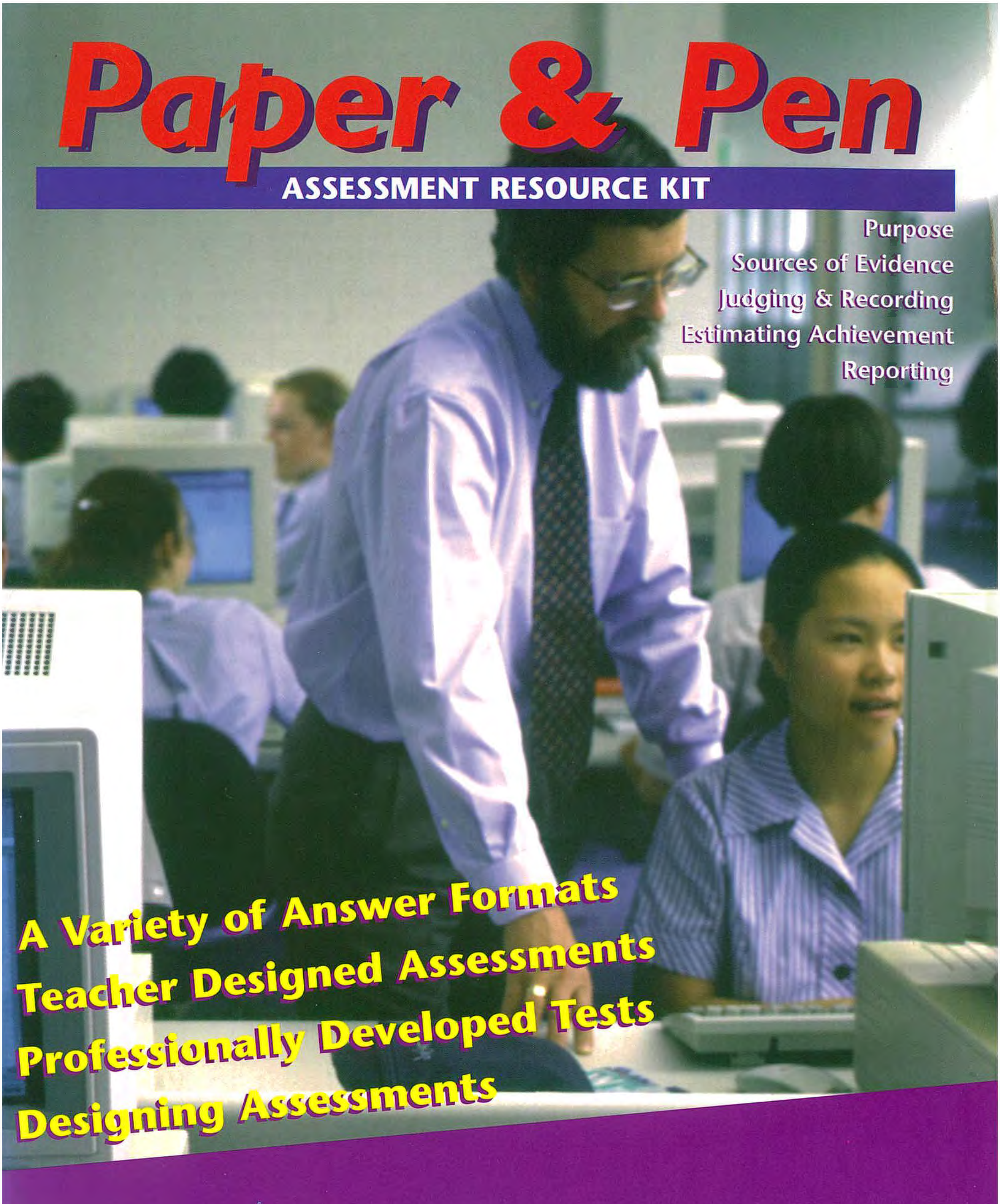


# Paper & Pen

## ASSESSMENT RESOURCE KIT

Purpose
Sources of Evidence
Judging & Recording
Estimating Achievement
Reporting

A Variety of Answer Formats
Teacher Designed Assessments
Professionally Developed Tests
Designing Assessments

ISBN 086431230-X
9 780864 312303

ARK

Margaret Forster & Geoff Masters

# Contents

First published 1999 by The Australian Council for Educational Research Ltd.
19 Prospect Hill Road, Camberwell, Melbourne, Victoria 3124, Australie.

This project was supported by a grant under the Curriculum Development Projects Program from the Commonwealth Department of Employment, Education,Training and Youth Affairs. The views expressed here are those of the authors and do not necessarily represent the views of the Department.

ISBN 0 86431 230 X

| | |
|---|---|
| *Designed by* | Susanna Murray |
| *Printed by* | Brown Prior Anderson (BPA) Pty Ltd |
| *Photographs by and courtesy of* | Symonetta & Yuri Kouzmin<br>Caulfield Grammar School<br>Sandringham Secondary College, Highett Campus |

# WHAT IS PAPER & PEN ASSESSMENT?

A common method of assessment is to present students with a series of questions or prompts and to use their written responses as evidence of individuals' levels of knowledge, competence, understanding or attitude. We refer to assessments of this kind as 'paper and pen' assessments.

Various kinds of questions have been developed as sources of evidence about achievement or attitude. These question types include—but are not limited to—multiple-choice questions, short answer questions, long answer questions, true-false questions, cloze items, essay questions, Likert-style questionnaire items, semantic differentials, and prompted self-reflections.

In answering questions, students do not always write their responses. Sometimes they colour bubbles and shapes, mark positions on lines, scratch shields covering printed text, and draw diagrams.

In fact, when questions of this kind were first used, students may not have used paper and pen at all. They may have recorded their responses on parchment or slate using crayons, pencil or chalk. And, with the advent of computer technology, questions of the kind discussed in this magazine may be presented on screens, with students answering through the keyboard and their responses (eg essays) being printed for assessment, or scored automatically. Nevertheless, because in most classrooms students currently respond to written questions and prompts on paper, we have chosen to continue to refer to assessment based on this range of question types as 'paper and pen' assessment.

The defining feature of the method of assessment discussed in this magazine is that students respond to a series of questions or prompts.

## Other assessment methods

Other methods of assessment also may involve the use of paper, pencils, crayons, pens or computers:

Student projects completed over a period of time and involving the collection, organisation, evaluation and presentation of material or data usually are more open-ended than the methods of assessment discussed here. Project assessment is addressed in the ARK magazine PROJECTS.

Collections of student work (eg portfolios of their writing) can be used as a basis for assessment. Portfolio assessment is addressed in the ARK magazine PORTFOLIOS.

Drawings and paintings done on paper are used in the assessment of achievement in art and technical drawing. The assessment of artwork produced by students is addressed in the ARK magazine PRODUCTS.

