Module

7

John Izard

Trial testing and item analysis in test construction
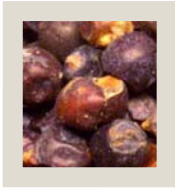
UNESCO

Content

Introduction

I

Assessment involves selecting evidence from which inferences can be made about current status in a learning sequence. The tasks that are chosen to provide that evidence have to be effective in distinguishing between those who have the required knowledge and those who do not. Trial testing (sometimes called pilot testing) involves giving a test under specified conditions to a group of candidates similar to those who will use the final test. Subsequent analysis of the data from the trials examines the extent to which the assessment tasks performed as expected under practical conditions. The shading in *Figure 1* below, indicates the position of trial testing and item analysis in the overall test construction process.

In order to assess the capacity of each question or task to distinguish between those who know and those who do not, the trial group of candidates should possess a range of knowledge from those with good knowledge to those lacking it. Typically one does not have definitive evidence on this (and if we did have, we probably would not need to construct the trial test). Therefore we need to depend on teacher advice and our experience to choose suitable trial test candidates. Note that this applies to both criterion-referenced and norm-referenced tests. In the former case we need some candidates likely to meet the criterion and some who do not. In the latter case we need some candidates who score well relative to their peers, some who score around the average relative to their peers, and some who score poorly relative to their peers.

Figure 1.   **Trial and analysis in the context of test construction**

Decision to gather evidence

↓

Decision to allocate resources

↓

Content analysis and test blue print

↓

Item writing ←

↓

Item review 1

↓

Planning item scoring

↓

Production of trial tests

↓

Trials

↓

Item review 2

↓

**Amendment
(revise/replace/discard)**

↓

More items needed? ⟶ **Yes**

↓

No

↓

Assembly of final tests
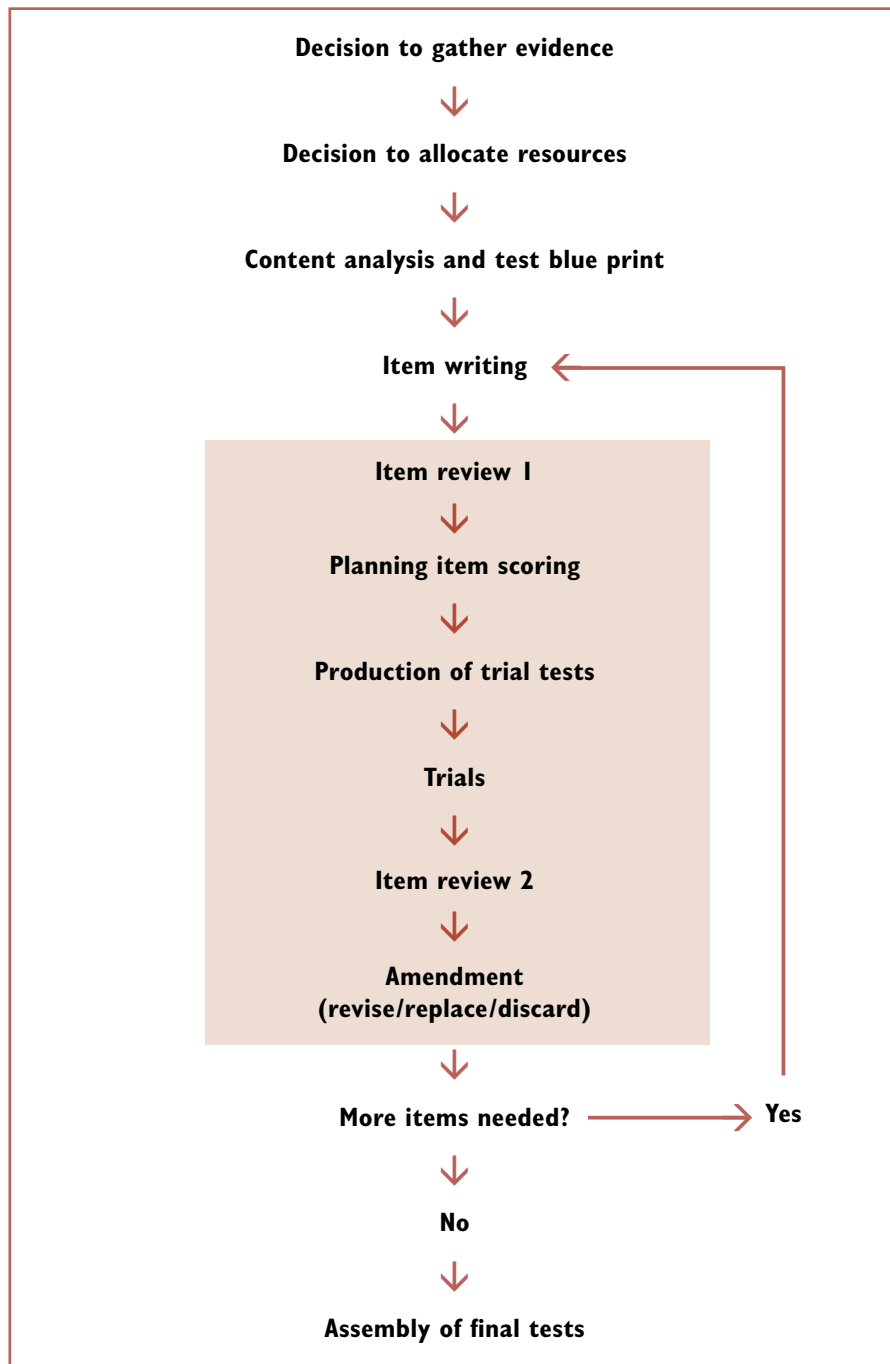
Preparing for trial testing ②

Before undertaking a trial test project, we need to make some important checks. Trial testing uses time and resources so we must be sure that the proposed trial test is as sound as possible so that time and resources are not wasted. The team preparing the trial tests should have prepared a content analysis and test blueprint. A panel should review the trial test in terms of the content analysis and test blueprint to make sure that the trial test meets the intended test specifications. It is also necessary to review each test item before trial testing commences.

Content analysis

A content analysis provides a summary of the intentions of the curriculum expressed in content terms. Which content is supposed to be covered in the curriculum? Are there significant sections of this content? Are there significant subdivisions within any of the sections? Which of these content areas should a representative test include?

Test blueprint

A test blueprint is a specification of what the test should cover rather than a description of what the curriculum covers. A test blueprint should include the test title, the fundamental purpose of the test, the aspects of the curriculum covered by the test, an

indication of the students for whom the test will be used, the types of task that will be used in the test (and how these tasks will fit in with other relevant evidence to be collected), the uses to be made of the evidence provided by the test, the conditions under which the test will be given (time, place, who will administer the test, who will score the responses, how the accuracy of scoring will be checked, whether students will be able to consult books (or use calculators) while attempting the test, any precautions to ensure that the responses are only the work of the student attempting the test, and the balance of the questions.

Comparing the test blueprint with the content analysis of the curriculum should show that the test is a reasonably representative sample of what the curriculum is about (at least as far as content is concerned). Test blueprints may include other dimensions too. For example, the blueprint may indicate the desired balance between factual recall questions and questions which require interpretation or application to a particular context. Or the blueprint may show the desired balance between different item formats (constructed responses as compared with recognition responses). When the test blueprint has several dimensions it is possible to see how the evidence to be collected combines these dimensions with other sources of information by means of a grid (or series of grids), and how account is to be taken of the importance of that evidence.

Item review

Why should the proposed trial test be reviewed before trial? The choice of what to assess, the strategies of assessment, and the modes of reporting depend upon the intentions of the curriculum, the importance of different parts of the curriculum, and the audiences needing the information that assessment provides. If we do not select an appropriate sample of evidence, then the conclusions we draw will be suspect, regardless of how accurately we make the assessments. Tasks chosen have to be representative so that:

- dependable inferences can be made about both the tasks chosen for assessment and the tasks not chosen;

- all important parts of the curriculum are addressed;

- achievement over a range is assessed (not just the presumed narrow band where a particular selection decision might be required on a single occasion).

The review panel has the responsibility of ensuring that the assessment tasks are appropriate, representative, and extensive. For example, the range of complexity of tasks should be at least as wide as the expected range of achievement for the students being assessed if evidence of learning is required about all students. Just as a team of item writers can produce a better range of items to consider for trial, a team of item critics (including item writers – they need the feedback) can provide better and more constructive comments on proposed trial items. Item review without the benefit of interaction with colleagues is generally inefficient and tends to be too idiosyncratic, representing only one person's limited view of the topic to be assessed. The review of assessment tasks by a review panel is essential before trial testing commences. Sometimes the item seems clear to the person who wrote it – but the item may not necessarily be clear to others. The review panel will ask questions like:

- Is the task clear in each item? Is it likely that the person attempting an item will know what is expected?

- Are the items expressed in the simplest possible language?

- Is each item a fair item for assessment at this level of education? Is the wording appropriate to the level of education where the item will be used?

- Are there unintended clues to the correct answer?

- Is the format reasonably consistent so that students know what is required from item to item?

- Is there a single, clearly correct (or best) answer for each item?

- Is the type of item appropriate to the information required?

- Are there statements in the items which are likely to offend?

- Is there content which reflects bias on cultural or other grounds?

- Are the items representative of the behaviours to be assessed?

- Are there enough items to provide an adequate coverage of the behaviours to be assessed?

This part of the review before the items are tried should help avoid tasks which are expressed in language too complex for the idea being tested, and/or contain redundant words, multiple negatives, and distracters which are not plausible. The review should also identify items with no correct (or best) answer and items with multiple correct answers. Such items may be discarded or re-written. Only good items should be used in a trial test. (The subsequent item analysis helps choose the items with the best statistical properties from the items that were good enough for trial).

Other review issues

Some tests provide items so that candidates can do under supervision in order to be sure that they know how to record their responses. Some candidates will have had more experience in attempting tests. In situations where tests are to be used for

selection purposes it may be necessary to provide more detailed information about the test. For example, an information leaflet about a test can be useful in reducing test anxiety, and in avoiding some of the unsavoury effects of test coaching (by providing a simple form of coaching for all candidates rather than advantaging those who can afford to pay private tutors). Here are some of the important questions.

Will the students be told how the items are to be scored? Will they be told the relative importance of each item? Will they be given advice on how to do their best on the test?

Will there be practice items? Do students need advice on how they are to record their responses? If practice items are to be used for this purpose, what types of response do they cover? How many practice items will be necessary?

Will there be a separate answer sheet? Recording responses on a separate answer sheet may allow a test booklet to be used again. If there is to be a separate answer sheet, have plans been made to recycle the test question booklets? (If so, resources may be required to have each page checked very carefully to make sure that there are no marks left by previous candidates who used the test). Will this use of a separate sheet add to the time given for the trial test? What information should be requested in addition to the actual responses to the items? (This might include student name, school, year level, sex, age, etc.).

Has the scoring been arranged for efficient scoring (or coding) of responses? Are distracters for multiple-choice tests shown as capital letters (less confusing to score than lower case letters)? One long column of answers is generally easier to score by hand than several short columns.

How much time will students have to do the actual test? What time will be set aside to give instructions to those students attempting the test? Will the final number of items be too large for the test to be given in a single session? Will there be a break between testing sessions when there is more than one session?

What type of score key will be used? Complex scoring has to be done by experienced scorers, and they usually write a code for the mark next to the test answer or on a separate coding sheet. Multiple-choice items are usually coded by number or letter and the scoring is done by a test analysis computer programme.

What test administration information will be given to those who are giving the trial test to students? Will the students be told that the results will be returned to them? Are the tests to be treated as secure tests (with no copies left behind in the venue where the test is administered)?

Review of trial test, presentation and layout

Some very practical working rules should be adopted. The front page should explain briefly which group has prepared the test, give the purpose of the test, and give instructions to the candidate about:

- the number of items;

- the time available for them to attempt the test;

- how they are to show their answers (whether on the test paper, or on a separate answer sheet); and

- what to do if they change their mind about an answer and wish to alter it.

The options in multiple-choice items should be arranged in some logical order (for example, from the smallest to the largest). The items should be placed in order from the easiest to the most difficult (to encourage candidates to continue the test).

The layout of items should avoid patterns in the correct answers such as three or more of the same letter in a row, or other patterns like ABCD or ABABAB (which might lead to 'correct' responses for the 'wrong' reasons).

Any materials required during the administration of the trial test should be listed so the candidates know, explicitly, what they should have for the testing session. Candidates must be informed that all test materials must be returned to the testing supervisor.

If the test is to be expendable, there must be space for the trial candidate's name, location or department (so that scores can be returned if appropriate, and to give those conducting the trial test, information about the diversity of the trial sample). If the candidate is not to write on the test but is to write on a separate answer sheet, that answer sheet must have the candidate's personal identification details instead of the test itself.

When several trial tests are being given at the same time (and this is usually the case) it is important to have some visible distinguishing mark on the front of each version of the test. Then the test supervisor can see at a glance that the tests have been alternated. If distinguishing marks for each version cannot be used, then a different colour of cover page for each version is essential.

The trial test pages should not be sent for reproduction of copies until the whole team is satisfied that all possible errors have been found *and corrected*. All corrections must be checked carefully to be sure that everything is correct! [Experience has shown that sometimes a person making the corrections may think it better to

retype a page rather than make the changes. If only the 'corrections' are checked, a (new) mistake that may have been introduced will not be detected.]

