concepts, or those gaps where only one response is logically right. Here is the same passage in 'modified cloze': even here some alternative insertions are possible.

One subset of . . . . items might be especially mentioned. These are so-called ' . . . . items', where candidates' . . . . knowledge is tested by asking . . . . to insert appropriate . . . . in regularly spaced gaps . . . . a passage of prose (e.g. where . . . . fifth word has been deleted, ... a space left).

d Extended responses (paragraphs and essays)

These tasks can either be specifically tied to informational stimulus given on the paper, or rely on the candidate's bringing knowledge and understanding of a topic into the test-room in order to respond to the directive on the test-paper.

> *Read the material presented opposite and prepare a balanced and detailed critical response to the ideas about education presented by the various writers and cartoonists. Your piece of writing should be between 600 and 1000 words.*

> *Write a paragraph which summarises Fairbank's view of the First Opium War as expressed in the passage.*

> *Write an essay which traces the impact of the economic theories and ideas put forward by John Maynard Keynes on the current economic policies of this country.*

Extended responses give the item writer a much greater chance to emphasise the **complexity of knowledge** and ideas and, where appropriate, the **sequence of events** or **logical connections** between ideas. In all these items, later cognitive operations in the minds of the candidates may depend heavily on right choices having been made early in the test experience. Hence it is important to stress (in the instructions) the need for students to undertake **planning and drafting** of their work, and if possible allow time for these somewhere during the test session.

These formats often bring into sharp focus the problem of **choice between different tasks** on differing aspects of course work. If it is at all possible, choice should be avoided and all candidates asked to perform the same task. Where, because of the extent or complexity of course learning, choice is unavoidable, care should be taken to make the various choices as nearly **approximate in conceptual difficulty** and **ease of execution**. This may not be possible without field testing of the various options beforehand.

3. Problems for solution

The problems which test-writers set for solution by candidates actually represent a sub-category of constructed responses, but seem important and common enough to be discussed separately. In a sense, setting an essay task for extended response has a problem element built into it, especially if the candidate is required to read a large amount of informative stimulus and choose some sort of personal response to be developed during the test. For example, we might look again at an example (given earlier under the heading 'Extended Responses') with this in mind:

> *Read the material presented opposite and prepare a balanced and detailed critical response to the ideas about education presented by the various writers and cartoonists. Your piece of writing should be between 600 and 1000 words.*

Other problems may not require words as a response, but quantitative reasoning and expressions for their solution. Or perhaps it is manipulation of materials that is required, as in a test conducted in a laboratory or workshop or art studio.

Mathematical problems are probably the most common kind, from simple primary examples to more complex ones at higher levels of schooling. As in the first example, they too might involve manipulation of objects as an aid to their solution.

John's boss has told him to set up six jars as a display for a jelly bean promotion. John sets up six jars on a shelf, three full ones then three empty ones.

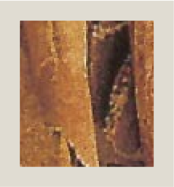
"How does it look?" says John, about to leave for lunch.

"Well, I'd like it better if you alternated full and empty jars', said the boss.

John is hungry and in a hurry. What's the least number of jars he has to move?

Look at the photograph and plan of the room. You will notice none of the corners is a right angle (90 degrees).

Prepare a measured drawing of a piece of furniture (cupboard, table, etc.) which would be appropriate to the room, and would fit the angle of one of the corners exactly.



6 Selecting item types for test purposes

As stated earlier, the basis for any selection from within the wide range of possible formats broadly categorised in the previous section lies in the learning to be covered. The test writer must achieve the best possible match between this learning and the potential instrument, as revealed in the specification.

All formats have their advantages and disadvantages. A summary of these is given in ***Figures 10*** and ***11***. It is, however, merely a summary. Multiple-choice items, for example, have a large number of other problems associated with their construction. For example, plausible distractors are hard to write; the items present wrong information (the distractors) as if it were right, perhaps reinforcing a wrong view in the student's mind which gets taken away from the test room; the format ignores the students whose developmental stage leads them to be half-right, half-way to full understanding – students get no credit for this. Moreover, if a curriculum suggests that we should be testing student performance with regard to aesthetic and affective criteria, or the making of critical or value judgements, multiple-choice item writers will find it hard to represent these satisfactorily in their items. The format also disallows a plurality of answers, to represent the complexity of much student learning, unless the items themselves become very strained.

Above all perhaps is the fact that multiple-choice items are 'closed'. That is, they do not allow for the diversity of human opinion which legitimately informs much of our appreciation, apprehension and interpretation of the things we learn. A right answer, one right answer only, is chosen for the candidate. Sometimes this doesn't matter, when there is only one possible right answer, in the objective sense. But at other times the 'right answer' might be a quite subjective matter of the test writer's opinion, and the candidate might have a different – and legitimate – view. For instance, in making the critical judgements mentioned above, human beings often differ in what they value or recognize as being important. This is normal and legitimate – we are entitled to our views if they are based on fact and reasonable interpretations of what we read or see. Multiple-choice items cannot handle this variety.

Many of these problems are solved by the selection of an 'open' extended response format, such as an essay. But these formats too bring some of the problems associated with subjectivity, though not when the student is preparing the response and can argue for his or her own point of view. The problems come later, when someone has to sit down at the task of assessing the value of the work done. Any assessor brings to this task a view of what are appropriate personal opinions and responses to the 'topic' of the student's writing, sometimes amounting to bias or outright prejudice.

Before we take fright at all these problems, and allow our testing to degenerate into mere assessment of general knowledge or the most simplistic kinds of understanding and cognitive processing, we should remember that testing of abilities in analysis, synthesis and higher-order evaluations is possible, and should be done (as an earlier section pointed out). It is just that one format can't do them all. We might never completely solve the guessing problem ('did the student **really** know this?') but a range of formats in one test will allow us to get a detailed picture of a wide range of knowledge, understanding and ability – even creativity – as learning outcomes. It's not always easy to achieve this picture, but it can be done, and it is often worth the attempt.

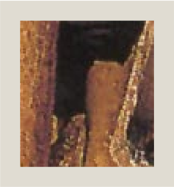
Especially in examinations. In this circumstance, there is a wide range of purposes for the testing, some of which will require different formats. Not only will variety help the students to make their way more easily through the experience, but (as ***Figures 10*** and ***11*** indicate) certain formats lend themselves more readily to certain purposes than others might. For example, testing of a range of detailed factual information is readily done using matching or classification items (***Figure 10***), whereas testing the candidates' capacity to perform integrated, higher-order skills such as synthesising knowledge or attempting evaluations might mean setting a complex problem for solution, or need an extended response to do those skills justice (***Figure 11***).