Computer-based simulations (eg flight simulations for the assessment of trainee pilots) we describe as performances. The assessment of student performances is addressed in the ARK magazine PERFORMANCES.

# Purposes and contexts

**Some purposes for paper and pen assessments**

- diagnosing students' strengths and weaknesses
- evaluating students' knowledge, skills and understandings
- certification
- selection
- monitoring standards

Paper and pen assessments are used for a variety of purposes including the diagnosis of students' special strengths and weaknesses, the evaluation of students' mastery of skills, selection into courses with limited intake, skill certification, and the monitoring of educational standards.

Contexts for paper and pen assessments include classroom formative assessment, classroom summative assessment, and 'external' summative assessment.

Formative assessment occurs in the course of everyday teaching. Assessments can be teacher-designed or purchased, professionally developed tests to inform the teaching program. Formative assessment is sometimes called 'low-stakes' assessment.

Classroom summative assessment occurs at the end of a course or semester to 'sum up' students' levels of achievement. These assessments also may include professionally developed tests. The audience for summative assessment results includes teachers, students, and parents.

External summative assessment occurs at the end of a course to determine, for example, which students will gain certification, be awarded scholarships, or be offered university places. Summative assessments of this kind usually are administered by an authority outside the institution in which the student is being taught. These assessments are sometimes called 'high-stakes' assessments.

A teacher who uses a professionally developed, standardised test to assess a student's spelling strengths and weaknesses in order to plan an individual spelling program is using paper and pen assessment for diagnostic purposes, in the context of classroom formative assessment. An example of diagnostic spelling items used for this purpose is shown on page 3.

A teacher who administers a professionally developed test to provide foreign language learners with information about their levels of achievement in a second language is using paper and pen assessment for evaluation purposes, in the context of classroom summative assessment. An example of an item from a reading test for students learning Indonesian as a foreign language is shown on page 4.

A university administering paper and pen assessments of moral reasoning as part of a screening instrument to select students into medical school is using paper and pen assessment for selection purposes in an external, high-stakes context. An example of a question from such an assessment is shown on page 4.

# Using paper and pen assessments for classroom formative assessment purposes

Teachers using the UK diagnostic spelling dictation assessment administer a graded dictation test to students.[1] They categorise students' responses word by word on a diagnostic 'grid'. Three spelling categories are used: 'plausible', 'invented', and 'random'.

Students whose spellings are largely random (spellings which bear no resemblance either in sound or structure to English) or invented (spellings which are invented from the sound of the word, with little or no reference to letter sequences in English) receive special tuition.

A section of 'Dictation Four' and a student's categorised responses are shown below. Dictation Four is designed to be suitable for children in Year 6 in England and Wales, P7 in Scotland and Y7 in Northern Ireland.

"A strange shape was approaching from the southern valley. The children standing there noticed an aeroplane of advanced design circling above. The machine touched down with precision in the rough mountainous region without even scratching the surface......"

| **PLAUSIBLE** | | **INVENTED** | **RANDOM** |
|---|---|---|---|
| **Readable**<br>Words conforming to English spelling which are readable in the context of the passage. | **Unreadable**<br>Words with some structural resemblance to the stimulus word, which may not be readable in the context of the passage. | Words invented from the sound of the word, with little or no reference to letter sequences in English. | Words that bear no resemblance, either in sound or structure, to English. |
| air o plane | stage (strange) | aposing (approaching) | sree (shape) |
| bass (base) | chider (children) | afase (advanced) | stome (southern) |
| | stading (standing) | roatch (rough) | vany (valley) |
| | niced (noticed) | mintinesne (mountainous) | dige (design) |
| | a (an) | evel (even) | mincing (machine) |
| | how (who) | chides (children) | then (touched) |
| | fight (flight) | saroudid (surrounded) | palhon (precision) |
| | dask (desk) | our off tube (altitude) | nesin (region) |
| | compter (computer) | faseing (freezing) | saires (scratching) |
| | childer (children) | telphecs (telephoned) | saish (surface) |
| | washed (watched) | rescusing (requesting) | putus (pilot) |
| | wine (when) | devied (delivered) | grised (explained) |
| | hard (heard) | ingmently (immediately) | damieshis (damaged |
| | | egrement (excitement) | scsisoyis (anxious) |
| | | | inffing (increasing) |
| | | | onfen (conditions) |
| | | | amding (alpine) |
| | | | specy (spare) |
| | | | cnopdy (component) |
| | | | ensed (electronic) |
| | | | resne (repairs) |
| | | | figeted (flight) |
| | | | plant (pilot) |
| | | | diepec (departed) |

## Using paper and pen assessments for classroom summative purposes

In the National Australia Bank Language Certificates program, Australian and New Zealand students complete a series of reading tasks at one of three levels: First, Beginners, or Intermediate.[2] Beginners level is designed for secondary school students who have studied the target language for between 80 and 200 hours. This summative assessment is administered by classroom teachers and leads to the award of certificates to encourage students in their learning of another language. Papers are marked by trained markers. Students receive certificates indicating their level of achievement.

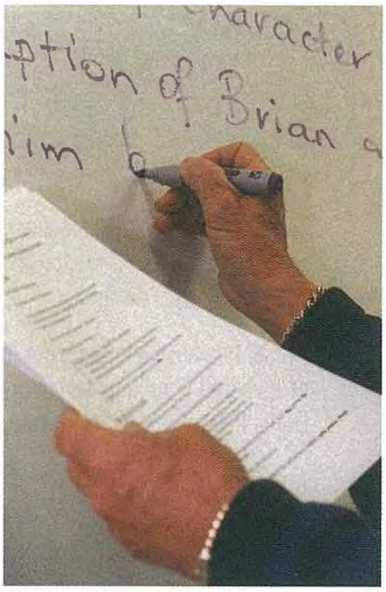
The question below is taken from a Beginners level paper in Indonesian.

Minangkabau Women's Dress

Kalau di Jawa para ibu memakai kain dan kebaya di Sumatera mereka biasanya memakai sarung dan kebaya panjang. Sarung mereka adalah 'kain songket' dengan sedikit benang perak atau emas.

'Kain songket' is pretty because it is decorated with

□ batik.
□ beads.
□ embroidery.
□ gold or silver thread.