When those responsible for constructing the questions, assembling the trial test, and reviewing it, are satisfied that each question meets the criteria for relevant, reasonable, valid and fair items, the test is ready for trial. Only items which have survived this review should be subjected to trial with candidates like those who will eventually attempt the final version of the test.
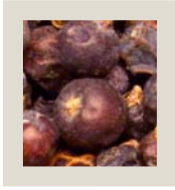
Planning the trial testing (3)

Empirical trial testing provides an opportunity to identify questionable items which have not been recognised in the process of item writing and review. At the same time, the test administration instructions are able to be refined to ensure that the tasks presented in the test are as identical as possible for each candidate. (If the test administration instructions vary then some candidates may be advantaged over others on a basis unrelated to the required knowledge which is being assessed by the test).

The trial testing will:

- establish the difficulty of each item;

- identify distracters which do not appear plausible;

- assist in determining the precision of the test and suggest the number of test items for the final test;

- establish the contribution of each item to the discrimination between candidates who achieve at a high level and those who do not;

- check the adequacy of the administration instructions including the function of any practice items and the time required for most students to complete the test;

- identify misconceptions held by the students through analysis of student responses and, where possible, the questioning of some students as to their reasons for making these responses.
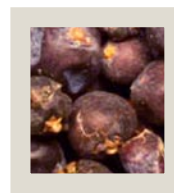
(4) Choosing a sample of candidates for the test trials

The size of the trial testing group for each trial test should be around 150 to 250 persons, covering a wide range of ability, geographic dispersion, and should be roughly representative of the various groups likely to attempt the final versions of the tests. It is usual to try to have approximately equal numbers of male and female candidates for the trials, with males and females each meeting the target group requirements.

The target audience for the final form of the test should guide the selection of a trial sample. If the target audience is to be a whole nation or region within a nation, then the sample should approximate the urban/rural, sizes and types of school and age level mix in the target audience. This type of sample is called a judgment sample, because we depend on experience to choose a sufficiently varied sample for trial purposes. The choice of sample also has to consider two competing issues: the costs of undertaking the trial testing and the need to restrict the influence of particular schools. The more schools are involved in the trial testing and the more diverse their location, the greater the travel and accommodation costs. The smaller the number of schools the greater the influence of a single school on the results.

Judgment samples often have to take into account the following categories of schools:

- Government/Private

- Co-educational/Boys/Girls

- Major Urban/Minor Urban/Outer Urban/Rural

- Primary/Secondary/Vocational

- Selective/Non-selective

As a consequence, those choosing the judgment sample need to know how many students (at least approximately) there are in each category so that the judgment sample can approximate the national or regional target audience for the final form of the proposed test. In some nations and regions, test security concerns result in trial testing being conducted in another nation or region.

Preparing the codebook

When a trial test is prepared it is necessary to document where an item appears on the test, which area of content and which skills are being assessed, the name assigned to the item (if one is assigned), the number of options, the code used for missing data, any coding values for particular responses, and any notes that provide necessary information about the item.

The document which is a collation of such item information and associated trial sample description is known as a codebook. (This label is also applied to a machine readable file with the same information).

The test specification grid (part of the test blueprint) will help in the preparation of this documentation (see *Figure 2*). For example, the content and skill objectives of a basic statistics test are shown in the grid below.

Figure 2. Test specification grid for a *basic statistics* test

Content	Objectives			
	Recall of facts	Computational skills	Understanding	Total
Frequency distributions	2 items	-	4 items	6
Means	2 items	4 items	2 items	8
Variances	2 items	4 items	2 items	8
Correlation	4 items	4 items	12 items	20
Relative standing	4 items	-	8 items	12
Total	**14**	**12**	**28**	**54**

Figure 3. Item specification grid for a *basic statistics* test

Content	Objectives			
	Recall of facts	Computational skills	Understanding	Total
Frequency distributions	items 1,4	-	items 6, 9, 12, 16	6
Means	items 2,7	items 8, 10, 19, 22	items 13, 18	8
Variances	items 3,5	items 11, 15, 20, 24	items 14, 17	8
Correlation	items 21, 25, 32,36	items 23, 27 35,41	items 28, 31, 34, 37, 39, 43, 45, 47, 49, 50, 52, 54	20
Relative standing	items 30, 42, 44, 53	-	items 26, 29, 33, 38, 40, 46, 48,51	12
Total	**14**	**12**	**28**	**54**

The code book should show which items appear in each cell. One way of doing this is to show the specification grid with the item numbers in place and show the score key below the grid (see *Figures 3* and *4*).

Figure 4.   Codebook details for Figure 3

Basic Statistics correct answers

***   .... 1 .... 2................ 3................ 4................ 5.......

***   12345678901234567890123456789012345678901234567890 1234

key   1532423252543514151523131532345414524521332454321123 15

Missing data: coded as 6. Multiple answers: coded as 7.

What to consider in arranging for a test to be given

Experience has shown that those involved in the construction of test items should also be involved in the trial of those items. Test constructors need first-hand feedback on the qualities of their test items; students attempting a new test can help in providing that direct feedback.

The institutions which have agreed to allow trial tests to occur on their premises should be contacted in advance. They should be informed of the number of candidates that are required from that institution, whether they be from different year levels (if in training) or from different employment levels (or equivalent) if already working. [It is usually wise to have as diverse a group as possible, particularly in the context of testing for selection purposes.]

There may need to be a preliminary visit to each institution to establish whether the trial tests will be done in one large room, and/

or several smaller rooms (such as classrooms). Each testing room needs a test supervisor! The supervisor introduces the test to the trial candidates, explains any practice items, and has to ensure that candidates have the correct amount of time allowed to attempt the test, that any last minute queries are answered (such as informing trainees that the results of this trial testing are to be used to validate the questions and will not have any effect on their current course work), and gather all test materials before candidates leave the room.

Test materials should be sorted into bundles before entering the testing room so that different trial test forms can be alternated. All bundles should have three or four spare copies of each trial form in case of printing or collating errors. No candidate should be sitting beside another candidate doing the same form of the trial test. Candidates may sit in front of or behind other candidates attempting the same form of the trial test, unless the test is being done in a sloping-floored lecture theatre which permits one person to see the paper of the person in front.

Preparing test administration instructions

A sample set of administration instructions is given in Panel 1. These may be used as a model for writing such instructions. Other issues to consider, include the provision of practice examples (particularly if the format of the test, is expected to be unfamiliar to those students in the trial group), provision of pens or pencils in two colours so that after a given period of time candidates can be instructed to change to the other colour (particularly if the variation in the number of items completed in that time needs to be determined), and advice to the test administrator on alternating the versions of the trial tests so that adjacent candidates are attempting different versions.

Panel 1. **A sample set of test administration instructions**

Instructions for Administration

These instructions assume the candidates can read. The tester should have a stopwatch, or a digital watch showing minutes and seconds or a clock with a sweep-second hand. Make sure each candidate has a pen, ballpoint pen, or pencil. All other materials should be put away before the test is started.

Give each candidate a copy of the test, drawing attention to the instruction on the cover.

do not open this book or write anything until you are told.

Instruct the candidates to complete the information on the front cover of the test, assisting as necessary. Check that each candidate has completed the information correctly. (Year of birth should not be *this* year; number of months in the age should be *11 or less*; first name should be shown *in full* rather than as an initial.) Ensure that the test booklet remains closed. Read these instructions (which are shown on the cover of the test), asking candidates to follow while you read.

Say:

> **Work out the answer to each question in your head if you can. You can use the margins for calculations if you need to. You will receive one mark for each correct answer.**

> **Work as quickly and accurately as you can so that you get as many questions right as possible. You are not expected to do all of the questions. If you cannot do a question do not waste time. Go on to the next question. If there is time go back to the questions you left out.**

> **Write your answer on the line next to the question. If you change an answer, make sure that your new answer can be read easily.**

Check that everybody is ready to start the test. Tell candidates that they have **30** (thirty) minutes from the time they are told to start to answer the questions. Note the time and tell candidates to turn the page and start question one.

After **30** (thirty) minutes tell candidates to stop work and to close their booklets.

Collect the tests, making sure that there is one test for each candidate, and thank the candidates for their efforts.

---
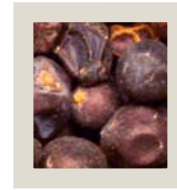
5 Conducting the actual trial testing

Finding the appropriate place where the testing is to be held in an institution unfamiliar to the supervisor, may mean that the supervisor has to arrive at that institution well in advance of the planned testing time. Each testing room must have a supervisor. The supervisor for each room has to have a complete set of testing materials (since testing rooms may not be adjacent or even in the same building). It is more efficient for all the supervisors to start the testing at the same time, rather than go from room to room starting the testing on a staggered timetable.

The supervisor makes sure that all candidates are seated, introduces him/herself, explains briefly what will happen in the testing session and answers queries, distributes the test and associated papers to each person according to the agreed plan, and ensures that each candidate has a fair chance of completing the trial test without interruption. The supervisor must enforce the test time limits so that candidates in each testing room have essentially the same time to attempt the items.

After the test has been attempted, it is usual for all test materials to be placed in an envelope (or several if need be) with identification information about the trial group and the location where the tests were completed. If there is time, the trial tests can be sorted into the different test forms before being placed in the envelope. The envelope should be sealed. The test supervisor for a room is responsible for ensuring that *all* the test papers (used and unused) are returned to those who will process the information.
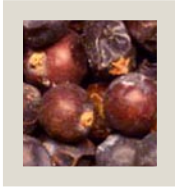
Processing test responses after a trial testing session 6

When the trial tests arrive back at the trial testing office they should still be in their sealed envelopes or packages. Only one envelope is opened at a time, as it is important to know the source of every test paper. When an envelope is opened, the trial tests are sorted into stacks according to the test version.

Identification numbers are assigned to the tests in the package, and written clearly on the tests. For example, some digits of the numbers may be assigned according to the institution that provided the trial test candidates. The first institution numbers may be prefixed with '1', the second with '2', and so on. It is important to check whether the intended trial group became the actual trial group. If the actual trial group differs substantially from the intended group, interpretation of trial data will be made more difficult because the group will be less representative. For example, trial groups should have both urban and country representation. Data for country trials may be slower in returning for processing. If country data are not included, the analyses will not be representative of country and urban groups. That is, there will be no evidence of the usefulness of the items for distinguishing between more able and less able respondents in country areas.

Scoring procedures

- **Multiple-choice**

Multiple-choice items present a task and provide a number of options from which the candidate has to choose. The candidate's task is to identify the correct or the best alternative. Judgments of experts are needed to establish which option is the best (or correct) answer for each item. Once these correct answers have been decided, the score key can then be used by clerical staff or incorporated in machine scoring. Scoring becomes a mechanical task and many test analysis software packages for personal computers can score and analyze test data in a single processing run. The correct score key is crucial. Errors in score keys create interpretation problems. In such a case the total score obtained is not the best measure of what the test is measuring, items which are sound are queried, and candidates do not receive appropriate credit for their achievements. Further, since test software packages require the score key in the files to be kept on disk, there is a need for the computer containing score keys, to be kept in a secure place and for there to be restrictions on access to the computer.

- **Constructed response**

There are potential difficulties in scoring prose, oral, drawn and manipulative responses. An expert judge is required because each response requires interpretation to be scored. Judges vary in their expertise, vary over time in the way they score responses (due to fatigue, difficulty in making an objective judgment without being influenced by the previous candidate's response, or by giving varying credit for some correct responses over other correct responses), and also vary in the notice they take of handwriting, neatness, grammatical usage and spelling.

One technique for avoiding or minimizing such problems is to train a team of scorers. Such training often involves a discussion of the

key issues that have to be identified by a candidate. The scorers should then apply what they have learned by scoring the same batch of anonymous real samples of responses. It is important to have a range of real samples. (The training is to ensure that scorers can tell the difference between high quality, medium quality, and low quality answers and assign marks so that the higher quality answers will get better scores than the medium quality answers, and medium quality answers in turn will get better scores than low quality answers). These results are then compared (perhaps graphically) and discussed. It is not expected that identical results will be obtained by each scorer. Rather, the aim is to improve the agreement between scorers about the quality of each response. We expect that there should be greater agreement between the scorers where the responses are widely separated in quality. Making more subtle distinctions, consistently, requires more skill. To achieve consistency, each paper (or sample of papers) should be remarked without knowledge of the other assessment. If large differences occur in such a case, training is required until the interpretations tend to agree. Members of the scoring team may differ in the importance they place on various aspects of a task and fairness to all candidates requires consistency of assessment within each aspect. Even when team members agree in the rank ordering of responses, the marks awarded may differ because some team members are lenient while others are more stringent.

A more subtle difference occurs when some judges see more "shades of grey" or see fewer such gradations (as in the tendency to award full-marks or no marks). Scorers should make use of similar ranges of the scale.

Short-answer items may require a candidate to recall knowledge rather than recognise it (to produce an answer rather than make a choice of an answer) or may be restricted to recognition. The former may be something like miniature essays (or the oral or drawn equivalent), or may require a word or phrase to be inserted (as in

cloze procedure or fill-the-gap). Recognition tasks may require a key element of a drawing/photograph/diagram/prose passage to be identified, as in the case of a proof-reading test of spelling, or choosing the part of a diagram or poster which has a safety message.