Figure 10

CRITERIA FOR CHOICE OF SELECTED RESPONSE ITEM FORMATS		
ITEM TYPE	**ADVANTAGES**	**DISADVANTAGES**
A True-false	• easy to write • easy to mark • easy to sample variety within a course	• guessing factor very high (50%) • limited to unequivocal choices • cannot test higher order skills
B Matching items	• useful for testing relationships • useful for testing factual information • easy to construct a large number	• the cluster approach destroys item independence • difficult to word instructions
C Classification items	• relatively easy to construct • easy to mark • useful for testing factual information • useful for testing simple relationships	• the cluster approach destroys item independence to some degree • limited to factual sorting • limited to unequivocal facts
D Multiple-choice items	• reduces the guessing factors • versatile – can be used to measure a wide range of cognitive processes • reduces problem of subjective scoring • analysis of results can provide much diagnostic information • easy to mark	• little, if any, stimulus given to creative thought • expensive and time-consuming to construct • difficult to measure organization and presentation of ideas • plausible distractors hard to write • presents wrong information as if it was right

Figure 11

CRITERIA FOR CHOICE OF CONSTRUCTED RESPONSE ITEM FORMATS		
ITEM TYPE	**ADVANTAGES**	**DISADVANTAGES**
A **Short-answer items**	• excellent for testing factual knowledge • successful guessing is reduced • easy to write • easy to mark	• unsuitable for measuring complex learning • easy to respond in inappropriate ways
B **Fill-in-the-blank sentence completion**	• easy to test a range of factual knowledge • guessing factor is reduced • easy to write • easy to mark	• hard to measure higher-order skills • easy to respond in inappropriate ways
C **Cloze, modified cloze**	• easy to construct • a good measure of word knowledge • tests passage understanding	• some ambiguities hard to decide on • many opportunities for choice have little value • often little more than guesswork
D **Extended responses**	• a means of assessing higher-order skills • relatively easy to construct • stimulate creative and critical thought as well as learned responses • can measure learning in affective domain	• sometimes lead to inadequate sampling of learning done • time-consuming and expensive to mark • difficult to achieve inter-marker reliability
E **Problem solutions**	• a means of assessing higher-order skills • can measure complex learning outcomes • relatively easy to construct	• can be time-consuming to mark • sometimes difficult to establish stable assessment criteria



7 Item writing as a creative act

The introduction to this document suggested that item-writing was a truly creative act. We have set our specifications, considered the curriculum and its learning outcomes, reviewed the options for format-types, and made a selection. We are now in a position to start the creative process.

From the start, three sets of cardinal distinctions might be kept quite clearly in mind, as well as one golden rule. The distinctions – the first and third are particularly important for multiple-choice test writing – are between:

- simple comprehension and higher-level interpretations;
- 'open' items and 'closed' items;
- factual knowledge and inferential reasoning.

These simple distinctions will help improve the test quality overall. They will help us to remember to develop items which do not merely test factual knowledge and simple understandings but tap into higher order skills which students develop as they learn.

Some easy, knowledge-based, or simple comprehension items have a place in any test. They help ease students into the test-taking. But we should recognize that they sometimes don't require much more than simple cognitive processing – searching for facts which are fairly obviously in the stimulus material. The answering process is 'closed' – there is only one right answer. If the items never get beyond this level an important opportunity has been missed: there is more to learning than this.

We should also expect our test takers to be able to read ‘between the lines’ as it were – for example, to be able to:

- draw inferences which rely on implied or logical connections between events or facts;
- understand implications or key underlying assumptions which might not be stated at all in the material but which nevertheless are important things to learn or understand about it;
- analyse, summarise and develop a personal and evaluative critical perspective.

All these might come under the heading of ‘interpretation’. The stimulus material is printed. The candidate develops his or her own interpretation of it, and uses this in answering these more ‘open’ items. There will be no one right answer – many responses from quite differing personal perspectives will deserve, and rightly be awarded, full credit.

The golden rule, for all test-writers, is:

AVOID TRICKS AND TRIVIA

As item-writers we are not out to trick the candidates in any way. Our professionalism as educators should mean that we really want to find out what they know, rather than what they don’t. ‘Trick’ questions are out. Of course some questions may well be so difficult for a few students that they might regard them as trick questions, but that is not what is meant. Our intention should always be that any student who has engaged in the course being tested, or who has developed the relevant skills, has a reasonable chance of performing well on the item. By all means choose material or items which focus on substantial misunderstandings or mistakes which you know students might have or make, or which reflect on what you know to be a common and important difficulty for students doing the course. But make sure that it **is** common and important, not merely a devious or slippery misinterpretation you have invented to trap the unwary.

Trivial questions likewise should be avoided. In a test, we have little enough time to test the important facts or concepts or understandings of a course, without wasting this precious commodity on inessential or peripheral or simply irrelevant matters. This applies to distractors in multiple-choice items as well – don't waste everyone's time in preparing a distractor which virtually every test-taker is going to recognize as trivial, and hence to be ignored.

Beyond these there are a number of other central matters which an item writer needs to keep constantly in mind. These are summarised in ***Figure 12***.

The last two points in ***Figure 12*** will assist in avoiding 'test writer bias'. A single view of the curriculum to be tested might be narrow, or even faulty, even if the test constructor is experienced and an acknowledged expert. In school-based test development it is often impossible to achieve this variety, where only one person really knows the course as taught. Elsewhere it often is possible, and is a good principle.

Assuming that all that has been done, the team has been set up, curriculum has been clarified and formats decided upon by the group, the next stage of the work is individual – drafting and revising the item material. As an example, we might look at an individual working on development of items in one particular format – perhaps the most difficult of all – multiple-choice.

Figure 12

PRINCIPLES AND PROCEDURES DURING ITEM DEVELOPMENT

- Use the specification, don't ignore it – have it handy throughout the development.
- Allow (and take) as long as possible for the whole process.
- Once the format is chosen, select the material and give yourself time to ponder it: familiarise yourself with its main points and other minor ones which might form the basis for good items and distractors.
- Set up the scoring procedures and develop any assessment criteria simultaneously with the test development.
- Develop the items using co-professional input – other teachers who might be involved in the course, or who will use the results – during material selection, item review and editing.
- If possible arrange for more than one person to work on actually developing items, and use a range of items from these different sources in the final test product.

Figure 13

THE PROCESS OF MULTIPLE-CHOICE ITEM DEVELOPMENT

1. First, search for informative stimulus material related to the course or the objectives of the test. In rare cases, the material might have to be written or otherwise prepared by the item-writers themselves. However, more usually it exists already and can be chosen from books, periodicals and other sources. Keep records of where you found the materials.

2. Look for a variety of stimulus material on a single topic. The range of types of stimulus includes written, pictorial, graphical and tabular material. Keep your candidate audience in mind.

3. A decision is made about whether one stand-alone piece of stimulus material will do to test a curriculum element, or whether several pieces in conjunction would provide a better test, allowing possibilities of comparative items.

4. Extra, relevant pieces should be selected, in case the first ones prove to be less impressive than they first appeared.

5. Read, and re-read (perhaps several times over) the material, and make one-line notes about possible testing-points discovered during these readings.

6. If no testing-points appear in some sections of the material, then look at the possibility of cutting the material to remove the extraneous section(s). But don’t hack it to pieces – meaning tends to get lost!

7. When you feel that the material and its possibilities have been fully comprehended or digested, the one-line notes from Step 5 are sketched as possible ‘stems’ for individual items.

...

…

8. Write these sketches at the top of individual sheets of paper or cards (one per item), and just under them some preliminary sketches for possible distractors might be made as they occur to you. Do these in pencil, not ball-point.

9. It is not necessary to 'finish' one item (complete with stem and distractors) before going on to the next. Ideas will emerge at different times (sometimes quite inconvenient ones), especially if enough time has been allowed and the process isn't rushed.

10. Once an item has been sketched, underneath it write a draft in the correct format. If you've used separate pieces of paper the items can be drafted in any order (or even left incomplete for the time being, if a third or fourth distractor just won't come to mind). Do all this in pencil, too.

11. Assemble the various pieces of paper (or cards) into what seems a reasonable order in terms of the difficulty of the items – the general rule is "easiest to hardest", but there are often exceptions to this. Place the stimulus material on top of the pile and have the sequence typed out.

Figure 13 summarises an enormously complex process, and needs a little elaboration here and there. In **point 1** the importance of keeping bibliographical records of where you found your material might be stressed – you will need this later when seeking permissions to re-publish it, and such details easily get lost.

At **point 2** you will need to consider the appeal of the material for various groups of test-takers: will it appeal to both girls and boys? Urban and rural dwellers? Is there some significant sub-group who will fail to understand it completely because of language or dialect problems? (Would it help to asterisk especially difficult words and define them at the foot of the passage?) Or are there particular emotional overtones (such as references to death or disasters) that might disadvantage particularly susceptible individuals?