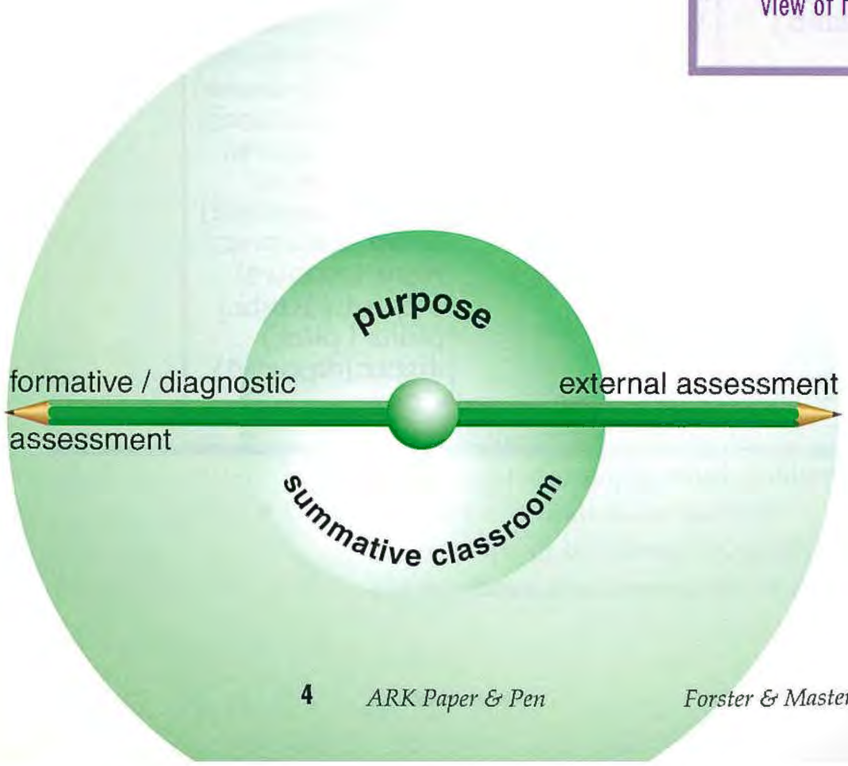


## Using paper and pen for external high-stakes assessments

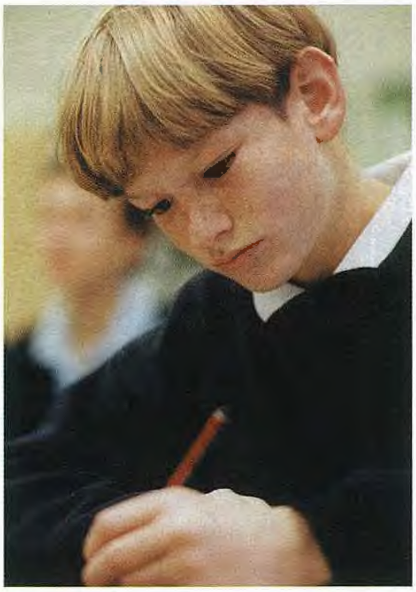
Obtaining high academic scores no longer guarantees students a place in medical studies at the Universities of Newcastle and Melbourne.[3] As well as achieving academically, students also need to do well on a 'personal profile' test in which they respond in writing to questions which demand considerations of ethics.

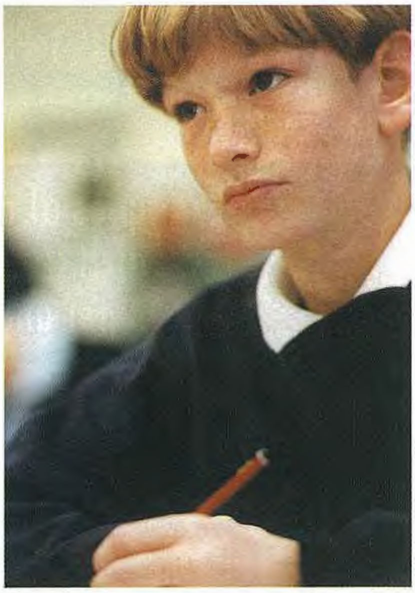
An example of a short answer question from the personal profile test is shown below.

A man steals money to buy his wife a dialysis machine she desperately needs. You are a doctor. What is your view of his crime?



It is useful to think about these different contexts and purposes for paper and pen assessment along a 'purpose' continuum, as shown here. Formative classroom assessment (including diagnostic assessment) is to the left of the continuum, 'high-stakes', external assessment, to the right. Summative classroom assessment is near the centre of the continuum.





## Developmental assessment

Developmental assessment is the process of monitoring students' progress though an area of learning so that decisions can be made about the best ways to facilitate further learning.

The backbone of developmental assessment is a progress map. A progress map describes the knowledge, skills, understandings, attitudes or values that students develop in an area of learning in the order in which they typically develop. The descriptions of progress along the map are sometimes called 'learning outcomes'.

In developmental assessment, students' progress in an area of learning is monitored against the learning outcomes described on a progress map.

Developmental assessment uses a range of methods for collecting evidence of student achievement, including portfolios, projects, products, performances and paper and pen assessments. Each method is appropriate for collecting particular kinds of evidence; that is, for addressing particular kinds of outcomes. For example, performance assessment is particularly useful for assessing 'doing' outcomes–outcomes for which it is necessary to observe students in action. Project assessment–the assessment of extended pieces of work completed over a period of time (involving the collection, organisation, evaluation and presentation of material or data)–is useful for assessing both subject specific knowledge and skills, and general skills such as report writing and working as a member of a group.

# Using paper and pen for developmental assessment

Paper and pen assessments are useful for collecting evidence about a wide range of learning outcomes in many curriculum areas. Examples of achievement outcomes which can be assessed using paper and pen assessments are shown in the example box on the right.

## Examples of achievement outcomes which can be addressed using paper and pen assessments

- understands and applies basic and advanced properties of functions and algebra (Mathematics)
- demonstrates competence in the general skills and strategies of the writing process (Language Arts)
- understands the promises and paradoxes of the second half of the 20th century (Global History)
- uses basic word processing, spreadsheet, database, and communication programs (Life skills)
- knows the kinds of forces that exist between objects and within atoms (Physical Sciences)
- understands aspects of substance use and abuse (Health) [4]

# Designing contexts for the assessment of learning outcomes

Paper and pen assessments can be used to assemble evidence about the outcomes on a progress map. Because outcomes usually are worded in a general way (eg 'writes stories or essays that show an awareness of intended purpose' [5]) it is possible to design an almost unlimited number of assessment tasks for the purposes of collecting evidence.

Assessment tasks provide specific contexts for collecting evidence about general outcomes. We use a student's performance in a specific context to collect evidence regarding a general outcome. That is, the particular contexts in which observations are made are never of interest in themselves: they are of interest only to the extent that they provide information about the dimension being assessed. As Ramsden et al (1993) note: 'Educators are interested in how well students understand speed, distance and time, not in what they know about runners or powerboats or people walking along corridors. Paradoxically, however, there is no other way of describing and testing understanding than through such specific examples.' [6]

Some examples of specific contexts for collecting evidence about general outcomes are shown on page 8.