Scoring short responses (whether production or recognition in format) has some of the difficulties of scoring more extended responses but it is generally easier for judges to be consistent, if only because the amount of information to be considered is smaller and likely to be less complex. However item analysis is still a necessary part of the scoring arrangements for short responses as a quality assurance process.

Scoring trial papers

If the test needs to be scored by expert judges before analysis, this scoring is done next. If there are essay-type items, two approaches can be used. The first requires the marker to obtain scores on distinct aspects such as completeness of evidence, logical organization, and effectiveness of explanation. This analytic method may be time consuming and errors may creep in if the marks awarded to each aspect are not added correctly. The second approach requires a general unanalyzed impression. This approach depends upon rapid global judgments leading to sorting of samples of work into a number of groups. For example, the first sorting might set up three groups: poor, average, and good.

When this sorting has been finished the essays in each group are checked quickly to ensure that they are in the correct group. The essays in each group are then sorted into two further groups and checked again. For both approaches essays should be assessed as anonymously as possible.

Regardless of the approach that is chosen, it is necessary to decide in advance what qualities are to be considered in judging the adequacy of the answer. If more than one distinct quality is required in an essay, separate assessments are needed. It may be useful to prepare an answer guide in advance, showing what points should be covered. Where there are several essays in an examination paper it is good practice to mark the first essay all the way through the stack of test papers. Then shuffle or rearrange the papers before starting to score the next essay. Repeat this process after each essay has been marked.

When all items have been marked, the scores are entered into a computer file. If the test is multiple-choice in format, the responses may be entered into a computer file directly. (The scoring of the correct answers is done by the test analysis computer programme). The next envelope of tests is not opened until the processing of the first package has been assigned. This is to ensure that tests do not get interchanged between packages. [*Sending the wrong results to an institution reflects very badly on those in charge of the test trials and analysis.*] Data entry can be done in parallel provided that each package is the responsibility of one person (who works on that package until all work on the tests it contains is completed). The tests are then returned to their package until the analysis has been completed, and the wrapping is annotated to show which range of candidate numbers is in the envelope and the tests for which the data have been entered. (If a query arises in an analysis, the actual test items for that candidate must be accessed quickly and efficiently).

The analysis can be done as soon as all of those particular trial tests have been processed and the resulting data files have been combined. (Remember to check that blank or duplicate lines have been taken out of the combined data file. Leaving such lines in may lead to spurious discrimination and difficulty indices).
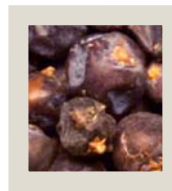
(7) Aknowledging co-operation

Empirical trial is the only satisfactory method of finding the difficulty of a test item for a particular group. Without the co-operation of those managing the trial sites, and the trial group of candidates, this information could not be obtained.

If the results from the trial tests are to be sent back to the institutions which co-operated in the trials, the results should be accompanied by some advice on interpretation. This advice should include something like this.

> *These results are from the trial testing conducted on <date>. Since these results are based on trial tests some caution should be exercised in interpretation of the results. For example, the trial tests administered may have differed in difficulty so the same score on each test may not represent equivalent achievement.*

Appropriate thanks for co-operation should also be given.
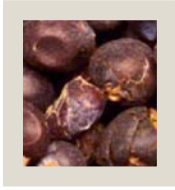
Analysis in terms of candidate responses

When candidate responses are available for analysis, trial test items can be considered in terms of their psychometric properties. Although this sounds very technical and specialized, the ideas behind such analyses are relatively simple. We expect a test to measure the skills that we want to measure. Each item should contribute to identifying the high quality candidates. We can see which items are consistent with the test as a whole. In effect, we are asking whether an item identifies the able candidates as well as can be achieved by using the scores on the test as a whole.

Two main indices are obtained from a traditional analysis of student responses to test items. These are an index of *item difficulty (or facility)* and an index of *item discrimination*. Also, further information can be gained from an analysis of the choices in a multiple-choice context. Many software packages provide summary statistics such as the mean, standard deviation, reliability or internal consistency index, and a frequency distribution of scores, for the test as a whole as well.

- **Item difficulty**

Empirical trial of a test is the only satisfactory method of finding the difficulty of a test item for a particular group. The index of difficulty, which is reported for a particular test administered to a particular group, is a function of the skills required by the questions *and* the skills achieved by those attempting the test. *Item*

facility is the opposite of *item difficulty*. As the difficulty increases, fewer candidates are able to give the correct response; as the facility increases, more candidates are able to give the correct response. In general, between 90 per cent and 100 per cent of students should complete all items unless the purpose is to test speed itself, as in the case of a *speed of reading* test.

- **Item discrimination**

Traditional test analysis considers the extent to which a single item distinguishes between able and less able candidates in a similar way to the test as a whole. Items which are not consistent with the other items in the way in which they distinguish between able and less able candidates (*as measured by this test*) are considered for deletion, amendment, or placement on a different test. Modern test analysis techniques consider other factors as well. (These are discussed later).

For a test of many items, it is common practice to assume that the total score on the trial test is a reasonable estimate of achievement for that type of test. Criterion groups may be selected on the basis of total score (if that type of analysis is being done). When such an assumption is made, we expect candidates with high total scores to have high achievement and candidates with low total scores to have low achievement.

The procedure investigates how each item distinguishes between candidates with knowledge and skill, and those lacking such knowledge and skills. Choosing items with an acceptable discrimination index will tend to provide a new version of the test with greater homogeneity. [However this process should not be taken too far because a test measuring a more complex area will be made less relevant if only one type of item is retained.]

Introduction to test analysis strategies

In this introduction, you will analyze a set of data by hand. Scores on a test are usually obtained by adding scores on each of the tasks. The validity of adding task scores depends upon the tasks belonging to some dimension that makes sense. Indeed, if those who tend to be good at one task do not tend to be good at another similar task, we question whether both tasks are assessing similar qualities. If the tasks are not assessing similar qualities we have no logical reason for adding the separate task scores together. If success on a task tends to be consistent with success on other tasks, we may infer that it is legitimate to add scores from each task, and that we are able to give meaning to scores on the resulting scale.

A data set is shown below in *Figure 5*. The student identification numbers are shown across the top of the columns. The item numbers are shown down the left hand side. (This layout is appropriate where each student answer strip is overlapped with other answer strips for an analysis by hand. Later in this module a different layout, appropriate for computer analysis of test items, will be used). Each column of non-bold numerals represents the responses of one student. The correct answers are shown by a 1; incorrect answers are shown by a zero. For example, student 18 was correct on the first 3 items and incorrect on the fourth item. Adding 1 mark for each correct item down the page gives the score obtained by a person. For example, student 7 has 10 correct responses. Adding 1 mark for each item across the page gives the score obtained by an item. For example, 5 students were correct on item 20.

Test analysis investigates the patterns of responses for both persons and items. Some of the techniques will be illustrated initially with this set of data (see *Figure 5*).

Figure 5.　Matrix of student data on a twenty-one item test

Items																			

	Students

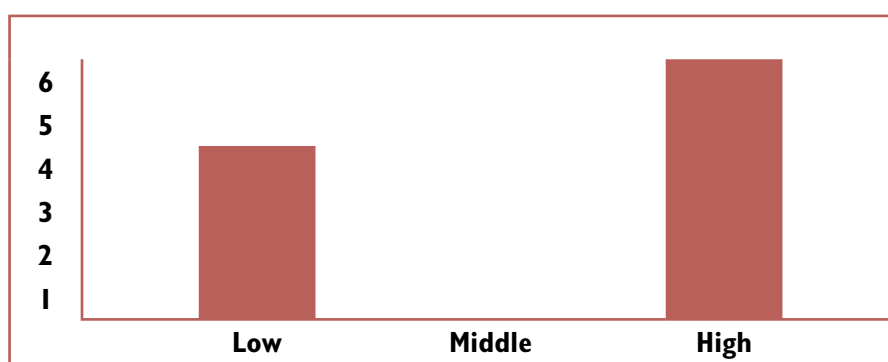| Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	15
2	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14
3	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	14
4	0	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	0	0	11
5	1	0	0	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1	13
6	0	0	0	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	11
7	0	0	0	0	0	0	0	1	0	1	0	1	1	1	1	0	1	1	8
8	0	0	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	13
9	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1	9
10	0	1	0	0	0	0	1	1	0	0	0	0	1	1	1	0	0	1	7
11	0	0	0	0	1	0	0	1	0	1	0	0	1	1	1	0	1	0	7
12	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	6
13	0	0	0	0	0	1	1	0	0	0	1	1	0	0	0	1	1	1	7
14	0	0	1	0	0	0	1	1	1	1	1	1	0	1	0	1	1	1	11
15	1	0	1	1	0	0	0	0	1	1	1	1	1	0	1	1	1	1	12
16	0	0	0	1	0	0	0	0	1	0	0	1	1	0	1	1	1	1	8
17	0	0	0	0	0	1	0	0	1	0	1	1	0	1	0	1	1	1	8
18	0	1	1	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	6
19	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	4
20	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	1	1	5
21	1	1	1	1	1	0	1	0	0	0	1	1	1	1	0	0	0	0	10
	3	4	5	7	7	9	10	10	12	12	14	14	14	14	15	15	17	17	199

When the data were entered into the table, the data for the student with lowest score was entered first, then the student with the next lowest score, and so on.

In *Figure 6*, the position of the rows (item scores) has been altered so that the easiest item is at the top of the matrix and the other rows are arranged in descending order. Notice that the top right corner of the matrix has mostly entries of 1s, and the lower left corner has mostly entries of 0s.

In *Figure 6*, the students have been assigned to 3 (equal) groups. The highest 6 scorers will be called the **High** group; the lowest 6 scorers will be called the **Low** group; the **Middle** group of 6 has been shown *underlined*. (Note that to form three groups of equal size, the middle group has some students with the same score as students in the high group).

We can investigate the patterns of success for each item (in an approximate way) by graphing the success rate of the **Low** group and the corresponding success rate of the **High** group. (We will ignore the **Middle** group for the moment). (See *Figure 6*).

Figure 6.   **Students divided into three groups according to score**

Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
																			Students
1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	15
2	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14
3	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	14
5	1	0	0	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1	13
8	0	0	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	13
15	1	0	1	1	0	0	0	0	1	1	1	1	1	0	1	1	1	1	12
4	0	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	0	0	11
6	0	0	0	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1	11
14	0	0	1	0	0	0	1	1	1	1	1	1	0	1	0	1	1	1	11
21	1	1	1	1	1	0	1	0	0	0	1	1	1	1	0	0	0	0	10
9	0	0	0	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1	9
16	0	0	0	1	0	0	0	0	1	0	0	1	1	0	1	1	1	1	8
17	0	0	0	0	0	1	0	0	1	0	1	1	0	1	0	1	1	1	8
7	0	0	0	0	0	0	0	1	0	1	0	1	1	1	1	0	1	1	8
13	0	0	0	0	0	1	1	0	0	0	1	1	0	0	0	1	1	1	7
10	0	1	0	0	0	0	1	1	0	0	0	0	1	1	1	0	0	1	7
11	0	0	0	0	1	0	0	1	0	1	0	0	1	1	1	0	1	0	7
12	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	6
18	0	1	1	0	0	1	0	0	0	0	0	0	0	0	1	0	1	1	6
20	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	1	1	5
19	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	4
	3	4	5	7	7	9	10	10	12	12	14	14	14	14	15	15	17	17	199

Consider item 1 (with the data as shown in *Figure 7*).

Figure 7.   Responses for item 1

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	15

The **Low** group has 4 successes; the **High** group has 6 successes.
You can draw a graph like the one shown in *Figure 8* for item 1.

Figure 8.   Correct answer responses for item 1

Next consider item 4 (*Figure 9*)

Figure 9.   Responses for item 4

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
4	0	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	0	0	11

The **Low** group has 2 successes; the **High** group has 4 successes.
You can draw the graph like the one shown in *Figure 10*.

Figure 10.   **Correct answer responses for item 4**
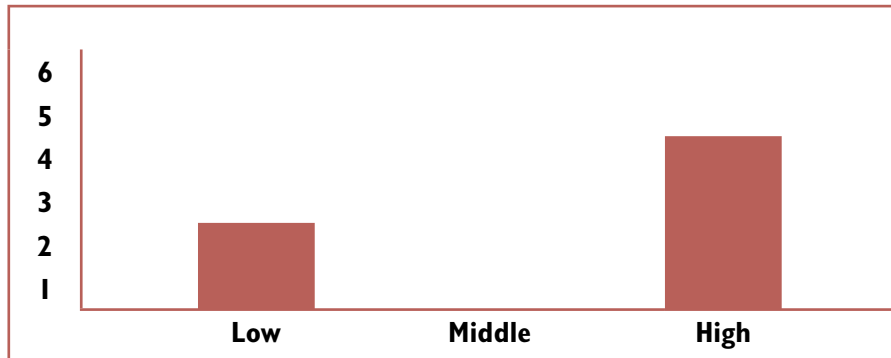
Note that in each case, although the actual numbers differ, the low group had less success than the high group. This is the expected pattern for correct answers if the item measures the same skills as the whole test. Now look at the pattern for item 19 (*Figure 11*).