At **point 5**, you might ask yourself: 'Why does this piece appeal to me? What would I hope my candidate-readers would learn from the experience of reading this piece or looking at this picture?' In these ways you might develop for yourself a fuller understanding of what makes the stimulus tick, as it were. Simultaneously, it will ensure that you see the central importance of the piece, as well as exploring some of the details that later will come in handy for item or distractor development.

At the stage of **point 7**, you will need to remember two things: the candidate's background knowledge is not expected to be an alternative to, or substitute for, using the stimulus material. The intention is that they cannot do the item without having read the stimulus. Similarly, they should be able to understand the problem from the stem, directly not have to search through the options before they understand what the question is actually about.

When you are sketching at **point 8**, if you are using a longish passage of text, you will probably find yourself including line references to help you 'key' your draft options back to the passage. Give some consideration as to whether you might do this in the final test also. Too often, test-takers find themselves involved in a search-and-destroy mission trying to find where something mentioned in an item is also mentioned in the passage. This creates an artificial difficulty you can avoid for them. Save them the valuable time they would otherwise lose: it doesn't make it any easier to decide on the right answer!

During the final assembly of your draft items (**point 11**), you will also have an opportunity to review the content and quality of the items you have written. You will be looking for a range of difficulty. You will also be able to see the spread of the types of questions you have prepared: are there enough global questions? Too many particular questions focusing on no more than vocabulary? Is inferential reasoning well represented? If the mood of a passage is important, is it in an item in your pile? All these considerations will help you develop an effective sequence for the material as you present it during the panel meeting which is the next stage.

Here is an example of the process outlined in ***Figure 13***. It should be noted that the assumption is always that one passage yields a set of items, not just one. Stand-alone multiple-choice items are hardly an economic proposition in terms of space and time, and require too many mental gymnastics on the part of the candidates who face them. They also sometimes give the impression of learning as being a matter of unconnected bits and pieces – hardly an impression we want our students to carry away with them from the test room.

stimulus material – Steps 1-3

Passage

Multiple-choice item writing is a difficult art at the best of times but how often we make it more difficult than it need be! We start off thinking it will be easy and go at it very quickly before all the matters we need to keep in mind have been considered. Rushing the process won't help – items do not spring fully-formed on to the page. We need to ponder, review, get to know what it is we want to ask and what the material might let us ask. Quickly prepared items are often a waste of time – we find that out when we show them to another person, and that person merely says: 'trivial' or 'faulty'.

one-line note – Step 5

main point

sketch for a stem – Step 7

What is the main point of the passage?

sketches for distractors – Step 8

- *the difficulty of doing it*
- *time*
- *logical order*
- *plausible distractors*
- *getting a good critic*
- *avoiding trivia*

draft of stem and revised distractors – Step 10

1. ***What is the main point of the passage?***

A *the time it takes*
B *the logical order of item development*
C *plausible distractors hard to find*
D *getting a good critic to look at the item*
E *finding enough distractors*
F *avoiding trivial questions*
G *the difficulty of item writing*

It would also be possible to construct a negative item using much the same material, to emphasise something of the complexity and variousness of the views expressed in the passage – but only if we saw that to be a point worth testing:

draft of negative stem and revised distractors – Step 10

1. ***Which of the following aspects of item-writing is NOT considered in the passage?***

A *the time it takes*

B *the logical order of item development*

C *plausible distractors hard to find*

D *getting a good critic to look at the item*

E *finding enough distractors*

F *avoiding trivial questions*

G *the difficulty of item writing*

Opinion varies on the value of such negative approaches to stimulus material. They are certainly harder for students than straightforward approaches, but just occasionally they are unavoidable. They should never include another, double negative in any of the options.

As an item-writer moves through the process and constructs a number of items on the chosen material, a number of subsidiary issues emerge as needing consideration. Some in the following summary have been mentioned before, some are new:

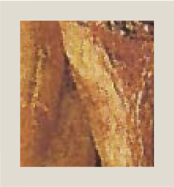
- Is the passage too long for the items written?
- Are there words in the passage which are too hard for the candidates?

- Are the stems clearly worded?
- Are the items independent or will the answer to number 4 be dependent on getting some other item right?
- Are the items independent or will the distractors in item 10 give away the answer to number 2?
- Is this distractor ambiguous?
- Do all the distractors share the same syntax and grammar?
- How plausible are these distractors?
- Do all the distractors in a given item refer directly to the stem or are some wild?

This might also be the place to consider just how many options (four or five) in a multiple-choice item one is going to have in the final test. Professional opinion amongst item-writers varies about this issue. Five options does cut down the correct-guessing rate, but the number does make life harder for both the candidate and the item writer. The feeling of the present writer is that four options will do – a larger number of elegant, straightforward items might appear on our test-papers, without the fifth, strained or irrelevant distractor that is all we can think of. However, by all means present five (or six, even) possibilities to the panel who will review your drafts.

Two other hints are worth mentioning: one is to read through the items aloud in the voice of a 'student' who might do the test, thinking all the time of how they might respond to the material – this helps one to distance oneself from the material over which one has sweated long and hard. The second hint is to allow sufficient time to leave the sequence of items alone for a few days, and then come back to it afresh.

But to solve these (and many other) problems ultimately we need help – fresh minds and critical eyes, just like those of the students we will be testing. So begins the next stage of the development process.



8 Panelling or moderating drafted items

Critical review of items is essential. And it is essential before the items and the item-writer get locked into a situation where they have to make do with what they've got. Two words are sometimes used for this part of the process – 'panelling' draws attention to the need for a panel of reviewers, or more than one critic; 'moderating' draws attention to the actual process of the meeting – moderating the more extreme efforts of the writer by exposing the work to other views.

Though it may seem wasteful, panels should always be supplied (as far as is possible) with more than is eventually needed: more items and often more distractors for each item. However the items themselves should be in as good a condition as possible before they are duplicated and sent to panel members:

- worked over as much as time permits, not just raw drafts
- typed, not hand-written;
- numbered and ordered, not random;
- complete with stimulus;
- any line references made have been indicated clearly in passage and items;
- laid out in standard form.

The standard for multiple-choice items is as in:

Figure 14

STANDARD LAYOUT FOR A MULTIPLE-CHOICE ITEM

item number in bold, and indented stem	**23** ***How many people are in the group?***
option letters in bold capitals, and options further indented	**A** *less than ten* **B** *between ten and twenty* **C** *more than twenty* **D** *The passage does not say.*

- Any option which is a full sentence should have a full stop or period, as in **D** above.
- Option letters should not have full-stops or brackets.
- If a negative word is used in the stem, such as **NOT** or **EXCEPT**, it should be printed in bold capitals.

No indication should be given in advance to the panellists of the author's suggested right answer. Their task is to work through the items as if they were doing a test, and come up with their view of what is 'right'. However the keys must be marked on the item-writer's copy to be used during the panel meeting – panellists get justifiably restless if in the heat of the moment the item-writer can't remember what was to be keyed.

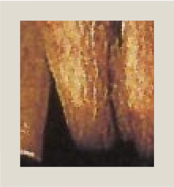
The number and choice of panellists will depend on circumstances and resources – panellists wherever possible should be paid! The need for confidentiality must be stressed. There will be a panel chairperson (not an item-writer whose work is up for review – there is too much else to do), who might also act as a co-ordinator for distribution of the materials well in advance. Other panellists are chosen for their expertness, the variety of viewpoints they can contribute, and their number should include some representation from those who will eventually use the results. Gender balance should be maintained where possible. Each item writer whose work is being reviewed acts as secretary to the panel while those items are being discussed.

Running a panel is not as easy as it might seem. Confidentiality must be ensured before and during the sessions – some programs ask panellists to sign legal declarations that they have and will observe this need. In the sessions themselves, time very rapidly runs away as people present and defend their points of view, or argue about detail. Chairpersons have to be very firm about closing off trivial discussions!