The first example on page 8 illustrates a specific context for assessing the general outcome 'understands the relationship between speed, distance and time for moving objects'. In this task students are asked to demonstrate their understanding by analysing a graph showing distance run during a cross-country race against time taken.

The second example on page 8 illustrates a context for assessing students' understanding of how one additional score relates to the average of a set of scores. In this task students are asked to help 'Dr Maths' answer a perplexing number question. The answer to the question will assist 'Devious Dave' to avoid completing a piece of work.

The third example on page 9 illustrates a context for assessing the writing outcome 'produces texts in a fluent and legible style and uses computer technology to present these effectively in a variety of ways'. Students are asked to develop a written explanation about an aspect of a topic studied and to construct this explanation on computer screen.

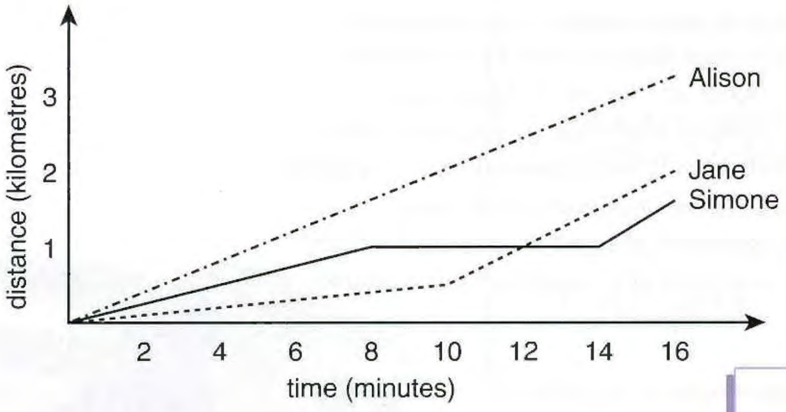
# Developing a context for addressing a general learning outcome

## Example 1: Science

Assessment tasks provide specific contexts for collecting evidence about general outcomes. This science assessment task addresses the outcome: 'describe the relationship between speed, distance and time for moving objects'.[7]

Context for observation: The cross-country run

The graph below shows the distance run by students during a cross-country race.



**1** Complete the table below to show how far each student travelled in 10 minutes.

| **Name** | **Distance run** |
|---|---|
| Alison | |
| Jane | |
| Simone | |

**2** For each student, describe the changes in speed during the run.

Alison: ________________________________________

________________________________________

Jane: ________________________________________

________________________________________

Simone: ________________________________________

________________________________________

**3** Plot your own distance-against-time graph for two pet owners walking their dogs in a park. Show and label the pet owners (a) moving slowly, (b) moving faster, and (c) at rest. Use the back of this sheet.

## Example 2: Mathematics

Assessment tasks provide specific contexts for collecting evidence about general outcomes. This mathematics task addresses the outcome: 'show an understanding of how one additional score relates to the average of a set of scores'.[8]

Context for observation:
Help Dr Maths answer this perplexing question:

Dear Dr Maths

All of our school maths projects are marked out of 20. After four projects, my average score is 13. If I can increase my average to at least 15 with a good mark on my 5th project, I won't have to do the 6th one. What kind of mark will I need to get? Please explain the mathematics to me as well.

Yours hopefully,

Devious Dave

Dear Devious Dave,

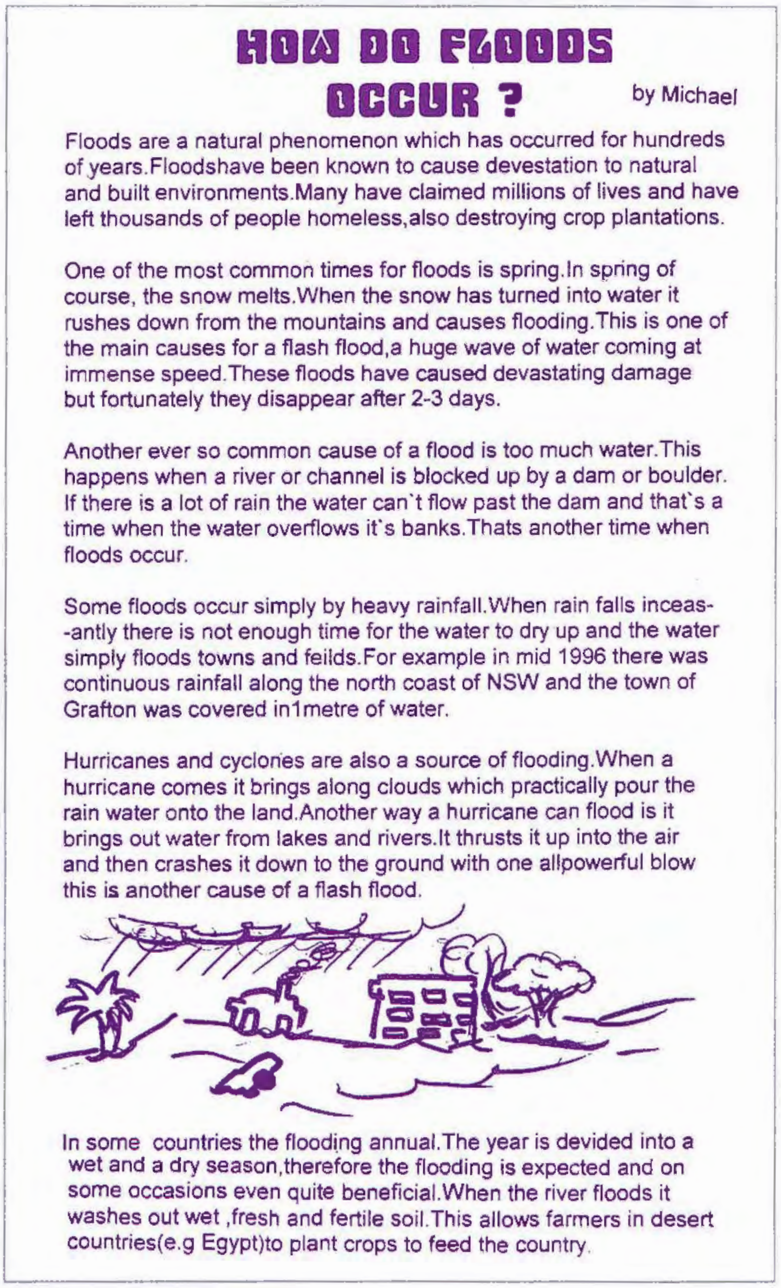
## Example 3: Writing

Assessment tasks provide specific contexts for collecting evidence about general outcomes. This writing task addresses the outcome: 'produces texts in a fluent and legible style and uses computer technology to present these effectively in a variety of ways'.

Context for observation[9]:

Students discuss the format and structure of an explanation in relation to a unit on 'water'. After researching a topic with a partner they construct an explanation independently on screen.

One student's final draft is shown below.