Figure 11. **Responses for item 19**

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
19	0	0	0	0	0	1	<u>0</u>	<u>0</u>	<u>1</u>	<u>0</u>	<u>0</u>	<u>1</u>	0	0	0	0	1	0	4

The **Low** group has 1 success; the **High** group has 1 success. You can draw a graph like the one shown in *Figure 12*.

Figure 12. **Correct answer responses for item 19**
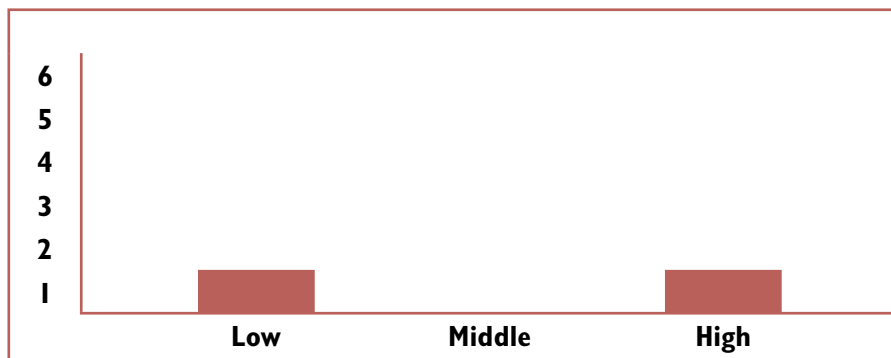
In this case the columns are equal. If these data were from a larger sample and gave this pattern, we could conclude that item 19 was not consistent with the rest of the test. Further, if the low group did better than the high group we would think that there was something wrong with the item, or that it was measuring something different, or that the answer key was wrong. Test analysis can identify a problem with an item but the person doing the analysis has to work out why this is so.

Look again at the graphs *for correct answers* for items 1, 4, and 19 (as shown in *Figure 13* below). Trend lines have been added. Items performing as expected have a rising slope from left to right for the correct answers. Item 19 does not show a rise; the data for this item show no evidence that the item distinguishes between those who are able and those who are not (where the criterion groups are determined from scores on the test as a whole). For item 21 (*Figure 13* below) there is evidence that this item distinguishes between those who are able and those who are not (as determined from the test as a whole) but *not in the expected direction*. Those who are less able are *better* on this item than those who are more able. It may be that the score key has the wrong 'correct' answer, that the item is testing something different from the other items, that the better candidates were taught the *wrong* information, and/or only the weaker candidates were taught the topic because it was assumed (incorrectly) that able students already knew the work. Item analysis does not tell you which fault applies. You have to speculate on possible reasons and then make an informed judgment.

Figure 13. Correct answer responses for items 1, 4, 19, and 21

Item 1

Item 4

Item 19

Item 21
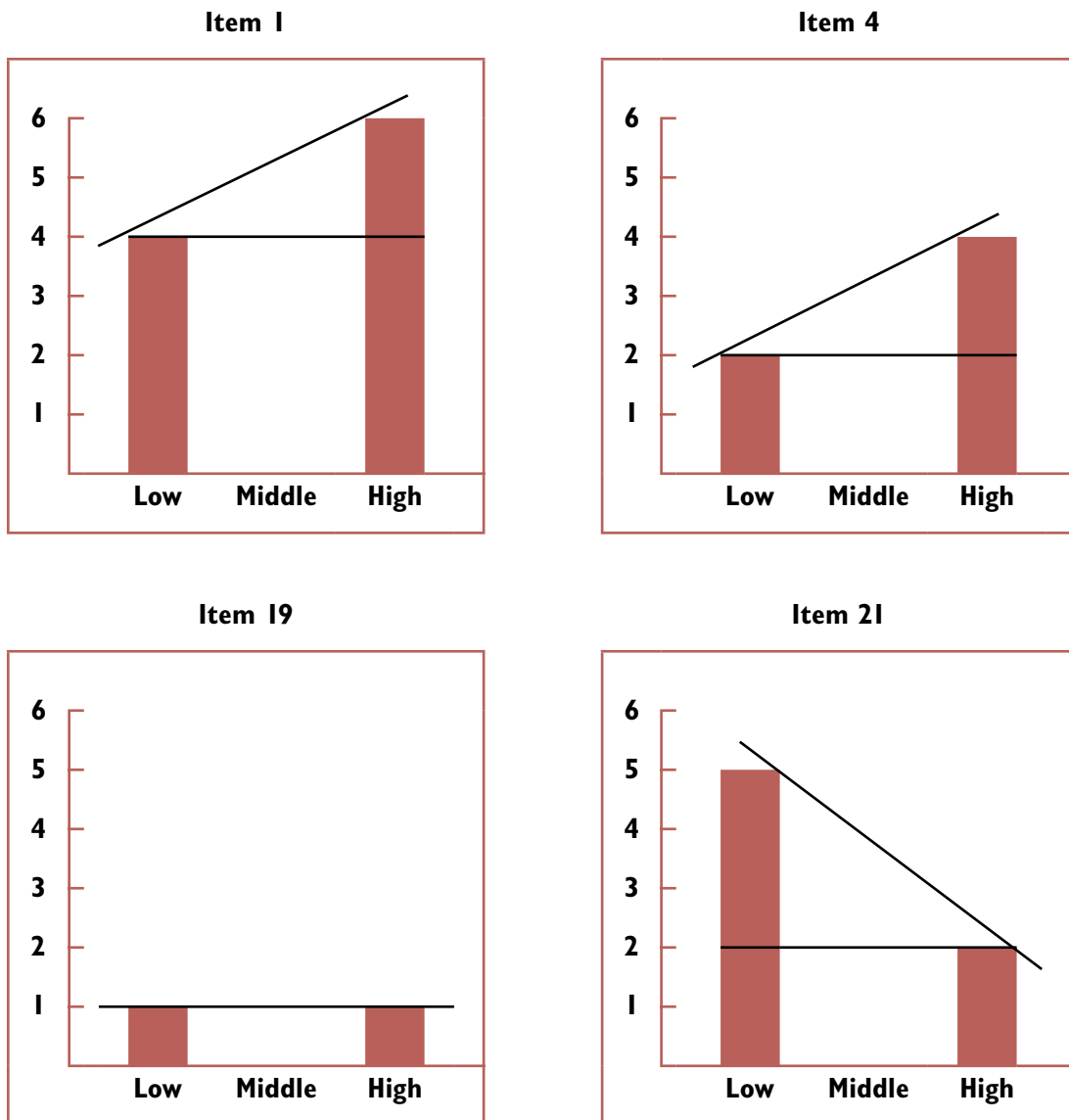
Items with correct answer patterns like items 1 and 4 distinguish between those who are low scorers on the test as a whole, and those who are high scorers. Such items are called positive discriminators; the gradient of the trend line is positive.

Items with patterns like item 19 fail to distinguish between high and low scorers. Such items are called non-discriminators; the gradient of the trend line is zero or close to zero.

Items with patterns like item 21 also distinguish between high and low scores but in the wrong direction. Such items are called negative discriminators; the gradient of the trend line is negative.

Now we return to considering the **middle** group. (If the **middle** group was not exactly the same size as the **low** and **high** groups, we would plot the proportion of candidates in each group). For item 1 there were 5 successes in the middle group; for item 4 there were 5 successes, and for item 19 there were 2 successes. For item 21 there were 3 successes.

The additional information provided by the middle group data allows us to consider how well the item distinguishes between the **Low** and **Middle** groups, and between the **Middle** and **High** groups. Items which perform as expected will have a correct answer option graph with positive discrimination. Items which do not perform as expected have a correct answer option graph with zero or negative discrimination, or have a correct answer option graph with positive discrimination in one part and not in another part.

The graphs in *Figure 14* illustrate some items with correct response patterns where caution must be exercised. The initial information based only on the high and low groups suggested that items 1 and 4 were acceptable but consideration of middle group responses showed that item 4 was problematic.

Figure 14. Correct answer responses for L, M and H groups on items 1, 4, 19, and 21

Item 1

Item 4

Item 19

Item 21
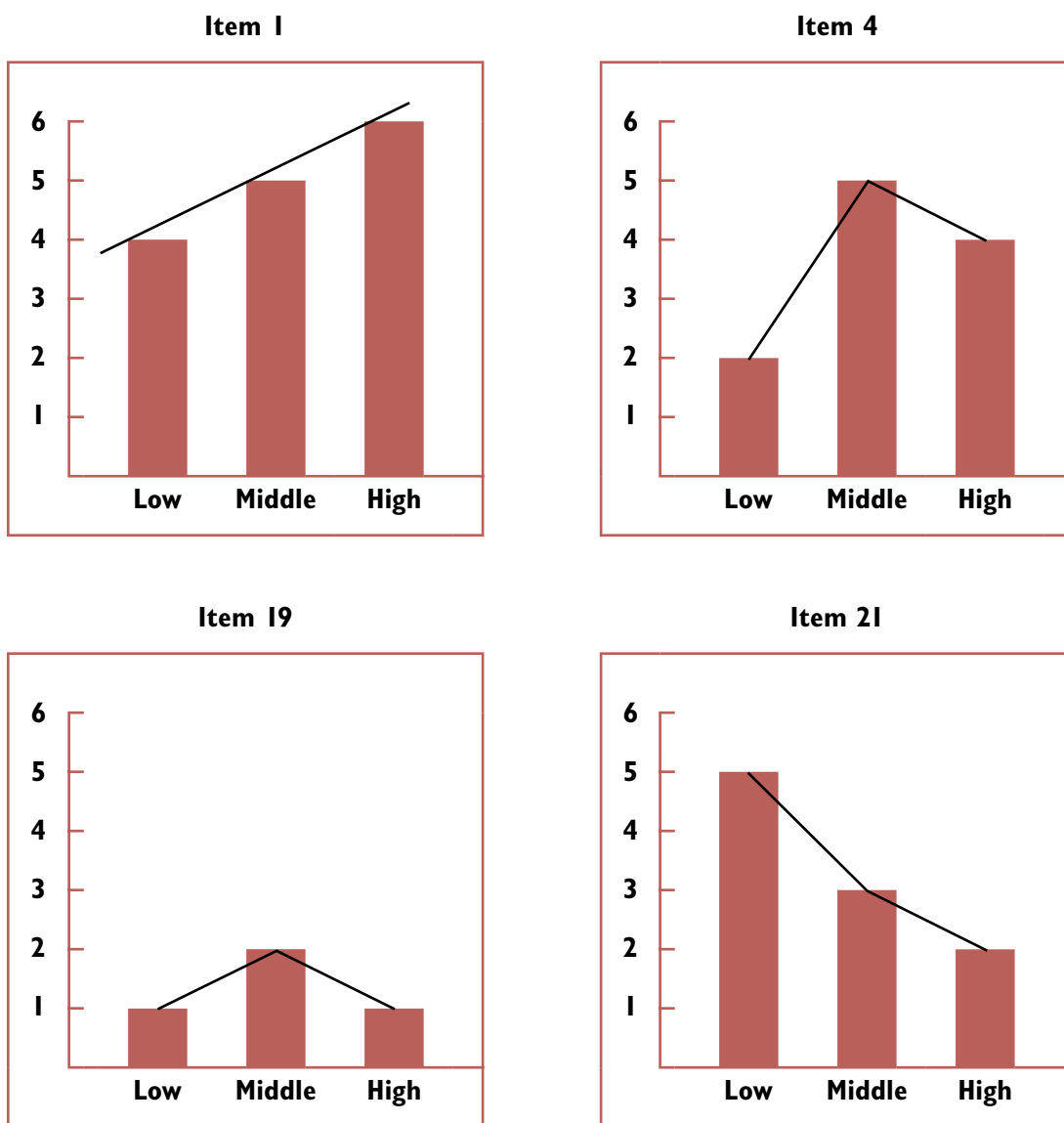
Figures 15 and *16* show some possible patterns from analyses. *Figure 15* shows one pattern that is acceptable and two patterns that may be acceptable, and *Figure 16* shows patterns that raise concern.

Figure 15. Acceptable and may be acceptable correct answer response patterns
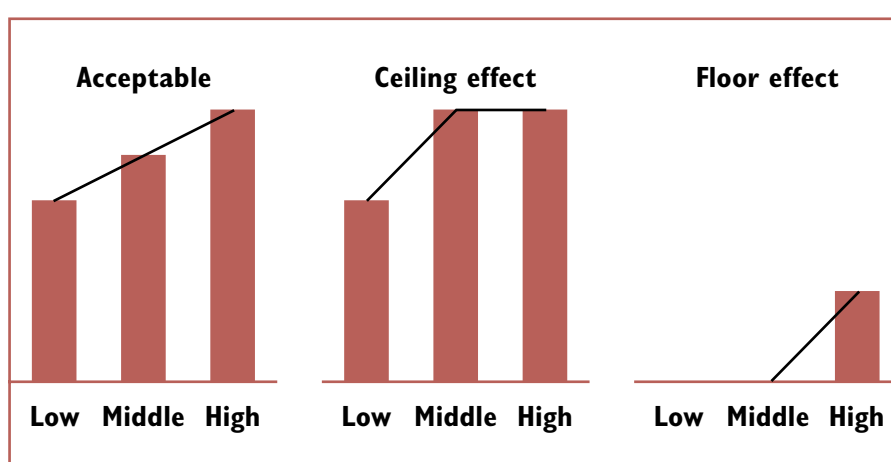
In each of these patterns, success rate improves with ability. The middle graph represents easy items where there is *no evidence* of the item distinguishing between middle and higher groups. The item may be acceptable but may not be; we reserve judgment until further evidence is obtained. Such an item is said to have a ceiling effect; the high group cannot distinguish their achievement from the middle group because the item was so easy and a trend line cannot go beyond 100 per cent correct. The right-hand graph represents difficult items where there is *no evidence* of the item distinguishing between lower and middle groups. Such an item is said to have a floor effect; the middle group cannot distinguish their achievement from the low group because the item was so difficult and a trend line cannot show less than zero percent correct.