The sessions often also give rise to quite sharp inter-personal conflicts, especially where the item-writer is new to the business. It can be very difficult for such a person not to feel personally attacked – his or her intellectual capacity, or even personality! It is natural for the item-writer to be intent on acting as defendant of as many of those hard-won items as possible. But this should be tempered by the knowledge that things **will** have gone wrong – even the most experienced item-writers mis-read passages or leave out essentials. When such matters do come to the attention of the panel (and that is why it is there, after all) it is up to the chairperson to keep tempers cool and comments constructive rather than personal. Participants are there to criticise the work, not the person, and comments from an outsider such as 'I just don't understand how you [the item writer] think' are totally unacceptable.

Where possible, tape-recording the session should be undertaken as well as the keeping of written records. Sometimes, in the heat of intellectual argument, points get made which might be missed by the item writer-scribe but be quite valuable during the later editing of the item concerned.

Opinion varies as to how long a panel meeting should last. The standard of item criticism seems to fall away quite sharply after about two hours. However, it is often more economical to go on a little longer to finish the work rather than re-convene everyone at a later time, and the chairperson might need to keep that in mind in reaching a decision about when to stop.



9 Stage One – editing or vetting

The item writer using the panel meeting records, written or taped, to edit or vet the draft items as meticulously and slowly as possible. The panel will have engaged in a variety of discussions and made many suggestions: they will have offered hunches about the validity or difficulty of items, have given their perceptions about the plausibility of distractors, and have pointed out actual errors of fact or language use. It is now up to the item writer to respond and accommodate to as many of these comments as seem sensible from a professional viewpoint. Not all will be – panellists are sometimes wrong! – but most will be invaluable in getting the final form of the items just right.

The vetting process involves some or all of the following:

- cutting or adding to the stimulus material where the panel feels this is necessary;
- choosing words in the stimulus material for asterisking and defining at the foot of the passage, if the panel thinks the candidates won't know them but they are vital for understanding the passage;
- deleting whole items the panellists think are trivial or too difficult;
- writing whole new items suggested as replacements or additions by the panellists;

- deleting the less successful draft distractors, or ones the panel feels are implausible;

- writing new distractors suggested by the panel;

- checking the language at every turn – accuracy, consistency, clarity;

- establishing a final order for the items in the set or the whole test.

Much of this activity will seem like unnecessary wastage, but it is (on the contrary) essential – only the best distractors, keys, stems and stimulus material should be taken out into the field for pre-testing. Sometimes the circumstances of this field testing mean that slightly variant versions of the same item can be tried on different groups. This is sometimes a useful procedure where the item-writer cannot make up his or her mind as to whose hunch to follow – one's own or a panel member's.

As an example of what might happen to an item during the vetting process, let's look again at the stimulus and draft item presented earlier.

Passage

Multiple-choice item writing is a difficult art at the best of times but how often we make it more difficult than it need be! We start off thinking it will be easy and go at it very quickly before all the matters we need to keep in mind have been considered. Rushing the process won't help – items do not spring fully-formed on to the page. We need to ponder, review, get to know what it is we want to ask and what the material might let us ask. Quickly prepared items are often a waste of time – we find that out when we show them to another person, and that person merely says: 'trivial' or 'faulty'.

1. ***What is the main point of the passage?***

   A *the time it takes*
   B *the logical order of item development*
   C *plausible distractors hard to find*
   D *getting a good critic to look at the item*
   E *finding enough distractors*
   F *avoiding trivial questions*
   G *the difficulty of item writing*

During the panelling process, we need to inspect these suggestions very carefully. The item writer has offered us more options than we will need, so we might begin with them. What does each offer as a summary statement of the process we have learned about in the passage?

Time is certainly mentioned (**A**), as is the logic of the process (**B**). Even though we might know that plausible distractors are hard to find (**C**) or indeed enough distractors (**E**), the passage doesn't actually mention these matters, so they are the first candidates for removal – relevant to the issue of item writing, but not to the passage. Trivial questions do get a mention (**F**), as does the difficulty of item writing in general (**G**).

So our item now looks like this:

1. ***What is the main point of the passage?***

   A *the time it takes*
   B *the logical order of item development*
   D *getting a good critic to look at the item*
   F *avoiding trivial questions*
   G *the difficulty of item writing*

Next, we might take a cold, hard look at the language used. The stem, for example: is it explicit enough? We might try a slightly more elaborate wording, such as «What is the main point the author is trying to make in the passage?» If we used that (or even the original), **A** doesn't make much sense – what is "it" that is taking time? Both **A** and **B** would need a verb: 'indicating' for **A** and 'following' for **B** would help us understand things more readily – **D** and **F** both have such verbs. **G** doesn't, so we might insert 'recognising'. What does our item look like now?

> **1.** ***What is the main point the author is trying to make in the passage?***
>
> **A** *indicating the time that item writing takes*
> **B** *following the logical order of item development*
> **D** *getting a good critic to look at the item*
> **F** *avoiding trivial questions*
> **G** *recognising the difficulty of item writing*

However this is still not right – the five options need something to hold them together. We could do this by extending the stem to point out something all the options have in common. The passage is being written **for** item writers, so we might adjust all our wording yet again to include them, in what is called a run-on stem:

> **1.** ***What is the main point the author is trying to make in the passage? Item-writers should***
>
> **A** *allow for the extensive time that item writing takes*
> **B** *follow the logical order of item development*
> **D** *get a good critic to look at the item*
> **F** *avoid trivial questions*
> **G** *recognize the difficulty of item writing*

Now, what's the right answer? **B**, **D** and **F** are all mentioned in the passage so they are reasonably plausible as options. But is "the" answer to be **A** or **G**? Back to the stem: we might ask ourselves 'Is there only one main point in the passage?' Time is certainly a recurring problem – it gets mentioned or implied several times – but beyond that it is the overall difficulty of the whole process which is being indicated, with 'not rushing' being a major factor. If we only need four options for our final item, we could collapse **A** and **G** into a single option, such as: 'preparing items slowly and methodically', and put the word 'difficulty' (or something like it) into the run-on stem. Now we might have:

1. ***What is the main point the author is trying to make in the passage?***
Item-writers will overcome the difficulties of item-writing by

- **B** *following the logical order of item development*
- **D** *getting a good critic to look at the item*
- **F** *avoid trivial questions*
- **G** *preparing items slowly and methodically*

A final check: does the run-on stem agree syntactically with each of the options? No: we've forgotten to put back the participle-form into **F**: it needs to be 'avoid**ing**'. And since the run-on stem forms four different sentences, we'll need full-stops for each option. We need too to adjust the option letters now the item is finished, with the key now being **D**. Also note one other last-minute change: options should all be about the same length in terms of words, so an addition ('all faulty') to the new option **C** has been made from the passage. The item now looks like this:

1. ***What is the main point the author is trying to make in the passage?***
Item-writers will overcome the difficulties of item-writing by

A *following the logical order of item development*

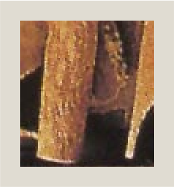
B *getting a good critic to look at the item*

C *avoiding all faulty or trivial questions*

D* *preparing items slowly and methodically*

Figure 15 offers a summary of the processes (and processing) which would enable an item-writer to check the edited or vetted items for the trial testing which now follows.

It should be pointed out that although the discussion in this and the previous section has focused on the multiple-choice format as the basis for the process description, the same sequence of procedures should be followed with regard to other item formats. They too need intense and critical review. about curricular relevance, the wording used, and their position within the pattern of activities which runs through the whole test.



10 Advance preparation for final formatting

Two other activities should also be taking place simultaneously with detailed vetting. One is to begin obtaining publishers' and/or authors' legal permissions for use of any text or other material which is copyright. Field testing of materials should not take place until these have been approved – it would be a waste to test a passage or other stimulus material in the field only to find later that permission was not forthcoming and all the item-writing had been wasted!