HOW DO FLOODS OCCUR ?

by Michael

Floods are a natural phenomenon which has occurred for hundreds of years.Floodshave been known to cause devestation to natural and built environments.Many have claimed millions of lives and have left thousands of people homeless,also destroying crop plantations.

One of the most common times for floods is spring.In spring of course, the snow melts.When the snow has turned into water it rushes down from the mountains and causes flooding.This is one of the main causes for a flash flood,a huge wave of water coming at immense speed.These floods have caused devastating damage but fortunately they disappear after 2-3 days.

Another ever so common cause of a flood is too much water.This happens when a river or channel is blocked up by a dam or boulder. If there is a lot of rain the water can`t flow past the dam and that`s a time when the water overflows it`s banks.Thats another time when floods occur.

Some floods occur simply by heavy rainfall.When rain falls inceas--antly there is not enough time for the water to dry up and the water simply floods towns and feilds.For example in mid 1996 there was continuous rainfall along the north coast of NSW and the town of Grafton was covered in1metre of water.

Hurricanes and cyclones are also a source of flooding.When a hurricane comes it brings along clouds which practically pour the rain water onto the land.Another way a hurricane can flood is it brings out water from lakes and rivers.It thrusts it up into the air and then crashes it down to the ground with one allpowerful blow this is another cause of a flash flood.

In some  countries the flooding annual.The year is devided into a wet and a dry season,therefore the flooding is expected and on some occasions even quite beneficial.When the river floods it washes out wet ,fresh and fertile soil.This allows farmers in desert countries(e.g Egypt)to plant crops to feed the country.

# A variety of answer formats

A variety of answer formats can be used in paper and pen assessments. For example, students' understandings can be assessed through multiple-choice items, short answer questions, essays and concept maps, and their attitudes and values through Likert-style questionnaires.

Some answer formats are better suited to the assessment of particular kinds of outcomes than others. Examples of outcomes and examples of paper and pen answer formats that might be used in their assessment are shown below.

## Examples of answer formats that might be used to assess selected outcomes

| **Outcome** | **Answer format** |
|---|---|
| Produces texts clearly, effectively and accurately, using the sentence structure, grammatical features and punctuation conventions of the text type (English, Australia) | essay |
| Interprets and compares displayed information (Mathematics, Ontario) | short answer |
| Uses abstract ideas about physical phenomena in explanations (Science, United Kingdom) | short answer |
| Identifies events that affect balance in an ecosystem (Science, Australia) | multiple-choice |
| Demonstrates discernment on ethical issues and recognises the need for truthfulness and integrity (Values, Western Australia) | Likert-style questionnnaire |

# Judging and recording students' responses

Depending on the answer format used for gathering evidence of achievement, responses to paper and pen tasks may be machine-scored, scored by hand, or judged using provided marking guides or rating scales.

Two important considerations in the judgement of students' responses are:

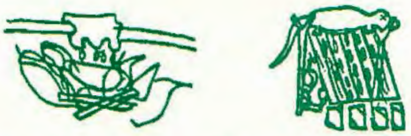
- the method of judgement; and
- the comparability of judgements.

## Partial credit scoring

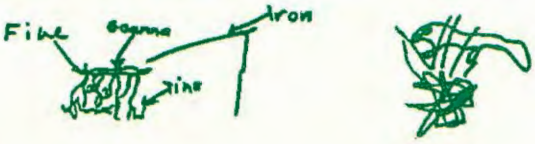
Partial credit scoring allows markers to recognise and record students' varying levels of success—their partially correct understandings and strategies. In this example, markers credit students with 1 score point if they demonstrate partial understanding of the described process. They are assessed for their reading comprehension, not drawing skill.[10]

Marconi got the goanna fat from people who lived in the bush. One way they used to extract the fat was to roost the goanna over a fire. They put a piece of corrugated iron on a slight slant next to the fire. The hot liquid fat ran down the grooves in the iron and was collected at the bottom in tins.

Draw a picture showing this method.



**2 points** goanna cooking over fire and attempt to show collection of fat in a plausible fashion



**1 points** goanna cooking over fire only or implausible collection

**0 points** drawing showing neither of above eg fire or goanna only

## Method of judgement

Some paper and pen answer formats (eg multiple-choice items and true/false items) do not require markers to use their judgement when assessing students' answers. A judgement about the correct answer has been made during the construction of the question. Other answer formats require markers to judge the quality of students' responses using scoring guides or rating scales.

Scoring guides sometimes are used for judging the quality of students' responses to short answer questions. For example, partial credit scoring guides allow markers to recognise and record students' varying levels of success—their partially correct understandings and strategies. An example of a partial credit scoring guide for a reading comprehension question is shown on this page.

Rating scales usually are used to assist markers to make judgements of extended responses. In using rating scales, markers judge the quality of student work against specified criteria. Markers make holistic judgements of work when they make a single rating based on an overall impression of the work. They make analytic judgements when they rate different aspects of the work. A piece of writing, for example, might be rated for the quality of ideas as well as for control over the surface features of the writing (spelling, punctuation, and grammar). An example of a holistic rating scale for the assessment of open-ended science items is shown on page 12. An example of an analytic rating scale for a narrative sequel is shown on page 13.

## Holistic rating scale

In using rating scales, markers judge the quality of student work against specified criteria. Markers make holistic judgements of work when they make a single rating based on an overall impression of the work.

*Scoring rubric for Connecticut Academic Performance Test (CAPT) science open-ended items.*[11]

**3**

The response is an excellent answer to the question. It is correct, complete, and appropriate, and contains elaboration, extension, and/or evidence of higher-order thinking and relevant prior knowledge. There is no evidence of misconceptions. Minor errors will not necessarily lower the score.

**2**

The response is a proficient answer to the question. It is generally correct, complete, and appropriate, although minor inaccuracies may appear. There may be limited evidence of elaboration, extension, higher-order thinking and relevant prior knowledge, or there may be significant evidence of these traits, but other flaws (eg inaccuracies, omissions, inappropriateness) may be more than minor.

**1**

The response is a marginal answer to the question. While it may contain some elements of a proficient response, it is inaccurate, incomplete, and/or inappropriate. There is little if any evidence of elaboration, extension, higher-order thinking, or relevant prior knowledge. There may be evidence of significant misconceptions.

**0**

The response, although on topic, is an unsatisfactory answer to the question. It may fail to address the question, or it may address the question in a very limited way. There may be no evidence of elaboration, extension, higher-order thinking, or relevant prior knowledge. There may be evidence of serious misconceptions.

## The need for comparability of judgements

In the classroom context, whether another teacher would make the same judgement of a student's piece of work using the same marking guide usually is not of great concern. In high-stakes contexts, however, inter-marker reliability is crucial.