Figure 16. **Unacceptable correct answer response patterns**
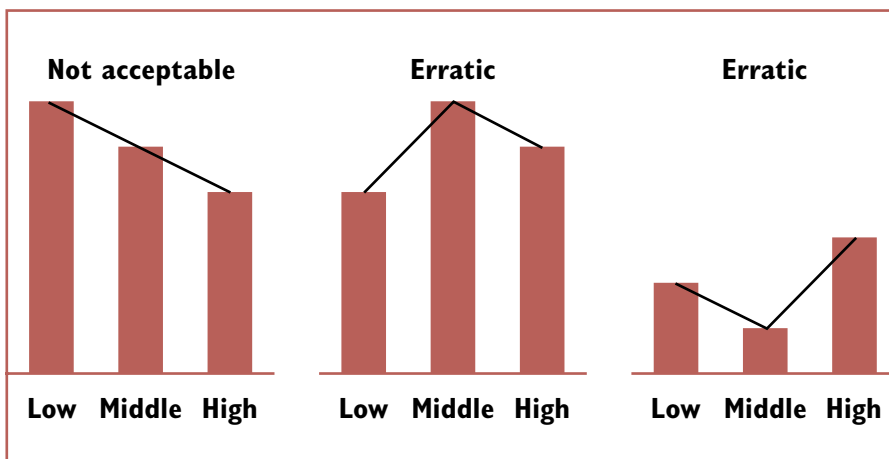
The patterns for correct responses are summarised diagrammatically in *Figure 17.*

Figure 17. **Patterns for correct responses**
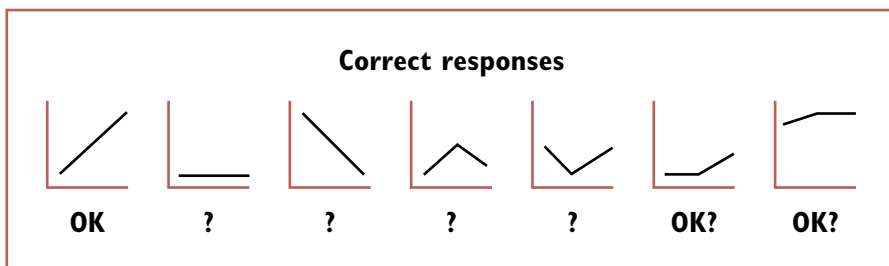
Doing an item analysis 'by hand'

These relatively simple ideas can form the basis for understanding item analysis. We will now look at some of these ideas with another data set. Doing an analysis by hand may take a longer time but it will help you understand the analysis process. (It is more efficient to let the computer do the analysis provided that you know what you are doing).

The data for analysis are shown below (*Figure 18*). In this figure the candidates are listed in the left column. Each row shows the responses to the items. Acceptable responses (correct and incorrect) are *1, 2, 3, 4, 5* and *6*. The first five acceptable responses are multiple-choice options for each item. (In this example the responses have been entered as numerals, but they could have been entered as letters such as A, B, C, D, E and F). The key (the list of correct answers in the correct order for this test) is supplied at the bottom of the response data. The *6* indicates that the question was omitted, but candidates had sufficient time to attempt all items. To help line up columns, the last two lines show the item numbers.

On a copy of this table of data, use a coloured pencil to highlight each correct answer. For example, in the first column after the candidate identification code (item 1), each 5 should be highlighted. Other responses such as 6 and 2 should **not** be highlighted. Repeat this procedure for each item in turn. Then count the number of highlighted numerals to obtain a total score for each candidate; write each total at the right-hand end of each row. Then count the number of highlighted numerals to obtain a total score for each item; write each total at the bottom of the column for that item.

Count the number of candidates. Use the candidate totals to identify the top one-third; mark these to show they are in the **High** group. Identify the bottom one-third; mark these to show they are in the **Low** group. Identify the middle one-third; mark these to show they are in the **Middle** group. (You might find it useful to cut up your piece of paper into rows with one candidate's results to a row. Then paste each slip of paper in order of total score).

X03, X08, X19, X21, X22, X24, X26, X14, and X15 will be in the high group; X05, X25, X09, X06, X12, X13, X20, X11, and X23 will be in the middle group; and X16, X27, X07, X01, X02, X04, X10, X17, and X18 will be in the lower group).

Make some tables like *Figure 19*. Use one table for each item. Taking each item in turn, count how many from the **High** group chose 1, how many chose 2, how many chose 3, and so on. As you complete each count, write the result in your table for that item.

Figure 18. Responses on a multiple-choice test of 30 items

X01	6	2	4	2	3	6	5	4	3	5	1	3	2	3	2	2	3	3	5	2	4	2	1	4	3	2	1	2	2	2
X02	5	2	4	2	2	5	5	1	3	4	1	4	2	5	1	1	3	2	9	9	5	5	1	9	3	1	3	2	2	2
X03	5	2	4	2	5	5	5	4	3	1	1	4	5	5	5	2	3	3	3	5	4	5	4	1	4	3	9	2	1	2
X04	2	2	4	1	2	5	1	5	3	1	1	1	2	5	5	2	3	5	5	4	2	5	5	4	3	2	2	2	1	2
X05	5	2	4	3	4	5	5	5	3	5	1	4	2	5	3	2	3	3	5	5	4	5	4	2	4	2	3	2	3	3
X06	5	2	4	2	3	9	1	1	3	1	1	4	2	5	4	2	3	3	5	4	4	5	4	1	3	2	2	2	1	2
X07	5	2	4	1	2	1	5	1	3	1	1	4	2	5	5	3	3	3	3	1	1	5	4	2	1	4	3	2	4	2
X08	5	2	4	2	5	5	5	4	3	5	2	4	2	5	5	2	3	3	3	5	4	3	4	1	5	4	5	2	1	2
X09	5	2	4	1	5	1	2	4	3	2	1	4	2	5	5	2	1	1	5	5	4	5	4	3	4	3	3	2	1	2
X10	5	1	4	1	5	1	2	4	3	1	3	4	2	3	2	1	3	3	4	1	5	5	2	1	3	2	4	3	1	2
X11	5	2	4	1	5	1	5	2	3	2	5	3	2	2	5	2	1	3	3	5	4	5	5	1	1	2	3	2	1	2
X12	5	2	4	2	5	5	1	4	3	1	1	4	2	5	5	1	3	3	5	5	4	2	4	2	1	4	1	4	1	2
X13	5	2	4	2	2	5	5	5	9	1	1	4	2	5	5	2	3	3	3	9	9	5	9	9	9	2	9	2	1	2
X14	5	2	4	2	2	5	5	4	3	5	1	4	2	5	5	2	3	3	4	9	4	5	4	9	9	9	1	2	1	3
X15	5	2	4	2	2	5	5	2	3	5	1	4	2	5	5	2	3	3	5	9	4	5	1	4	1	2	2	2	1	2
X16	5	2	4	2	9	5	5	5	3	4	1	4	2	5	5	9	3	5	5	9	4	5	4	9	1	9	9	2	1	2
X17	5	2	4	5	3	2	5	2	3	5	5	1	2	5	5	5	1	9	5	5	4	5	5	3	9	9	9	2	1	2
X18	5	2	4	1	1	5	1	4	4	5	1	4	2	3	2	1	3	3	5	9	9	2	5	1	2	2	2	4	1	1
X19	5	2	4	2	5	5	5	3	3	1	1	4	2	5	5	2	3	3	3	9	4	5	4	1	9	9	4	2	1	3
X20	5	2	4	2	4	5	5	2	5	9	1	4	2	5	5	1	3	3	3	9	4	5	4	9	9	3	9	2	1	2
X21	5	2	4	2	5	1	5	9	3	5	1	4	2	5	5	2	9	3	3	5	4	9	4	9	4	9	9	2	1	2
X22	2	2	4	2	5	5	5	4	3	5	1	4	2	5	5	2	3	3	3	3	1	4	2	1	1	5	2	2	1	2
X23	5	2	4	2	5	5	5	4	3	9	1	4	2	2	4	1	1	3	2	4	4	5	3	1	4	3	3	2	2	2
X24	5	2	4	2	5	5	5	4	3	1	1	4	2	1	5	2	3	3	5	2	4	5	4	1	1	3	5	2	1	2
X25	5	2	4	2	2	5	2	9	3	5	2	4	2	5	5	2	3	3	3	9	4	5	4	9	3	9	5	2	1	2
X26	5	2	4	2	5	3	5	4	5	5	1	4	2	5	5	2	2	3	2	3	4	5	4	3	4	3	2	2	1	2
X27	5	1	4	2	2	5	2	4	2	3	1	4	2	1	5	2	3	3	1	3	2	5	4	3	4	4	4	2	2	2
Key	5	2	4	2	5	5	5	4	3	5	1	4	2	5	5	2	3	3	3	5	4	5	4	1	4	2	4	2	1	2
Item	^	^	^	^	^	^	^	^	^ 1		^	^	^	^	^	^	^	^	^	^ 2		^	^	^	^	^	^	^	^	^ 3
Num	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0

Figure 19. Blank data table for an item

Item								
No —		**Option**						
		I	**2**	**3**	**4**	**5**	**Other**	**Total**
	H	—	—	—	—	—	—	—
	M	—	—	—	—	—	—	—
	L	—	—	—	—	—	—	—
Total		—	—	—	—	—	—	—

**Figure 20. Data table for item I**

Item								
No –I–		**Option**						
		I	**2**	**3**	**4**	***5**	**Other**	**Total**
	H	—	–I–	—	—	–8–	—	–9–
	M	—	—	—	—	–9–	—	–9–
	L	—	–I–	—	—	–7–	–I–	–9–
Total		—	–2–	—	—	–24–	–I–	–27–

Item 1 has been completed (*Figure 20*) to show you how the results are recorded. The ***** indicates the option that was keyed as correct.

This processing of the data has to be accurate. If several people are working together on this analysis, each person may process a subset of items.

When all items have a completed table of data, the information for the *keyed responses* can be graphed. A graph for item 1 is shown below (*Figure 21*), together with a blank graph (*Figure 22*). These graphs can be compared with those in *Figure 17.*

Figure 21. Graph for item 1

Item –1–

	Low	Middle	High
10			
9		▮	
8		▮	▮
7	▮	▮	▮
6	▮	▮	▮
5	▮	▮	▮
4	▮	▮	▮
3	▮	▮	▮
2	▮	▮	▮
1	▮	▮	▮
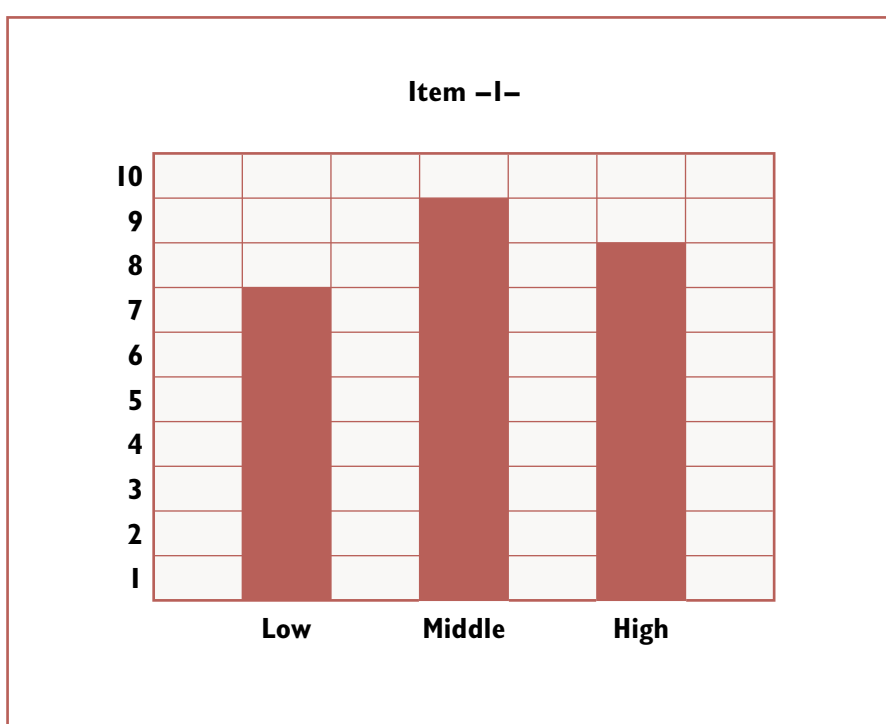
Figure 22. **Blank graph for an item**