The other activity is to design the test paper and answer sheet layout and format, so that the field test version is as close as possible to the real design envisaged for the final test. Layouts and instructions to candidates need to be field-tested as well as items.

Figure 15

THE PROCESS OF MULTIPLE-CHOICE ITEM DEVELOPMENT

1. The stimulus material to which the item refers will contain material which is relevant to the objectives of the test. Considering and responding to the material will be a worthwhile educational experience for the test-taker.
2. The stem and the keyed answer together will represent a meaningful and worthwhile response to a key or central issue in the stimulus material, not a merely peripheral one. Doing the item will not be a merely trivial experience for the test-taker.

...

…

3. Each of the distractors will be plausible: that is, they will represent a possibly relevant view of the matters raised in the stimulus and the stem.

4. The item will be independent. Finding the keyed answer will not depend on successful answering to any other item in the test. No clues as to the keyed answer will be given anywhere else in the test.

5. A successful response to the item-stem will depend on the test-taker understanding a key issue in the stimulus, not eliminating distractors to find a 'best answer' or merely recognising a stated fact.

6. The question stated or implied in the item will be positively worded. Where an important issue in the material unavoidably requires negative wording if it is to be tested at all, this will be in the stem, printed in bold capitals ('NOT'; 'EXCEPT'). No additional negatives will be used in any of the options.

7. The item will contain four or five options for answering, and be laid out in standard form.

8. Each of the options will be roughly the same length. If this is impossible, then two groups of options will be of similar length (e.g. two short and three longer).

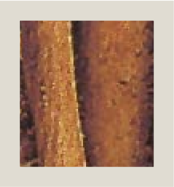
9. The item will have been trial-tested and found to have a facility between 20 and 80 percent.

10. In trial-testing, the keyed answer to the item will have been found to discriminate positively, and distractors to discriminate negatively.

Once again a variety of layouts might be possible, so that later options for choice are kept open. One or (better) two specimen or 'practice' questions should be prepared for the front cover, in case candidates have not seen this style of question before – this is a particularly important issue where multiple-choice items are used.

A first review of keyed answer order should also take place for multiple-choice tests. Two consecutive items may have the same key letter, but not more than two. Also, an approximately equal number of items should be assigned to each option letter (**A, B, C, D**) over the whole test. There is often a tendency amongst item-writers to try to 'bury' keyed answers by assigning them to **C** or **D**. If these two letters are overused, the candidate who guesses using them has a more than 25 percent chance of getting each item right.

Incidentally, the wise item writers don't throw away the successive draft and re-workings of the items they have prepared. For example, the final version of item 1 above may not 'work' in trial testing, and may need to be later revised. The rejected wordings and extra options may make that task easier if that happens.



Field or trial testing

11

In a school setting, a trial test of items on a population which will not do the final test is often impossible – the only students who could do the trial are the ones who have been taught the course, and for whom the test has been written. However in larger programs, field testing is indispensable, for four reasons:

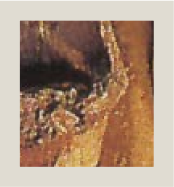
- 'debugging' the whole test, to get rid of errors;
- as a check on the conceptual difficulty of the whole test;
- as a check on test-length and timing;
- as a check on the vocabulary level used in the test stimulus and item material.

The sizes of the trial population needed for various tests will vary considerably – the rule-of-thumb might be 'the largest possible population, given the available resources'.

Another rule-of-thumb might be to choose a population as close in make-up to the one for which the test is ultimately intended – age; level of schooling; gender composition; experience of the matter which forms the basis of the test. Security might play a part in the decision about who to use, too. The more people who see the test, the more likely copies are to disappear. Thus it is essential that only people connected with, or hired by, the test project team, should handle papers and supervise trial sessions.

Ample time should be allowed in the session for all students to complete the test: trial data are needed on the items at the end of the test as well as the early ones. If in early sessions it becomes obvious that only a few students are reaching the end of the test and doing these late items, some special arrangement might need to be made in later sessions for some students to work through the paper 'backwards' as it were.

What if trial testing is not possible? This throws additional responsibility on the test developer to be ultra-meticulous in checking the test before it's used. Every effort must be made to get other professionals to look at draft forms of the test and offer critical comments.



Item analysis

12

There is not room in a paper of this size to canvass and describe all the different options which exist for the analysis of item data. Papers for other modules will explore the matter in some detail. School-based tests will use simple rather than complex strategies to obtain (and indices to express) this information. Larger test programs will probably have the resources to engage in quite lengthy and complex processing.

Broadly speaking, for multiple-choice items in large programs, the analysis consists of statistics which show the facility of each item:

- the percentage of the whole test-taking population which got it right;
- the discrimination index of each item: how well the keyed answer distinguishes between students of high ability and those less able;
- the response level for each item: how many actually attempted it, right or wrong;
- the criterion score on each item: the mean score of all those who did attempt it;
- whether any distractors did not function well: attracted too few candidates, or a preponderance of those of high ability.

However, from the item-editor's point of view, it should be noted that at least five other things are under analysis during this vital stage of the whole development process.

Information is obtained about:

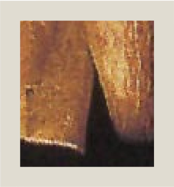
1. the parts of an item, especially in multiple-choice format with its stems, keyed answers and distractors;
2. the integrity or worthwhileness of the item as a whole;
3. the performance of the item as a discrete test element;
4. the performance of an item with regard to other items in the same set or test;
5. the integrity or worthwhileness of the test as a whole.

The distinctions between 2, 3, and 4 (which seem on the surface to be saying much the same thing) are important, especially for a multiple-choice test. Each, for the item editor, might offer a slightly different reason for rejecting or retaining an item in the final test. An item might be worthwhile as a measure of a particular higher order skill (2). It might operate well as a discrete test element, discriminating satisfactorily between the most able and the less able (3). But it might be simply far too hard by comparison with all the other items in the test (4) and deserve exclusion on that ground.

The interpreter of the item analysis sheets is faced with these sorts of trade-off situations all the time. Here's another. Two items may prove to have keyed answers which show excellent powers of discrimination (3). Each item may be well within acceptable boundaries of 'easy to hard' (4). One deals with a higher-order skill and has two 'dead' distractors, the other tests no more than factual knowledge, though each of the distractors contributed well to the overall performance of the item (2). Keep one item only? Which?

Keep both? Exclude both? There are no decision rules to cover adequately the permutations and combinations which occur, or the wide range of choices which becomes necessary. However, a sharp eye should always be kept on the curriculum and the objectives to be tested – the existence of the specification should not be forgotten.

The test length should also be checked. If a large number of trial candidates are found to have omitted one particular item, or failed to complete the full test, then this tells us something about test length. If everyone completes all items, this suggests that the test time might be shortened or the number of items increased, particularly if the trial test supervisors confirm that large numbers of candidates sat around after finishing early in the trial session.



13 Stage Two – editing for publication

When the test has been taken into the field and tried out, and the item analysis has been completed, a second stage of editing then occurs. Using the analysis results, each item is scrutinised and decisions made about rejection or retention. There may be a little judicious re-writing, but this should be limited to minor changes to distractors **only**. If the stem and keyed answer need change, then a new item has resulted, and this would require further trial testing: it would be better to reject the item and use one that did work.

Once every item is clean, making up the final form actually begins. As in so many matters to do with test development, there is a sequence of activity which should be followed, in order to ensure that new bugs don't appear and the test as completely as possible meets the original specification. ***Figure 16*** suggests an appropriate sequence.

Figure 16

EDITING FOR PUBLICATION – A CHECKLIST OF ACTIVITIES

1. Read the item analysis and sort the items (or groups of items) into three piles:
   - a) ready to go;
   - b) needing editing;
   - c) possible rejects.
2. Edit items and establish a final pool of items for the test. Check the omit rate to establish optimum test length.