Usually, the greater the requirement for comparability, the more tightly the assessment criteria are specified. Markers are trained and the marking process is carefully monitored to ensure a high level of inter-marker agreement. In some high-stakes settings products are assessed by several judges.

## Analytic rating scale

In using rating scales, markers judge the quality of student work against specified criteria. Markers make analytic judgements when they rate different aspects of the work.

*Scoring guide for writing a sequel* [12]

| | Criteria | Not shown | Low | Med | High |
|---|---|---|---|---|---|
| 1 | The story is a logical continuation of the original. | □ | □ | □ | □ |
| 2 | The setting is vividly described. | □ | □ | □ | □ |
| 3 | The characters are well described. | □ | □ | □ | □ |
| 4 | There is a dialogue throughout the story. | □ | □ | □ | □ |
| 5 | There is a logical plot including exposition, rising action, climax, falling action, resolution. | □ | □ | □ | □ |
| 6 | The piece is mechanically correct. | □ | □ | □ | □ |
| 7 | The piece is especially interesting, creative and engaging. | □ | □ | □ | □ |

# Estimating and reporting locations on a progress map

Developmental assessment requires an on-balance decision (inference) about a student's location on a progress map based on the available evidence. The way in which this inference is made usually is determined by the purpose of the assessment: the higher the stakes, the greater the requirement for comparability, and the more tightly the 'inference process' is likely to be specified.

In developmental assessment, teachers monitor student progress against a pre-constructed map of developing knowledge, skills and understandings (see *ARK Developmental Assessment*). Teachers make observations relevant to the outcomes on a progress map and use these observations as 'evidence' to estimate students' levels of achievement.

Developmental assessment requires an on-balance decision (inference) about a student's location on a progress map, based on available evidence. Paper and pen assessments provide one kind of evidence. The way in which this evidence is used to infer a level of attainment usually is determined by the purpose of the assessment: the higher the stakes, the greater the requirement for comparability, and the more tightly the 'inference process' is likely to be specified.

## Subjective estimates

When the estimate of a student's level of achievement is made subjectively, there may be only a loose connection between the available evidence (judgements and records of performances on paper and pen tasks) and the resulting estimate of the student's location on a progress map.

## Objective estimates

In high-stakes situations, where high levels of inter-marker comparability are desirable, the way in which records and judgements of students' performance on paper and pen tasks are used to estimate a student's location on a progress map may be tightly prescribed to ensure that the estimate is made objectively. The inference may be made numerically on the basis of a marker's pattern of judgements across carefully defined criteria, for example.

# Multiple sources of evidence

In developmental assessment, paper and pen assessments often are used in conjunction with other assessment methods to provide multiple perspectives on students' understandings. The extract on page 15 is taken from a teacher support document. It provides a model for designing assessment tasks which address curriculum focus and outcomes in a particular unit of work. The unit on 'plate tectonics' makes explicit the relevant curriculum and learning outcome focus. Notice that the assessment tasks include a group presentation (performance assessment), a model with written explanation (product assessment) and a test of concepts (paper and pen assessment). In making a judgement of a student's level of achievement on a progress map, it might be possible to use evidence from all three sources.

# Following articles

The following articles explore paper and pen assessment in more detail.

The first article describes the variety of paper and pen assessments. What are the different forms of paper and pen assessments and for what purposes are they used? The second article explores teacher designed paper and pen assessments:

- considerations in planning paper and pen assessments, including issues of validity, reliability and fairness;
- the strengths and weaknesses of different paper and pen formats;
- how to write good paper and pen assessments;
- how to judge and record students' responses;
- how to estimate students' levels of attainment; and
- how to report achievement.

The third article explores common considerations in the development and use of professionally designed paper and pen tests including:

- issues in writing fair questions;
- procedures for standardising test conditions;
- procedures for standardising marking; and
- ways to interpret test results.

The final article provides an overview of the paper and pen design process. A checklist summary details steps in the process.

# Using paper and pen assessments in conjunction with other assessment methods

This extract comes from a teacher support document. It provides a model for designing assessment tasks which address curriculum focus and outcomes in a particular unit of work.[13] Paper and pen assessment (a test of concepts) is one method of assessment suggested.

**Sample of designing assessment tasks which address curriculum focus and outcomes in a particular unit of work.**

Science level 6: Plate tectonics

| **Summary of relevant curriculum focus** | **Relevant learning outcomes** | **Assessment tasks** |
|---|---|---|
| *At this level a student is involved in:*<br>Focusing on causes of folding and faulting in relation to plate and tectonic theory; investigating types of folds and faults in rocks and causes of earthquakes and volcanoes; evidence of theory of plate tectonics. | **The changing Earth**<br>• Discuss folding and faulting and associated landforms<br>• Discuss the evidence that supports the theory of plate tectonics. | **1** Students develop a puzzle map of Gondwana, at 120 million years ago, which splits up into the current southern land masses; depicting and explaining the associations between continental drift, matching rock types and species distribution.<br>**2** Students build a working model of a fold or fault, depicting and explaining how these phenomena occur. |

| **What do I set up?** | **What do I assess? In what strands?** | **How will I assess?** |
|---|---|---|
| ***Procedure***<br>**1** Where are they? volcanoes, earthquakes, ridges and trenches.<br>**2** Mid-Atlantic Ridge sea-floor spreading exercise<br>**3** Fossil evidence for plate tectonics<br>***Resources provided***<br>**1** World map<br>**2** Paper, scissors<br>**3** Geological time scale and examples of fossil types (actual or photo), trilobites, brachiopods, etc. | ***Earth and beyond***<br>• Recognition of types of folding and faulting<br>• Understanding of causes of folding and faulting, volcanic and earthquake activity in relation to plate tectonic theory<br>• Understanding of the use of fossil evidence to support plate tectonic theory | • Group presentation of model and explanation of fold or fault<br>• Model and written explanation of Gondwana<br>• Test of concepts of fossil deposition, folding, faulting and plate tectonics |

[1] Peters, M. and Smith, B. (1993) *Spelling in Context Strategies for Teachers and Learners,* Berkshire: NFER-Nelson pp. 70 and 79.

[2] Australian Council for Educational Research ( 1998) National Australia Bank Language Certificates Beginners Level Indonesian, Camberwell: ACER

[3] Australian Council for Educational Research

[4] Kendall, J. and Marzano, R. (1996) *Content Knowledge: A Compendium of Standards and Benchmarks for K -12,* Aurora CO: Education Mid-continent Regional Laboratory

[5] Kendall, J. and Marzano, R. (1996) p. 294.

[6] Ramsden, P., Masters, G., Stephanou, A., Walsh, E., Martin, E., Laurillard, D., & Marton, F. (1993). Phenomenographic research and the measurement of understanding: An investigation of students' conceptions of speed, distance and time. *International Journal of Educational Research,* 13, 1993, 301-316.