Item —

10						
9						
8						
7						
6						
5						
4						
3						
2						
1						

    Low   **Middle**   **High**
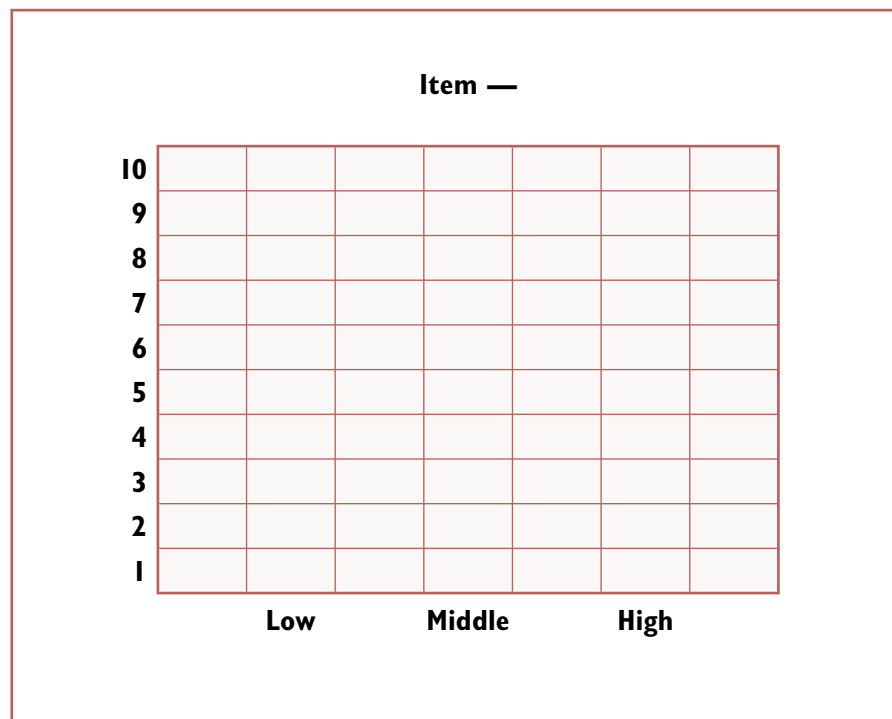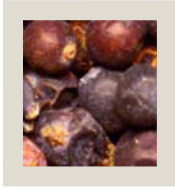
Item analysis approaches using the computer

9

There are two main types of approaches to item analysis used extensively in test research and development organizations. Some use one approach, some use the other, and some use both approaches in conjunction with each other. In this module the earlier approach will be called the *Classical* (or traditional) item analysis, and the more recent approach will be called *Item Response Modelling*.

The first step in an item analysis is to choose an appropriate criterion measure, which can be used to make judgments concerning whether an item discriminates between better performing students and poorer performing students. Many test research and development agencies assume that the total score on the test is the best criterion measure available. Criterion groups are set up on the basis of total scores on the test and each item's correlation with the total score is reported. (Note that there is a built-in spurious correlation here because each item is included in the total score. With tests of 20 items or more, the effect of the item contribution is ignored in practice).

The older classical approach to item analysis seeks to identify items which do not distinguish between high and low scorers in a similar way to a criterion measure. The extent of agreement between the item and the criterion measure in ordering the candidates is reported as a correlation coefficient, often the point-biserial correlation coefficient. (The phi coefficient, often estimated

by a graphical means, was used widely before the use of personal computers became common). These correlation indices range between -1 through 0 to +1.

Zero or low correlations and negative correlations identify items to be queried, discarded, revised, or replaced. High positive correlations identify items to be retained. The degree of success or failure for a particular item is usually measured either by the percentage of candidates *correct* or by the percentage of candidates *incorrect*. Both percentages have been called the *difficulty* of the item in various textbooks on measurement. For consistency and to avoid confusion, *percentage correct* should be called *facility* and *percentage incorrect* should be called *difficulty*.

The item response modelling approach to item analysis also seeks to identify items which do *not* distinguish between high and low scorers in a similar way to a criterion measure. However, this approach takes a more detailed look at the capacity of the item to distinguish between other subsets of the scorers. For example, to distinguish between low and middle scorers, and between middle and high scorers. Items are assigned a position on a scaled difficulty continuum from easiest to most difficult.

Candidates are assigned a position on a scaled ability or achievement continuum *in the same metric as the item difficulty continuum*. High achievers among the candidates and difficult items on the test are near the top end of the continuum; low achievers and easy items are near the bottom end of the continuum.

The actual pattern of responses resulting from the interactions of items with candidates is compared with a model pattern consistent with the observed marginal totals. The extent of agreement between the observed pattern and the model in ordering both the candidates and the items is reported in terms of fit statistics. Candidates and items with unusual patterns in the correct responses are identified

to be queried. Items may be discarded, revised, or replaced. Explanations are sought for unusual candidate patterns. There are several separate variations within item response modelling (sometimes known as Item Response Theory or IRT). In this module, only one of these variations will be used, the Rasch model (named after the Danish statistician who published his research findings in 1960).

We now look at these two types of analysis in turn, compare the approaches, showing where they agree on item quality and where they differ.

Classical strategies for item analysis

The high group/low group procedures used in the analysis of data by hand in the section on the introduction to item analysis strategies above are simplified examples of classical item analysis. Personal computers have made the task of scoring the test, counting the cases, calculating the percentages, and calculating the correlations between success on items and total score, easier, particularly for multiple-choice tests. The discussion of the clerical approach presented below has used the ITEMAN computer program to analyze the data presented in *Figure 18*.

The first part of the computer output from a traditional test analysis report for a multiple-choice test might look like *Figure 23*.

Figure 23. **Classical item analysis for data in** *Figure 18*

MicroCAT (tm) Testing System
Copyright (c) 1982, 1984, 1986, 1988 by Assessment Systems Corporation

Item and Test Analysis Program – ITEMAN (tm) Version 3.00

Item analysis for data from file iiepitm.dat   Page 1

		Item Statistics				Alternative Statistics			
Seq. No.	Scale -Item	Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
1	0-1	0.889	0.264	0.159	1	0.000	-9.000	-9.000	
					2	0.074	-0.051	-0.027	
					3	0.000	-9.000	-9.000	
					4	0.000	-9.000	-9.000	
					5	0.889	0.264	0.159	*
					Other	0.037	-0.531	-0.227	

The interpretation of this part of the printout is now described.

Each item discrimination (the measure of the extent of agreement between success on the item and success on the test as a whole) is shown opposite the item number. For example, the discrimination for item 1 is shown as **Point Biser. = 0.159**. (Usually reported as 0.16). The **Alt.** column shows the options. The first category can be 1 or A, the second 2 or B, and so on. The correct answer is shown by * . The **Prop. Endorsing** column shows the proportion of candidates who chose each option. The **Point Biser.** (Point biserial correlation coefficient) statistic shows the extent of agreement between the option and the test as a whole. The **Biser.** (biserial

correlation coefficient) statistic provides another statistic which also shows the extent of agreement between the option and the test as a whole. (Note that values of the point biserial correlation coefficient tend to be smaller in magnitude than if the same data are analyzed using the bisevial correlation coefficient).

(**-9.000** means *Not Applicable*. The extent of agreement cannot be calculated where no candidate has chosen an option). The last option (**Other**) indicates missing data – that is, no response at all.

The analysis for item 23 is shown in *Figure 24*. This item has many good qualities. It is in an appropriate range of difficulty (the proportion correct was **0.593**) and those who were incorrect are spread over each of the other options. The 'correct' option has a substantial positive agreement (**0.549**) with the test as a whole. All of the 'incorrect' options have negative agreements: 1 **-0.072**; 2 **-0.287**; 3 **-0.049**; and 5 **-0.515** with the test as a whole.

Figure 24. Analysis results for item 23

		Item Statistics				Alternative Statistics			
Seq. No.	Scale -Item	Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
23	0-23	0.593	0.695	0.549	1	0.148	-0.148	-0.072	
					2	0.037	-0.670	-0.287	
					3	0.037	-0.113	-0.049	
					4	0.593	0.695	0.549	*
					5	0.148	-0.792	-0.515	
					Other	0.037	0.026	0.011	

By contrast, item 27 (*Figure 25*) has a pattern of results suggesting that either the item has been mis-keyed, or the candidates have been taught incorrect information.

Figure 25. Analysis results for item 27

	Item Statistics				Alternative Statistics			
Seq. No.	Scale -Item	Prop. Correct	Point Biser.	Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
27	0-27	0.111	-0.145	-0.088	1	0.111	-0.086	-0.052	
					2	0.222	-0.071	-0.051	
	CHECK THE KEY 4 was specified, 5 works better				3	0.222	-0.298	-0.214	
					4	0.111	-0.145	-0.088	*
					5	0.111	0.568	0.342	
					Other	0.222	0.155	0.111	

The test analysis program has identified option 5 as a more likely correct answer (because the measure of agreement for that option is more positive than the keyed option). Note that the keyed option has a negative agreement (**-0.088**) with the test as a whole, while option 5 has a positive agreement (**0.342**). Either the item key is correct and a substantial proportion of the better candidates are misinformed, or the item key is incorrect. *If an error in the item key is found, it **must** be corrected and the analysis **must** be done again.*

Item 25 (*Figure 26*) is similar to item 27, but identifying the problem with the item may be difficult. The keyed option does have a positive agreement (**0.265**) with the test as a whole. However other options also have positive agreements (**0.030** and **0.403**). The test

analysis program has identified the largest positive agreement as a likely correct answer. However this type of pattern may occur when there is more than one correct answer. For item 25, it appears that the best correct answer may be option 5 and that option 4 may be another correct answer (that is, if mis-information is not a feasible explanation). Test construction experts often suggest that amendment is required so that there is only one correct answer for an item. *If an error in the item key is found, it **must** be corrected and the analysis **must** be done again.*

Figure 26. Analysis results for item 25

		Item Statistics				**Alternative Statistics**			
Seq. No.	Scale -Item	Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
25	0-25	0.259	0.358	0.265	I	0.222	0.042	0.030	
					2	0.037	-0.809	-0.347	
		CHECK THE KEY 4 was specified, 5 works better			3	0.222	-0.639	-0.457	
					4	0.259	0.358	0.265	*
					5	0.074	0.753	0.403	
					Other	0.185	0.081	0.056	

Sometimes an item has some options which work and some which contribute nothing to distinguishing between those who have knowledge and those who do not. In Item 7 (*Figure 27*), options 3 and 4 were not endorsed by any person, and no index of agreement with the test as a whole could be calculated (as shown by **-9.000**). In effect, only part of this item has worked; those who constructed the item need to provide two more attractive options.

Figure 27. Analysis results for item 7

	Item Statistics				Alternative Statistics			
Seq. No.	Scale -Item	Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
7	0-7	0.704	0.505	0.383	I	0.148	-0.500	-0.325	
					2	0.148	-0.256	-0.167	
					3	0.000	-9.000	-9.000	
					4	0.000	-9.000	-9.000	
					5	0.704	0.505	0.383	*
					Other	0.000	-9.000	-9.000	

In some cases, the item may have more than one fault. For example, item 13 (*Figure 28*) appears to be mis-keyed (or the better candidates are mis-informed) and some of the options do not attract.

Figure 28. Analysis results for item 13

	Item Statistics				Alternative Statistics			
Seq. No.	Scale -Item	Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
13	0-13	0.963	-0.861	-0.369	I	0.000	-9.000	-9.000	
					2	0.963	-0.861	-0.369	*
	CHECK THE KEY 2 was specified, 5 works better				3	0.000	-9.000	-9.000	
					4	0.000	-9.000	-9.000	
					5	0.037	0.861	0.369	
					Other	0.000	-9.000	-9.000	

Item 3 (*Figure 29*) is an example of an item which every candidate can do successfully. For this group of candidates, there is no evidence that this item is useful in distinguishing between able and less able candidates.

Figure 29. Analysis results for item 3

	Item Statistics					Alternative Statistics			
Seq. No.	Scale -Item	Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
3	0-3	1.000	-9.000	-9.000	1	0.000	-9.000	-9.000	
					2	0.000	-9.000	-9.000	
					3	0.000	-9.000	-9.000	
					4	1.000	-9.000	-9.000	*
					5	0.000	-9.000	-9.000	
					Other	0.000	-9.000	-9.000	

The next section provides a brief summary of the key aspects to consider when evaluating a set of test items.

Deciding whether an item is useful after trial with real candidates (classical analysis)

The steps are:

1. ***Find the correct option.***
   This is indicated in the **Key** column with the *.

2. ***Is the agreement index (Point-Biserial) positive?***
   If *Yes*, continue to **3**;
   If *No*, this is an unexpected result! Check why! Probably you need to change or reject the item. Check that the score key is correct!

   In practice, some positive agreement index values are small. Some are so small as to be effectively zero. The position of the cut-off between zero and non-zero index values depends upon the size of the candidate group. With a candidate group of 60, values less than about 0.249 are traditionally regarded as zero. The corresponding approximate values for larger group sizes are 80 (0.217), 100 (0.194), 120 (0.178), 140 (0.165), 160 (0.154), 180 (0.145), and 200 (0.138). If your trial test involved 200 candidates, then items with a correct-option point-biserial index of less than 0.138 would be rejected.