…

…

3. Check the specification against this pool for the number and qualities of the items available. Reinstate any usable ‘rejects’ if all the objectives are not satisfactorily covered.

4. Assign the items to a tentative order for the whole test and enter the scoring scheme at the end of each section of the test.

5. Check this order for:
   a) order of difficulty (for example, in a set of multiple-choice items, make sure some easy ones occur early, to give the candidate some confidence);
   b) keyed answer order and distribution;
   c) balance and variety of item type.

6. Write or insert appropriate instructions for candidates. Include suggestions for time to be spent on each section.

7. Assign the items to a preliminary paging of the test-paper: some may have to be moved to allow a more satisfactory layout. Allow sufficient space for answering if there is to be no separate answer sheet.

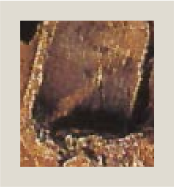
8. Number all separate items consecutively. Use letters to distinguish sections of the test if necessary.

9. Make a mock-up of the test paper (and answer sheet if used). Read the paper through from beginning to end, to check language, numbering (pages and any line-numbers used), all labelling and diagrams, and the layout in general.

10. Photocopy the mock-up and ask a colleague who knows the subject (but who has not seen the test) to ‘do the test’ on this copy.

11. Check the trial completion by the colleague.

12. If possible, put the test aside for a week and then do it yourself as a final check. Only then can you send the copy to the printer!



14 But is it a good test?

Figure 15 suggested some criteria by which we might judge whether a multiple-choice item was a good one. The same sort of evaluation should be made of whole tests, and so (by way of a summary of the preceding sections) here are some suggestions as to criteria we might use.

Hopkins and Antes (1990: 156-7) propose three main areas of review, which they call ‘Balance’, ‘Specificity’ and ‘Objectivity’. They attach to each area a pair of questions which reviewers (and test writers!) should ask themselves to obtain a picture of how good the test is. (The questions have been re-numbered from the original.)

balance

1. Are the items selected for the test representative of the achievement (content and behaviours) which is to be assessed?
2. Are there enough items on the test to adequately sample the content which has been covered and the behaviours as spelled out by the objectives?

specificity

3. Do the test items require knowledge in the content or subject area covered by the test?
4. Can general knowledge be used to respond to the test items?

objectivity

5. Is it clear to the test taker what is expected?
6. Is the correct response definite?

As a test-writer you rely on your test panel to verify positive answers to many of these questions, but you will need to keep them in mind as you go through the developmental process. The specification will help you to meet the demands of ***Question 1***. Distinguishing content and objectives as horizontal and vertical dimensions of your matrix (as in ***Figures 6***, ***6a*** and ***6b***) is an easy way to start. Not every cell has to be filled, but 'Balance' requires that every line and every column has something in it.

There is another point to be made about balance. Do we want our test to consist wholly of items in only one format? We might decide so: say, all essays or all multiple-choice items, for ease or economy of marking. However, where variety of format is a feature, an easy format (say, true/false) should not dominate the whole instrument to the point where we miss out on testing other more complex learning objectives.

The panel (or you own judgement) will help you accommodate to ***Question 2***. Where testing-time is insufficient to cover everything taught (and there rarely **is** enough time), achieving the best 'Balance' means that the sample should consist of the most important content and the most important behaviours.

Question 3 is really about relevance as an element of 'Specificity'. Here the panel will really help, by eradicating trivial items or questions about matters which are merely peripheral and wasting valuable testing time. As we said above with reference to multiple-choice items, the stem and the keyed answer together should represent a meaningful and worthwhile response to a key or central issue in the stimulus material, not a merely peripheral one. Whether multiple-choice or not, doing a test item should never be a merely trivial experience for the test-taker.

In a similar way, panels will help with ***Question 4***, with comments such as: 'It's relevant, but it's not specific to this course – even a child in a nursery knows that!'

Question 5 has a number of applications. Whether in the instructions to candidates, or in the wording of an item itself, the language you use must be clear, concise and explicit. Where you think that definitions will help understanding (but not give away answers), then print the definitions. Where there are rules or word limits to be observed, state them. This will help the candidates do their best, again without giving away answers.

Clarity and explicitness will also help achieve a positive answer to ***Question 6***. This doesn't mean that all correct responses will necessarily be the same: that is obviously an impossibility with responses to an essay task, for example. But it does mean that all candidates should know the boundaries for correct responses – how one might be achieved, even if they themselves can't.

To the overall appraisals above, we might add some more specific indications of good test practice. Most of these will apply to items in any format, whether selected-response or constructed-response.

- ***Provide enough information for the candidates to be able to understand what is expected of them, but not so much that they will become confused.***

  poor
  Write all you know about the Spanish Civil War.

  poor
  Some people say that the Spanish Civil War was really a chance for certain European countries to try out tactics and use new weaponry in advance of the outbreak of a larger European War. If you agree with this view, find some evidence for its truth and write this in about a page. If you don't agree, say why you don't in a piece of writing of about the same length.

  better
  "The Spanish Civil War was a dress rehearsal for World War II." Do you agree? In a short essay of about 500 words, support your view with evidence.

- ***Indicate, by leaving space, or giving some other indication, the amount which is expected in response to the item. If leaving space, be generous: some students have large handwriting.***

poor

Write a few lines about the floating markets of Bangkok.

better

Give two reasons why the floating markets of Bangkok are important to the city's economy.

1 ....
....

2 ....
....

- ***Group items of similar format or content, and where necessary label them. Arrange items in order from simple to more complex.***

poor

1. Find a word in the second paragraph of the second story that means 'briefly'.
2. Write a short character-sketch of one of the players in the first story.
3. Why didn't the captain perform very well?
4. Is it true that the Blue Team won the game?

better **PASSAGE 1**

1. The Blue Team won the game: true or false?
2. Write a five-line character-sketch of one of the football players mentioned in this story.

better **PASSAGE 2**

3. Write the word in Paragraph 2 that means ‘briefly’.

4. In three sentences, explain why the captain did not perform very well.

- ***Make all directions and questions explicit and precise: avoid ambiguity.***

poor

Do you know what the word prescient means?

better

The word “prescient” means ....

. . . .

poor

How did Tina feel in your own words?

better

Write two sentences which show, in your own words, how Tina probably felt during the hold-up.

. . . .

poor

Don’t attempt the essay until after you have written it out first.

better

Draft your piece of writing on the blank page, then write out a good copy on the lined page.

- ***Where particular rules are to be used, or specific units, designate these in the instructions.***

poor
What is the total elapsed time?

better
The journey took .... hours, ... minutes.

. . . .

poor
Since John walked right round the court once, how far did he walk?

better
In walking around the perimeter of the court, John travelled .... meters.

. . . .

poor
Solve the following problems.

better
Solve the following problems. Express your answers correct to two decimal places.

- ***Wherever possible, indicate the criteria to be used in assessment of an extended response item such as an essay or a problem for solution.***

  samples
  Up to five marks may be deducted if you do not show all your working.

  . . . .

  Your writing will be assessed according to the thought and content displayed in the piece, the structure and organisation of the whole, and your expression, style and mechanical accuracy.

- ***Wherever possible, present tasks which are new or unfamiliar to the student, but which remain centrally relevant to the learning which you expect to have been done.***

  (If the stimulus material to which an item or items refer is new to the candidate, nevertheless it should contain material which is relevant to both the objectives and the content of the course. Considering and responding to such new material should be a worthwhile educational experience for the test-taker: it should enhance their knowledge as well as providing fertile material for testing purposes.)

  samples
  Here are two documents relating to the study of Caribbean history you have done this year. Both will be new to you. Read them carefully and answer the questions which follow.

  . . . .