[7] Andrews, C. et al (1998) *Effective Assessment for Science,* Carlton & South Melbourne: Board of Studies and Addison Wesley Longman Australia Pty Limited p. 115.

[8] Beesey, C. et al *Effective Assessment for Mathematics,* Carlton & South Melbourne: Board of Studies and Addison Wesley Longman Australia Pty Limited p. 99.

[9] Board of Studies NSW (1998) *English K-6 Work Samples* p. 143.

[10] Forster, M., Mendelovits, J. & Masters, G. (1994) *Developmental Assessment Resource for Teachers DART English,* Camberwell: Australian Council for Educational Research.

[11] Hibbard, M. et al (1996) *A Teacher's Guide to Performance-Based Learning and Assessment,* Alexandria: Association for Supervision and Curriculum Development Alexandria Virginia p. 124.

[12] adapted from Hibbard, M. et al (1996) p. 63.

[13] Board of Studies Victoria (1996) *Using the CSF Assessment and Reporting,* Carlton: Board of Studies p. 33.

# A VARIETY OF ANSWER FORMATS

The defining feature of paper and pen assessment is that students respond to a series of questions or prompts. Responses to questions or prompts provide evidence of students' knowledge, skills, understandings or attitudes.

A variety of answer formats can be used with paper and pen questions and prompts. For example, students can be asked to respond to test questions measuring knowledge, skills and understanding using multiple-choice, short answer or long response formats. They can be asked to respond ro questionnaires measuring attitudes and values using open-ended items, Likert scales and semantic differentials.

This article describes a range of common paper and pen answer formats and discusses the kind of information each is designed to collect. Formats are presented in alphabetical order.

Examples of paper and pen answer formats

- cloze
- concept maps
- essay
- investigations
- matching items
- multiple-choice
- Likert-style questionnaires
- self-reflections
- short answer
- written retell

## Cloze procedure

Cloze procedure requires students to complete a text from which words have been deleted.

In traditional cloze, the introductory section of a text is left untouched to set the context for the reader. Then every fifth or tenth word is removed. In modified cloze, a particular word or phrase is removed to assess, for example, comprehension or knowledge of syntactical structures of English.

In the modified cloze example on the left, reading comprehension is assessed.

**Example: assessing students' comprehension of text**

**Extract from text;**[1]

'Snake!' Tony's mother yelled.

It was the first day of the family's camping holiday.

With one hand his mother grabbed up the baby from the grass by the tent flap. With the other she seized the stick of the beach umbrella. 'Keep back!' she called as Tony ran towards the tent.

Tony laughed and ran into the tent. 'It's only a lizard, Mum. I saw it walking through the grass.'

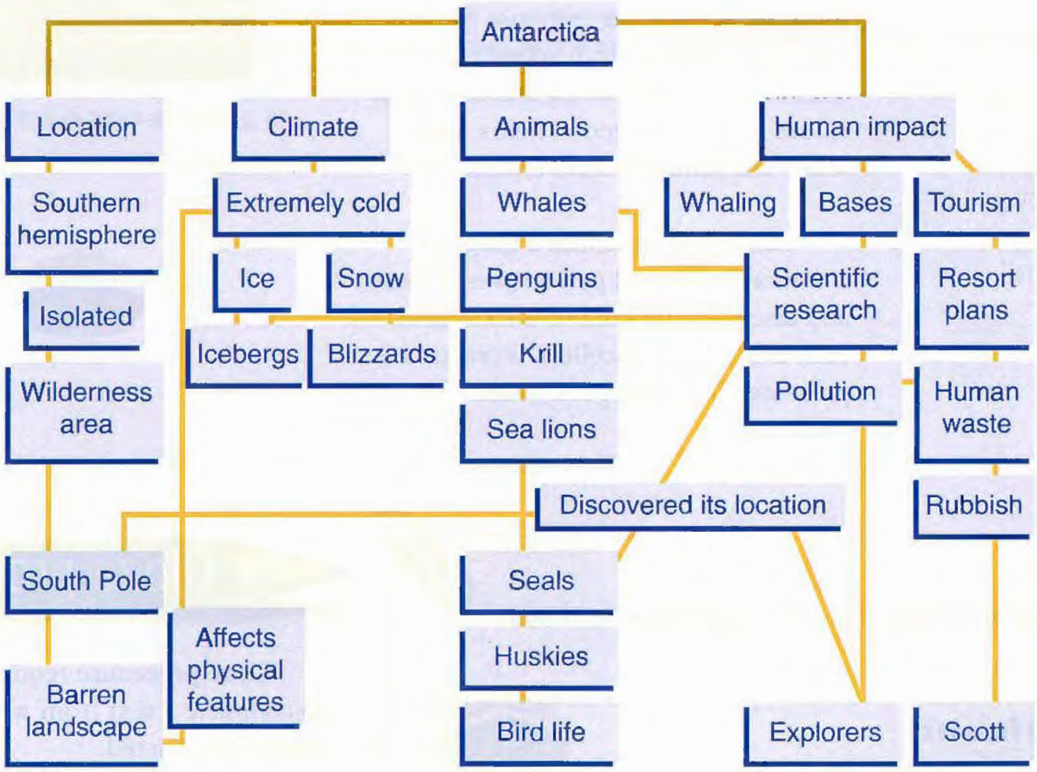
It was the first day of the family's camping holiday. Something crawled into the tent and ................................ thought it was a snake. Mother grabbed up the baby because she thought the snake might ................................ . Tony laughed because he knew the animal was a .................................. .

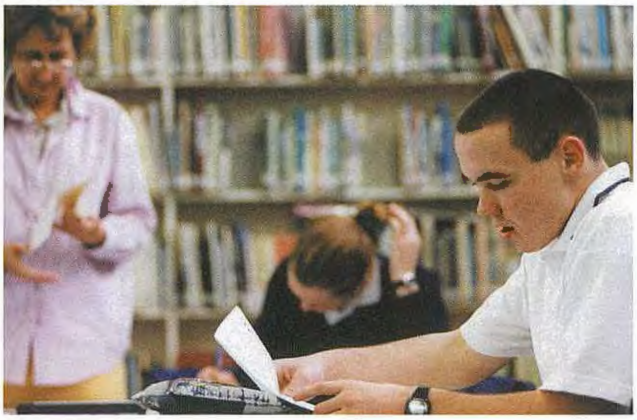
# Concept maps

Concept maps are diagrammatic representations of key concepts which relate to an area of investigation. The relationships between the concepts are shown with lines or arrows. Other names for concept maps include mind maps, semantic webs, and graphic organisers.

Concept maps are used to assess the level and complexity of students' understandings. Scoring criteria usually include the number of ideas generated, the classification of information, and the quality of conceptual links. An example of a concept map on Antarctica is shown below. [2]

**Example: assessing conceptual links**





## Constructed response:

see Short answer format (page 25)

# Essay format

Essays are long responses which students make to a given topic or 'prompt'.