   [Some classical test analysis programs provide a probability value associated with each option (sometimes called a **p-value**). The **p-value** shows the probability of the agreement index value occurring by chance. If the probability is higher than a chosen value, we treat the correlation as approximately zero. Traditional chosen values for the cut-off between 'zero' and 'acceptable' are p=0.05, p=0.01, and p=0.001. For p=0.05 we take a risk that for 1 in 20 cases we may accept an item as in agreement when it is only a chance agreement. For p=0.01, the risk is 1 in 100 and for p=0.001 the risk is 1 in 1000.

More conservative risk values result in more items being rejected. Many test developers use the correlation associated with p=0.05 as the cut-off; a lower point-biserial value (that is a higher p value) leads to rejection of the item. If the program provides a **p-value** then an additional question is asked: *Is the p-value 0.05 or less?* If *Yes*, continue to **3**; If *No*, probably change or reject the item.]

3. *Are the wrong option agreements negative?*
   If *Yes*, keep the item and continue to **4**;
   If *No*, consider each wrong option in turn. If an incorrect option has a positive correlation about the same as the correct option or higher, check the score key (this option may be an alternative correct answer that has not been credited as such). Options that are not chosen by any candidate are often replaced and the item is then retained for further trial. If there is no serious problem, keep the item and continue to **4**; otherwise change or reject the item.

4. *Assembling final forms of the test*
   We consider the position where the item belongs in the test specification table and the difficulty of the item. Each cell in the test specification table should have several discriminating items and a range of difficulties.

The other test items are considered in the same way. The final page of the ITEMAN test analysis looks like the information in *Figure 30*. Comments on the printout have been added.

Figure 30. Classical item analysis summary statistics

MicroCAT (tm) Testing System
Copyright (c) 1982, 1984, 1986, 1988 by Assessment Systems Corporation

Item and Test Analysis Program – ITEMAN (tm) Version 3.00

Item analysis for data from file iiepitm.dat   Page 7

There were 27 examinees in the data file.

Scale Statistics

Scale:	0	<-- This is the scale identification code.
N of items	30	<-- The number of items on this scale.
N of Examinees	27	<-- The number of candidates.
Mean	19.815	<-- The mean (or average) for this group of 27 persons (on 30 questions).
Variance	10.818	<-- A measure of spread of test scores for these candidates.
Std. Dev.	3.289	<-- Another measure of spread of test scores for these candidates. (The standard deviation is the square root of the variance
Skew	-0.111	<-- This index summarises the extent of symmetry in the disribution of candidates scores. A symmetrical distribution has a skewness of 0; negative values indicate more high scores than low scores and positive values indicate more low scores than high scores.
Kurtosis	-0.893	<-- This index compares the distribution of candidate scores with a particular mathematical distribution of scores known as the *Normal* or *Gaussian* distribution. Positive values indicate a more peaked distribution than the specified distribution; negative values indicate a flatter distribution
Minimum	14.000	<-- This is the lowest candidate score in this group.
Maximum	26.000	<-- This is the highest candidate score in this group.
Median	20.000	<-- This is the middle score when all candidates scores in this group are arranged in order.
Alpha	0.543	<-- This index indicates how similar the questions are to each other. The lowest value is 0.0 and the highest is 1.0. Provided that candidates had ample time to complete each item, higher values indicate greater internal consistency in the items. (See *Test Reliability* below).
SEM	2.224	<-- We use this index to estimate how much the scores might change if we gave the same test to the same candidates on several occasions (See *Test Reliability* below).
Mean P	0.660	<-- This is the average proportion correct for these items with these candidates.
Mean Item-Tot.	0.254	<-- This is the average point biserial correlation for these items.
Mean Biserial	0.338	<-- This is the average biserial correlation for these items.

Test reliability

The term *validity* refers to usefulness for a specified purpose and can only be interpreted in relation to that purpose. In contrast, *reliability* refers to the consistency of measurement regardless of what is measured. Clearly, if a test is valid for a purpose it must also be reliable (otherwise it would not satisfy the usefulness criterion). But a test can be reliable (consistent) without meeting its intended purpose. Test reliability is influenced by the similarity of the test items, the length of the test, and the group on which the test is tried. When we add scores on different parts of a test to give a score on the whole test, we assume that the test as a whole is measuring on a single dimension or construct, and the analysis seeks to identify items which contradict this assumption. In the context of test analysis, removing items which contradict the single-dimension assumption should contribute to a more reliable test. Where trial tests vary in length, the reliability index for one test cannot be compared directly with another. An adjustment to a common-length test of 100 items can be made using the Spearman-Brown formula:

$$\text{reliability}_{100\ item\ test} = \frac{\text{reliability}_{original\ test} \times (100/\text{number of items}_{original\ test})}{[1 + \text{reliability}_{original\ test} \times (100/\text{number of items}_{original\ test} - 1)]}$$

If the group of candidates is more diverse, the index obtained will be higher than for a less diverse group. For example, students at the one age-level will be less diverse than a group with students of several age-levels. A test reliability quoted for a sample of Grades 4, 5, 6 and 7 students is expected to have a higher value than a test reliability for the same test given to a similar size sample of a single Grade level (such as Grade 6).

There are a number of methods for estimating reliability; item analysis software programs generally only use one of these methods. There are four basic approaches.

- The same test can be given on two different occasions to the same sample of candidates; the reliability coefficient could then be calculated by correlating the scores on the two occasions.

- Two separate parallel tests can be given to the same sample of candidates; the reliability coefficient could then be calculated by correlating the scores on the two tests. (One variant is to delay the second test to assess stability over time).

- A single test can be split into two parts; the reliability coefficient could then be calculated by correlating the scores on the two parts. (In this case each part test is not as long as the complete test so an adjustment has to be made using the Spearman-Brown formula).

- The reliability can be calculated as an internal consistency from a single set of test data; this may be considered as equivalent to the average of all possible adjusted split-half coefficients. This is the approach used most often by item analysis computer programs.

The last two approaches only assess on one occasion so there is no assessment of stability over time.

Reliability is sometimes estimated in order to judge how precise a candidate's score might be. Various test analysis programs use different measures of item consistency. The reliability index may be described as an item homogeneity index, an internal consistency index, a Kuder-Richardson Formula 20 index, or a (Cronbach) Alpha index. For example, the ITEMAN program calculates an Alpha index which is a measure of the internal consistency of the test. In the item-analysis example above, the Alpha index is 0.543.

In practice, a reliability index for a test should be at least 0.7 and preferably higher than 0.8.

By making some assumptions about a particular candidate being similar to other candidates, the spread of scores of other candidates can be combined with the estimate of reliability to estimate a band of scores in which that candidate's score might fall if the test was given again. In the item-analysis example above this statistic (with a value of 2.224) is called the SEM, the standard error of measurement. For the item-analysis example above, we might expect that two thirds of the time the 'true score' of a candidate (the average score for an individual for an infinite number of test administration will fall within the candidate's observed score on the test plus or minus 2.224. Doubling the error limit provides a score range for the true score for 95 per cent of the time. To illustrate, we would expect that 95 per cent of the time the true score for a candidate who obtains 20 on the test would fall between 20 – (2x2.224) and 20 + (2x2.224).

Item response modelling strategies for item analysis

Part of the computer output from an item response modelling test analysis report for a multiple-choice test might look like *Figure 31*, showing how the items and the candidates are placed on the continuum. This output was produced by applying the QUEST computer program to an analysis of the data in Figure 18. Notice that item 3 is *not* shown. [If every person is correct on an item (or incorrect on an item), that item cannot be placed on the graph. Similarly, a person who has every item correct cannot be placed on the continuum. We know they are better than the next best person, but we do not know how much better. A more demanding test is needed to place such persons on the knowledge continuum.

A person with a zero score cannot be placed either. We have to find what they know as well as what they do not know to locate them on the graph.] *Figure 32* shows another part of the output that is a check on the fit to the model. *Figure 33* shows details of the individual items. Items are queried if they are well to the left of, or well to the right of, the vertical dotted lines in *Figure 32*. They may also be queried if the fit t-values in the last two columns of *Figure 33* are large.

Figure 31 also shows how the development of trial tests can result in more items in some difficulty ranges and less items in others. Most of the candidates have attainments (as judged by this test) higher than the average difficulty for the items. In other words, most items have difficulties below the attainment levels of the candidates.

In effect, this test is more powerful at detecting differences between candidates at lower levels within the range than at higher levels. More valid items in a particular range of difficulty lead to more precise distinctions between candidates within that range.

Figure 31. **Variable map for test data in** *Figure 18*

QUEST: The Interactive Test Analysis System			
Item Estimates (Thresholds)			
all on all (N = 27 L = 30)			
3.0		27 ←The most difficult item	
	x	←The top candidate	
2.0	x	25	24	26
1.0	xx	20	
	xx	19	
	xxx	10	
	xx	5	
	xxxxx	8	
	xxx	23	
	x		
	x		
0.0	x	6	16 ←Average item difficulty
	xxxx	4	7
The lowest candidate →	x	14	15	21
		17	22	29
		9	11	18
		12	30	
-1.0		1	28
-2.0		2	
3.0		13 ←The easiest item	

Each x represents one student

Figure 32. **Item fit map for test data in *Figure 18***

QUEST: The Interactive Test Analysis System						
Item Fit **all on all (N = 27 L = 30)**						
INFIT **MNSQ** 0.63 0.71 0.83 1.00 1.20 1.40 1.60						
1 item 1						
2 item 2						
3 item 3						
4 item 4						
5 item 5						
6 item 6						
7 item 7						
8 item 8						
9 item 9						
10 item 10						
11 item 11						
12 item 12						
13 item 13						
14 item 14						
15 item 15						
16 item 16						
17 item 17						
18 item 18						
19 item 19						
20 item 20						
21 item 21						
22 item 22						
23 item 23						
24 item 24						
25 item 25						
26 item 26						
27 item 27						
28 item 28						
29 item 29						
30 item 30						

Figure 33 shows the raw scores for each item and the maximum scores, the ability level on the continuum where the probability of success changes from less likely to be correct to more likely to be correct. The point is called the threshold for the item. Underneath each *threshold* numeral there is another numeral indicating the error associated with the threshold estimate.

Figure 33. Item estimates for test data in *Figure 18 (part only)*

QUEST: The Interactive Test Analysis System							
Item Fit **all on all (N = 27 L = 30)**							
Item name	**SCORE**	**MAXSCR**	**TRSH** **1**	**INFT** **MNSQ**	**OUTFT** **MNSQ**	**INFT** **t**	**OUTFT** **t**
1 item 1	24	27	-1.38 .63	1.00	1.29	0.2	0.6
2 item 2	25	27	-1.83 .75	0.96	0.66	0.1	-0.2
3 item 3	0	0	Item has perfect score				
4 item 4	19	27	-0.13 .44	0.83	0.74	-0.9	-0.8
5 item 5	12	27	1.04 .41	0.86	0.84	-1.1	-0.6
6 item 6	18	27	0.05 .43	1.02	0.99	0.2	0.1
7 item 7	19	27	-0.13 .44	0.95	0.86	-0.2	-0.3
8 item 8	13	27	-0.88 .41	1.11	1.12	0.9	0.6
9 item 9	22	27	-0.77 .51	1.07	1.04	0.3	0.2
10 item 10	11	27	1.20 .41	1.09	1.17	0.7	0.7
11 item 11	22	27	-0.77 .51	1.06	1.34	0.3	0.8
12 item 12	23	27	-1.04 .56	0.89	0.65	-0.2	-0.6
13 item 13	26	27	-2.55 1.03	1.10	5.51	0.4	2.3
14 item 14	20	27	-0.32 .46	0.94	0.88	-0.2	-0.2
15 item 15	20	27	-0.32 .46	0.87	0.77	-0.5	-0.6
16 item 16	18	27	0.05 .43	0.81	0.73	-1.2	-0.9
17 item 17	21	27	-0.53 .48	1.15	1.19	0.7	0.6
18 item 18	22	27	-0.77 .51	0.90	0.73	-0.3	-0.5

Deciding whether an item is useful after trial with real candidates (item response modelling analysis)

The steps are:

1. **Look at the variable map**

    Are the items (on the right, shown with numerals) spread over a similar range as the candidates (on the left, shown X)? If *Yes*, continue to **2**; If *No*, which are much higher, items or candidates? If *items*, further less complex items are required; if *candidates*, further more complex items are required. (The ranges of items and candidates should be similar).

2. **Look at the item fit map**

    Are any of the items shown well to the left or well to the right of the vertical dotted lines? If *No*, continue to **3**; If *Yes*, this is an unexpected result! Check why! You may need to change or reject the outlying items. Check that the score key is correct! If the score key is not correct, amend it and repeat the analysis. If the score key is correct, go to **3**.

3. **Are the fit t-values in the last two columns of the item estimates table larger than 3?**

    If *No*, keep the item and continue to **4**; If *Yes*, probably change or reject the item.

4. **Assembling final forms of the test**

    We consider the position where the item belongs in the test specification table and the threshold level (difficulty) of the item. Each cell in the test specification table should have several items over a range of difficulties.

Classical item analysis and item response modelling compared

In most situations, items rejected in the classical approach to item analysis will also be rejected by the item response modelling approach. However, it is the case that the item response modelling approach sometimes rejects items that are acceptable using the classical approach. This type of item is illustrated in *Figure 34*.