  Listen to the following extract from an early symphony by Mozart {No. 29, KV 201}. Although this work was not set for study this year, the questions which follow will require you to compare the piece with the other symphonies which you did study.

- ***If appropriate, indicate by means of a set of headings the structure or organisation you expect or suggest that students use in completing the task.***

  samples
  In 750-1000 words write a critical review of the various changes in the overall direction of Pablo Picasso's painting style from 1905 to 1945.

  You might mention:

  a. the early work;

  b. the impact of Cubism;

  c. post-Cubist developments.

  . . . .

  Choose any two poems by Schiller which you have studied this year. Write a critical comparison of the two poems, showing what each reflects about Schiller's poetic achievements, any essential differences in language or tone between the two, and your personal assessment of their qualities.

- ***Avoid choice of items as far as possible, though not necessarily choice within items. In the latter case, offer real, not meaningless, alternatives. If they may choose their own structure or form for presentation of their response, tell them they may.***

  samples
  "The recent economic history of Argentina might be well described as the nation staggering from one crisis to another."

  In a piece of writing of 750-1000 words, give your view of the country's economic history since 1960, emphasising one or more of the following:

- international trade;
- domestic monetary policy;
- the impact of the war in the Malvinas (Falkland) Islands.

Write about the man in the photograph above. You may write in any form you like: for example, a story, a letter, or a conversation.

- ***If separate stimulus material is used to prompt candidates to a response, a successful answer to the item should depend on the test-taker understanding a key issue or issues in the stimulus, not eliminating irrelevant material in order to find a 'best answer' or merely recognising a set of stated facts. Test taking should never become a search-and-destroy mission.***

poor
List the names of the characters in the story you have just heard.

better
Name the character in the story you have just heard whose actions contribute most to the build-up of suspense. In your own words, tell what he or she did that was decisive in achieving that build-up.

. . . .

poor
Which of the towns on the map is the third-largest in terms of its population?

better
Town X is the market town for the region shown on the map. One reason is that it is a railway junction. Use the map and its legend to identify three other reasons why it has become so important.

- ***Whenever you can provide helpful advice, print it, even at the expense of a few more words.***

poor

Candidates must do the questions in order.

better

You will give yourself your best chance if you work through the questions in the order of presentation. However, you may need to leave any particularly hard questions to come back to later.

. . . .

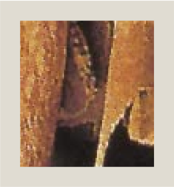
poor

TIME: One hour.

better

TIME: One hour.

Leave a few minutes at the end of the test-time in order to check your work thoroughly.



15 Training scoring teams

The test has been printed and the real population for whom it was intended has completed it. One further stage of the test program remains to which the item-writer is well-placed to make a contribution. This is the training of the team of people who become responsible for assessing the candidates' work. In some test programs, of course, the scoring is done by machine which scans and optically 'reads' the marks made by candidates in the appropriate places on an answer sheet. However, in other programs (particularly those which use extended essay responses as the prime format for response) the scoring will be the result of intense human activity, often by a large team of markers.

This team needs to be fully trained. And the training will need the prior production of a set of criteria to be applied by all team-members to the products of the testing, whether these be essays, painted or sculpted works of art, diagrams, plans or computer-output. To some extent all assessment is criterion-based in this way. Someone exercises a judgement with some criteria or other in mind. The need here is for the assessors to have common or shared criteria, as far as such a thing is possible. The best assessment is also criterion-referenced, in that the criteria not only determine the award of credit by markers, but also underlie the reports of student achievement which result from the assessment process. This is true even if those results have to be norm-referenced or standardised later for other purposes.

It is often possible to use trial test data to achieve the greater commonality of assessment criteria for training purposes. If, as should have happened, the extended response items were pre-tested, samples on each of the finally-chosen topics or questions will be available. In addition, the item constructor will have had a clear idea of what was intended or foreseen as a good or medium or poor response to the item – these foresights should yield a basic list of assessment criteria with which to begin the training sessions.

The governing principle for selection of an assessor should be expertise in and experience of the curriculum or course under review. This will have yielded a personal set of assumptions about learning and criteria for its assessment which the individual brings to the first training session. At that session the task is to get the item-writer's criteria, the assessors' criteria and the trial test samples into a dynamic situation which eventually yields a commonly held view about what should constitute the various levels of performance to be decided for the candidate population.

A simple set of base criteria with which to start any session, in any subject, might be the following:

1. the **thought** displayed by the candidate in preparing the work and the **content** offered in presenting the work;

2. the **structure** and **organisation** of the content of the piece of work, as finally presented for assessment;

3. the **expression, style** and **mechanics** of the finished piece.

The words in bold can be 'translated' to fit just about any criterion set. In language tests, **mechanics** might mean accuracy of spelling and punctuation and **organisation** might yield insight into paragraphing skills. In mathematics tests, **style** might mean the economy or elegance of the thinking which went into proposing a particular solution to a problem. What this set (or a similar one) might do is lead the assessors towards a fruitful discussion of the important criteria for their particular task. It might also help them to avoid a common assessment problem: concentrating on the surface features, to the exclusion of deeper, more important qualities of a student's work.

Figure 17

TRAINING A SCORING TEAM

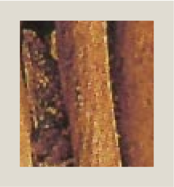
1. Find a venue which enables group discussions to take place.
2. Select only assessors who are expert and experienced in the appropriate subject or learning area.
3. Issue a basic list of criteria and a small set of student work from the trial test to each assessor in advance of the first training session.
4. At the first session, set up small-group interactions to review and revise the base criteria in the light of the test item and the advance reading.
5. In a large-group session, agree on a criteria list (or come as close to agreement as possible!)
6. Each individual applies the criteria to another small set of trial test materials (everyone does the same set).
7. In another small-group interaction, the results of the trial marking are discussed, and the whole group revises the criterion set as necessary.
8. Marking commences.

Figure 17 indicates a suggested sequence for organising and conducting a training program for assessors. The sequence emphasises prior knowledge of shared criteria amongst the assessors. Step 5, for example, can be easily done if a 'bundling' exercise is undertaken, where cards (each containing one possible criterion statement) are sorted into piles of like criteria, redundant cards are eliminated, and lists made of the remaining 'live' ones.

The sequence also emphasises the use of samples of work from 'real life' in assisting the definition of criteria. What it does not do is to tease out the wealth of possible strategies in applying those criteria to these samples. One such strategy would be for an assessor to use all the criteria simultaneously or holistically in coming up with a grade or score for each piece of work. Another would be to assess each piece analytically, element by element (or criterion by criterion), and come up with a set of part-scores which would then be added together. Yet another strategy would be to find and publish a criterion example or piece of work at each score or achievement level, and ask assessors to match each new piece of work to one of these given examples.

Whichever method is chosen (and each has its advantages and disadvantages), there must be provision of ample time for discussion and agreement by those involved. There will also need to be on-going analysis and follow-up of the actual performance of the assessors during the marking period. This on-task monitoring might also be accompanied by further occasions for group discussion during the process. Individual variations in the interpretation of criteria are only to be expected; score discrepancies are also to be expected. The aim is not to eradicate them completely – this would be impossible – but to lessen their eventual impact on the reliability and validity of the whole marking process. (This might mean, in extreme cases, removing poor assessors from the team.) Reliability and validity might also be enhanced in various other ways; for example, reducing the physical and psychological impact of a large workload by staggering the sessions to allow for leisure, and by batching the data in smallish bundles so that a sense of achievement is felt fairly constantly.

In the section above discussing the choice of item-types for test programs, mention was made of the subjectivity involved in the selection of 'right' answers for many multiple-choice questions. Subjectivity, of course, has a large place in the allocation of grades to pieces of student writing or other extensive responses to test items. Despite our best efforts to get assessors to agree, share and use criteria similarly, discrepancies of understanding and of grade allocation will occur. They are unavoidable. The aim is to minimise them. They do not invalidate the assessment process – they merely form part of the 'trade-offs' we might have to make when we can see no other way of assessing complex or extended responses to the learning the students have done.



Further reading 16

Below are citations to six books. Reference to them will take your understanding further. As the publication dates indicate some are quite elderly: nevertheless they are mentioned because this may improve the chances of accessing them through libraries. Page numbers are in **bold**.

Gronlund, N.E. (1976: third edition) **Measurement and Evaluation in Teaching**. New York: Macmillan.

preparing instructional objectives ***28-59***

multiple-choice formats ***188-209***

measuring complex achievements ***210-248***

Hopkins, C.D., and Antes, R.L. (1990) **Classroom Measurement and Evaluation.** Itasca: Peacock.

principles for writing good items ***210-18; 243; 265***

Lien, A.J. (1976: third edition) **Measurement and Evaluation of Learning**. Dubuque: Wm C. Brown Company.

characteristics of good instruments ***78-91***

construction and use of classroom tests (principles and steps) ***194-242***

Lindquist, E.F. (ed.) (1955) **Educational Measurement**. Washington, D.C.: American Council on Education.

summary of test planning ***175-184***

ideas for test items ***190-3***

formats ***193-212***

writing items ***213-227***

Mehrens, W.A. and Lehmann, I.J. (1984: third edition) **Measurement and Evaluation in Education and Psychology**. New York: Holt, Rinehart and Winston.

objective and extended response tests compared ***75-84***

general considerations in item writing ***84-89***

essay testing ***96-112***

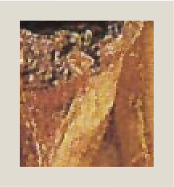
assembling, reproducing, administering and scoring ***177-202***

Miller, H.G., Williams, R.G., and Haladyna, T.M. (1978) **Beyond Facts: Objective Ways to Measure Thinking**. Englewood Cliffs: Educational Technology Publications.

multiple-choice items beyond the factual level ***33-56***

measuring predictive behaviours ***73-96***

measuring evaluative behaviours ***97-116***



Exercises

17

1. Select a curriculum area in which you are an expert or which is important for some professional reason.

   Using ***Figures 1*** and **2** as a guide, write a brief specification for a test or test program in this area. One page will do – **omit Step 8, the matrix.**

   In a small-group interaction (2-4 persons), review the specifications prepared by individuals.

2. Using ***Figures 3-6*** as a guide, use your specification, as revised after from **Exercise 1**, to develop a full matrix for the test or test program.

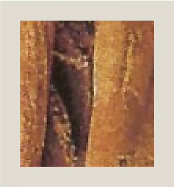
   Include a set of multiple-choice items as one of the items in the matrix.

   Repeat the small-group interaction undertaken in **Exercise 1**.

3. Use your revised matrix from **Exercise 2** to draft a small set (3 or 4) of multiple-choice items appropriate to the specification you have prepared.

   If you cannot find a piece of stimulus material, you may need to prepare or write a piece yourself.

   Set up a small panel and review the items.



18 Glossary of terms

answer sheet or booklet

a piece of test stationery separate from the question booklet, on which candidates record personal details and the answers to the test items.

constructed response item

a test item which allows or requires candidates to produce individual responses rather than merely select from a list of given options.

criterion score

the mean facility of an item taking into account only the performance of those candidates who actually attempted to answer: can also be calculated for whole tests or groups of items, using the same rule.

discrimination

the ability of an option to distinguish between those groups of candidates who had greater and lesser ability as indicated by their performance on the whole test. The indices used are usually expressed as positive or negative fractions of 1.0 and can be derived using a number of different formulae (e.g. point bi-serial; phi coefficients, etc.)

distractor

in a multiple-choice item, an option for choice which is not the keyed answer, but which has been written in such a way as to distract weaker candidates from selecting that key.

editing

preparation of refined versions of tests or items after other key stages in the development process, such as panelling or trial-testing. It is usually performed by the original item-writer(s).

extended response item

any test item which requires the production of a personal response by the candidate which is longer than a sentence or two.

facility

the index obtained by a multiple-choice item during testing which indicates the number of candidates who got it right: expressed as a percentage of the total number of candidates who sat the test. The index for an item to be used in a final test should always lie within the range 20-80 percent.

final form

the test instrument after it has been trial-tested, analysed, edited and prepared for publication.

instructions

- **to candidates**

information printed on the question paper or answer sheet which candidates need to be able to complete the test satisfactorily, but which does not actually form part of the stimulus material or questions: in some cases, these may also be read aloud by a supervisor.

- **to supervisors**

information provided for test supervisors or invigilators on how to conduct the test session correctly: includes a script of any instructions to candidates which are required to be read aloud.

item

an individual task which forms one component of a test instrument: usually applied in the context of a multiple-choice test to indicate a single question, but can be used more broadly.

key, keyed answer

in a multiple-choice item, the option which is designated to be correct, and for which a score is awarded.

key order

in a multiple-choice test, the sequence of letters attached to keyed answers, as in 1 **D**, 2 **B**, 3 **C**, etc. The same key letter should be used for no more than two consecutive items: viz. 3 **C**, 4 **C**, 5 **A**.

moderation

sometimes used to describe the process whereby expert panels meet to discuss and offer critical comment on test materials.

multiple-choice

an item format whereby a restricted number of optional responses is offered to candidates, from which they must select one as their answer.

omit rate

a tally of the number of candidates who did not answer a test item: especially important in estimating the performance of trial test candidates in the later items of the test, with a view to establishing an acceptable test length.

option

in a multiple-choice item, a set of responses (usually four or five) from which the candidates select their answer.

panel, panelling

a group of experts called together to discuss and evaluate draft items proposed for use in a test instrument.

question book(let)

a printed test instrument which contains instructions, stimulus material and test items for students to work through during the test session. Answers may be recorded in this book, or on a separate answer sheet.

selected response item

any item which prints a limited range of options from which candidates must select their answers.

specification

a document which specifies in some detail the nature and composition of a test program or instrument: sometimes called a 'blueprint'.

stem

in a multiple-choice item, the sentence(s) or part-sentence which indicate the testing point or question, which candidates use to select their answer from the options which follow.

stimulus

- **directive**

any information in a test which candidates need to understand the specific task which they are being asked to perform (e.g. the stem of a multiple-choice item, or a detailed essay topic).

- **instructive**

any information printed in a question booklet which candidates are expected to refer to when answering the specific questions which relate to it.

trial form

the test instrument after it has been developed, panelled and edited ready for administration to a trial population in the field.

vetting

the process of editing and arriving at a draft test form, using the discussions and evaluations of draft items by a panel as a guide.

Since 1992 UNESCO's International Institute for Educational Planning (IIEP) has been working with Ministries of Education in Southern and Eastern Africa in order to undertake integrated research and training activities that will expand opportunities for educational planners to gain the technical skills required for monitoring and evaluating the quality of basic education, and to generate information that can be used by decision-makers to plan and improve the quality of education. These activities have been conducted under the auspices of the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ).

Fifteen Ministries of Education are members of the SACMEQ Consortium: Botswana, Kenya, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania (Mainland), Tanzania (Zanzibar), Uganda, Zambia, and Zimbabwe.

SACMEQ is officially registered as an Intergovernmental Organization and is governed by the SACMEQ Assembly of Ministers of Education.

In 2004 SACMEQ was awarded the prestigious Jan Amos Comenius Medal in recognition of its "outstanding achievements in the field of educational research, capacity building, and innovation".

These modules were prepared by IIEP staff and consultants to be used in training workshops presented for the National Research Coordinators who are responsible for SACMEQ's educational policy research programme. All modules may be found on two Internet Websites: **http://www.sacmeq.org** and **http://www.unesco.org/iiep**.