Figure 34. **Correct answer response patterns where decisions vary**
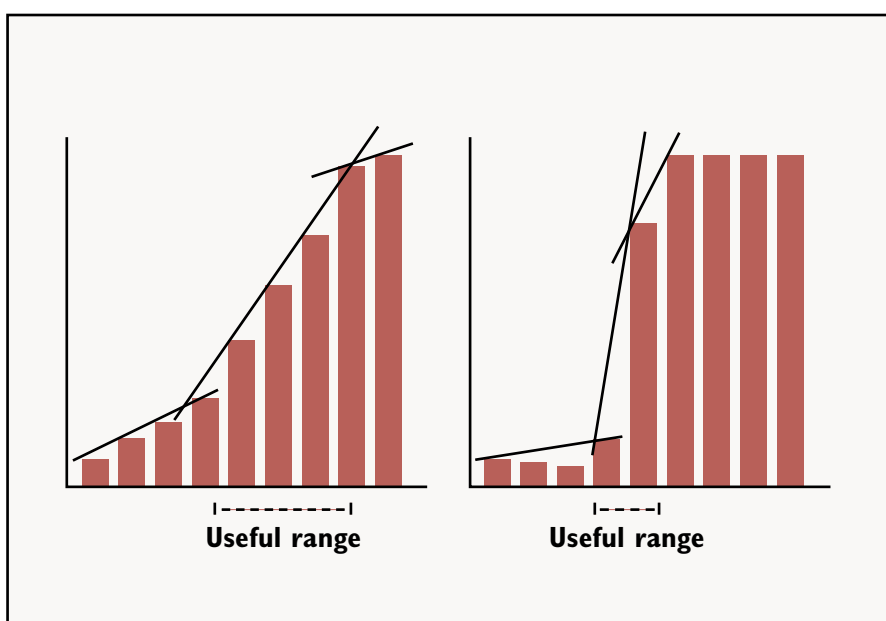
The first type of item is usually considered acceptable regardless of the analysis method. The second type of item is regarded as unacceptable by the item response modelling approach. Very steep gradients are regarded as inappropriate; such items are not useful over a reasonable range and often may be concerned with trivial content.
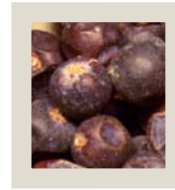
10 Maintenance of security

When trial tests are developed for secure purposes it is important that the secure nature of the tests be preserved. Copies of the tests for file purposes must be kept under lock and key conditions. The computer control files for the test analysis include the score key for each trial test so there has to be restricted access to the computers where the test processing is done.

The analysis reports (such as the item analysis, and the summary tables) will include the score keys and therefore those reports must be kept secure.
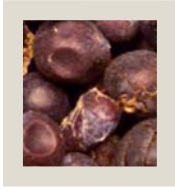
Test review after trials
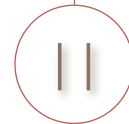
After the item analyses are complete, decisions have to be made about the items that will be retained, the items that will be modified, and the items that will be discarded. The test blueprint and associated specification grid must be consulted to ensure that enough items are retained to give a range of item difficulties within each cell of the grid.

An item may be **easy** because

- Wrong choices are not plausible;

- Most candidates know the work on which the items were based.

An item may be **difficult** because

- You have the wrong 'correct' answer;

- More than one answer is correct;

- The content is rare or trivial;

- The task is not well stated; and/or

- Candidates did not reach the item (other items may have been too complex, too lengthy, or too numerous).

An item may not **discriminate** because

- You have the wrong 'correct' answer;

- More than one answer is correct;

- The task is ambiguous;

- The 'correct' choice has flaw;

- The 'correct' choice is too obvious;

- The task is too difficult and candidates are guessing;

- The item is testing something different from the other items;

- The better candidates were taught the wrong information; and/or

- Only the weaker candidates were taught the topic because it was assumed (incorrectly) that able students already knew the work.

Cautions in interpreting item analysis data

Item analysis identifies questionable items which up until the trial stage had met our criteria for relevant, reasonably valid, and fair items. Item analysis may not necessarily identify faulty questions which should not have been included in the trial test because those criteria were not met. Some users of item analysis seek to reject all items but those with the very highest discrimination values. While this apparently gives increased reliability, this may be gained at expense of the validity of the final test. For example, a test of computation may have addition, subtraction, multiplication and division items. If items are progressively discarded through continued analysis it is likely that only one of the operations will remain (probably the one with the most items). The resulting test will be an apparently more reliable test but, because only one of the four operations is tested, it is no longer representative of all four processes, and hence not valid for the purpose of assessing the four processes.

Items which do not perform as expected can be discarded or revised. Test constructors should be aware of the possibility of distortion in the balance of questions when there are not enough items to satisfy requirements in all cells of the specification grid. If the original specification represents the best sampling of content, skills, and item formats, in the judgments of those preparing and reviewing the test, then leaving some cells of the grid vacant will indicate a less than adequate test. To avoid this possibility, test constructors may prepare three or four times as many questions that they think they will need for each cell in the grid. Test constructors have to avoid the tendency to test what is easy to test, rather than what is important to test.

## Assembling the final test and the corresponding score key

After trial, tasks may be re-ordered to take account of their difficulty. Usually the easiest questions are presented first. This is to encourage candidates to proceed through the test and to ensure that the weaker candidates do not become discouraged before providing adequate evidence of their achievements and skills. Minor changes to items may have to be made for layout reasons (for example, to keep all of an item on the one page of the test, or to avoid obvious patterns in the list of correct answers). Items representing a single cell within a test specification should vary in item content and difficulty. The position of the correct option in multiple choice items (A, B, C, D or E) should also vary and each position should be used to a similar extent. Some questions may have minor changes to wording, others may be replaced. The final test should be consistent with the test blueprint. The item review procedures described above are repeated (particularly important where stimulus material must be associated with more than one question) and each reviewer should work independently through the proposed test and

provide a 'correct' answer for each question. This enables the test constructor's (new) list of correct answers to be checked.

Preparation of final forms of a test is not the end of the work. The data from use of final versions should be monitored as a quality control check on their performance. Such analyses can also be used to fix a standard by which the performance of future candidates may be compared. It is important to do this as candidates in one year may vary in quality from those in another year. In some instances such checks may detect whether there has been a breach of test security.

It is customary to develop more trial forms so that some forms of the final test can be retired from use (where there is a possibility of candidates having prior knowledge of the items through continued use of the same test).

The trial forms should include acceptable items from the original trials (not necessarily items which were used on the final forms but in similar design to the pattern of item types used in the final forms) to serve as a link between the new items and the old items. The process of linking tests using such items is referred to as *anchoring*. Surplus items can be retained for future use in similar test papers.
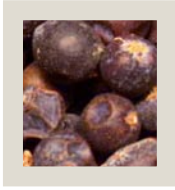
Confidential disposal of trial tests 12

It is usual to dispose of the used copies of trial tests by confidential destruction after a suitable time. [The 'suitable' time is difficult to define. Usually, trial tests are destroyed about one month after all analyses have been concluded and when the likelihood of further queries about the analyses is very low.]
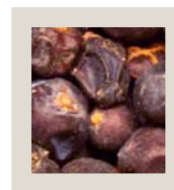
(13) Using item analysis software

In practice, test research and development agencies use item analysis software on a variety of computers to monitor the quality of their tests. Some useful software packages are listed in *Figure 35*. A • indicates that the software has the feature. Other software packages may provide similar coverage.

Figure 35. Coverage of item analysis software discussed in this module

Name	MS-Dos Version	Mac Version	Classical Analysis	Rasch (IRT) Analysis
QUEST	•	•	•	•
ITEMAN	•		•	
BIGSTEPS	•			•

Computer software

The **QUEST** computer program is published by The Australian Council for Educational Research Limited (ACER). Information can be obtained from ACER, 19 Prospect Hill Road, Camberwell, Melbourne, Victoria 3124, Australia.

The **ITEMAN** computer program is published by Assessment Systems Corporation. Information can be obtained from Assessment Systems Corporation, 2233 University Avenue, Suite 400, St Paul, Minnesota 55114, United States of America.

The **BIGSTEPS** program is published by MESA Press. Information can be obtained from MESA Press, 5835 S. Kimbark Avenue, Chicago, Illinois 60637, United States of America.

References

Adams, R.J. ; Khoo, S.T. (1993). QUEST: *The interactive test analysis system*. Hawthorn, Vic.: Australian Council for Educational Research.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmark Paedagogiske Institut.

Finding out more about trial testing and item analysis

1.   General strategies

Coffman, W.E. (1971). Essay examinations. In R.L. Thorndike (Ed.). *Educational Measurement*. (2nd Ed. (pp. 271-302). Washington, DC: American Council on Education.

Hake, R. (1986). How do we judge what they write? In K.L. Greenberg, H.S. Weiner, ; R.A. Donovan (Eds.), *Writing assessment: Issues and strategies*, (pp. 153-167). New York: Longman.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237-263.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R.L. Thorndike (Ed.) *Educational Measurement*. (2nd Ed.) (pp. 130-159). Washington, DC: American Council on Education.

Hopkins, C.D. ; Antes, R.L. (1990). *Classroom measurement and evaluation*. Itasca, Illinois: Peacock.

Hopkins, K.D. ; Stanley, J.C. (1981). *Educational and psychological measurement and evaluation*. (6th Ed.) Englewood Cliffs, NJ: Prentice-Hall.

Izard, J. (1991). *Assessment of learning in the classroom*. Geelong, Vic.: Deakin University.

Izard, J. (1995). *Module C.1 Overview of Test Construction*. Paris: International Institute for Educational Planning.

Low, B. ; Withers, G. (Eds.) (1990). *Developments in school and public assessment*. (Australian Education Review, No. 31). Hawthorn, Vic: ACER.

Mehrens, W.A. ; Lehmann, I.J. (1984). *Measurement and evaluation in education and psychology*. (3rd Ed.) New York: Holt, Rinehart and Winston.

Tinkelman, S.N. (1971). Planning the objective test. In R.L. Thorndike (Ed.) *Educational Measurement*. (2nd Ed.) (pp. 46-80). Washington, DC: American Council on Education.

Wright, B.D. ; Stone, M.H. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA Press.

Wright, B.D. ; Masters, G.N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

2. Broadening trial testing and item analysis strategies

de Lange, J. (1992). "Assessment: No change without problems". In M. Stephens ; J. Izard. (Eds.) *Reshaping assessment practices: Assessment in the mathematical sciences under challenge.* (pp. 46-76). Hawthorn, Vic.: Australian Council for Educational Research.

Griffin, P. ; Forwood, A. (1991). *Adult literacy and numeracy competency scales.* An International Literacy Year Project. Melbourne, Vic.: Phillip Institute of Technology.

Haines, C.R., Izard, J.F., Berry, J.S. et al. (1993). «Rewarding student achievement in mathematics projects». *Research Memorandum 1/93*, London: Department of Mathematics, City University. (54pp.)

Haines, C.R. ; Izard, J.F. (1994). "Assessing mathematical communications about projects and investigations". *Educational Studies in Mathematics, 27,* 373-386

Izard, J.F. (1991). "Issues in the assessment of non-objective and objective tasks". in A.J.M. Luitjen (Ed.), *Issues in public examinations.* (Proceedings of the 16th IAEA conference. Maastricht, The   Netherlands, 18-22 June 1990.) (pp73-83). Utrecht, The Netherlands: Lemma, B.V.

Linacre, J.M. (1990). *Modelling rating scales.* Paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA., USA, 16-20 April, 1990. [ED 318 803]

Wilson, M. (1992) Measurement models for new forms of assessment education. In M. Stephens ; J. Izard. (Eds.) *Reshaping assessment practices: Assessment in the mathematical sciences under challenge.* (pp. 77-98). Hawthorn, Vic.: Australian Council for Educational Research.

Applications of Item Analysis

Adams, R.J., Doig, B.A. ; Rosier, M.J. (1991). *Science learning in Victorian schools: 1990.* (ACER Research Monograph No. 41). Hawthorn, Vic.: Australian Council for Educational Research.

Doig, B.A., Piper, K., Mellor, S. ; Masters, G. (1994). *Conceptual understanding in social education*. (ACER Research Monograph No. 45). Melbourne, Vic.: Australian Council for Educational Research.

Masters, G.N. et al. (1990). *Profiles of learning: The basic skills testing program in New South Wales*, 1989. Hawthorn, Vic.: Australian Council for Educational Research.
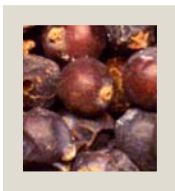
Exercises

1. Choose an important curriculum topic or teaching subject (either because you know a lot about it or because it is important in your country's education programme).

   • List the key content areas in that topic or subject.

   • List the important skills or behavioural objectives.

   • Show (in percentage terms) the relative importance of each key area.

2. Construct a test plan which has the content categories (from *Exercise 1*) at the left and the skills or behavioural objectives (also from *Exercise 1*) at the top. Adjust the numbers of items in each cell to reflect the percentage weightings you have chosen for each dimension.

3. Review an examination or test used in your country for the topic or teaching subject you chose in *Exercise 1*. Using your test plan as a guide, compare the examination or test with your test plan. Choose a topic in the curriculum which is not addressed by the examination or test and write some sample items to illustrate how item writers might satisfy this need.

4. Use the ITEMAN software to analyze the data given in *Figure 18*. Discuss the charateristics of each item in the test.