Essay format is used to assess students' ability to construct a considered response; for example, to develop a logically constructed argument or a controlled narrative with setting, character, problem and resolution.

Essay prompts can be structured (eg 'briefly explain', 'describe three methods for'...) or open-ended in their formulation. Examples of structured and open-ended essay prompts are shown on the left.[3]

**Example: assessing a logically constructed argument**

**Structured essay item:**

Select one of the following and discuss the environmental problems created:

**a** draining used oil on the ground

**b** pouring used oil down the drain

Keep your answer to no more than a half a page.

**Open-ended essay item:**

Critically evaluate the impact of anti-discrimination legislation on Australian workplaces.

# Investigations

Investigations are open-ended tasks that require students to apply familiar concepts and skills in unfamiliar situations. Students completing paper and pen investigations provide written reports of their findings.

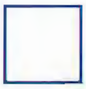
Investigations often are used to assess students' mathematical understandings, problem solving ability, and use of mathematical language. An example of a mathematics investigation is shown on the right.[4]

## Example: assessing mathematical understanding

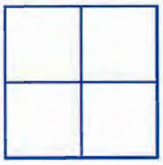
Investigations are open-ended tasks that require students to apply familiar concepts and skills in unfamiliar situations. This example can be used to assess students' mathematical understandings, problem solving ability and use of mathematical language.

### Use pattern blocks to investigate

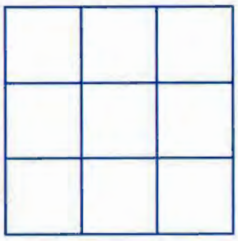
Use pattern blocks to investigate square, larger square, larger square.



Square



Larger Square



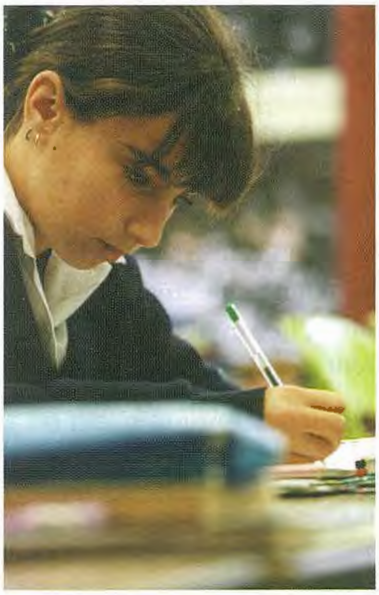
Larger Square

What patterns can be found?

What about other pattern block shapes?

## Likert scales

Likert scales are commonly used as the answer format in questionnaires designed to assess opinions, attitudes, values and interests. A Likert scale consists a set of ordered categories–usually Strongly Disagree, Disagree, Agree and Strongly Agree. Respondents record their responses to each question using these provided alternatives.



Students responding anonymously to the questionnaire designed by John XXIII College in Perth, Western Australia, to assess students' levels of 'conscience' record their responses to each statement using a 4-point rating scale 'Strongly Disagree', 'Disagree', 'Agree', 'Strongly Agree'. The extract below shows some of the statements on the 'Conscience' scale.[5]

| | **SD** | **D** | **A** | **SA** |
|---|---|---|---|---|
| I would feel bad if I had stolen something. | □ | □ | □ | □ |
| If I found a wallet with identification, I would try to find the owner. | □ | □ | □ | □ |
| If my friends were planning to steal, I would try to talk them out of it. | □ | □ | □ | □ |
| I would rather do my own work poorly than cheat and do well. | □ | □ | □ | □ |

## Matching items

Matching items consist of two lists: a list of premises and a list of responses. Directions explain the basis on which a match between the two lists is to be made.

Matching items often are used to assess the ability to recall knowledge and to understand relationships; for example, to recall terms and their definitions, or to understand principles and illustrations.

Directions: Column A lists features and Column B lists planets. Match the features with the planets by placing the appropriate letter in the space to the left of the number in Column A.[6]

A planet name may be used more than once.

| **Column A** | **Column B** |
|---|---|
| **1** closest to the sun | **a** Earth |
| **2** farthest from the sun | **b** Jupiter |
| **3** largest planet | **c** Mars |
| **4** many moons | **d** Mercury |
| **5** smallest planet | **e** Neptune |
| | **f** Pluto |
| | **g** Saturn |
| | **h** Uranus |
| | **i** Venus |

# Multiple-choice format

Multiple-choice items contain a question or incomplete statement followed by a series of alternatives from which students select the answer.

Sometimes the alternatives provide the student with several plausible answers and the right answer ('key') is the most plausible of the alternatives. Sometimes the alternatives to the key are wrong, although they will appear to be plausible to some students.

Multiple-choice items often are used to assess knowledge of factual information or literal comprehension. Examples 1 and 2 on this page are examples of multiple-choice questions of this kind. However, multiple-choice items also can be constructed to assess complex reasoning skills and sophisticated understandings. In Example 3 on the next page, students need to reason logically to answer the items. In Example 4, on page 24, students need to draw an inference from text.

One criticism made of multiple-choice items is that they do not require students to generate an answer and are therefore 'inauthentic' and invalid. This very feature, however, can allow the assessment of subtle understandings before students are able to express those understandings in writing.

Another criticism made of multiple-choice items is that they do not require markers to use judgement when assessing students' answers. This is true. A judgement about the correct answer has been made during the construction of the item.

### Example 1: assessing knowledge

The correct answer (key) is asterisked.

**19 + 122 =**

□ **131**

□ **141***

□ **312**

□ **212**

### Example 2: assessing literal comprehension

The correct answer (key) is asterisked.

My name is Anita. I am eight. I was born in Canada. I came to New Zealand when I was six. I have one cousin who lives in Wellington and another who lives in Auckland. I like holidays and the beach. I don't like school.

Where was Anita born?

□ Auckland

□ New Zealand

□ Wellington

□ Canada *

### Example 3: assessing logical reasoning

The correct answers (keys) are asterisked.

### Question 1

Tuan wants to make a cake and tries to remember the recipe. However, his first attempt at making the cake is not a success.

He tries again, this time adding two new ingredients: a teaspoon of bicarbonate of soda and an egg. There are no other changes to the recipe or the method he used in his first attempt. This time the cake is a success.

Which one of the following is the best conclusion about the success of the second attempt? [7]

The success is due to the adding of

□ either the bicarbonate of soda or the egg.

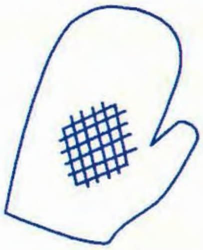
□ both the bicarbonate of soda and the egg together.

□ either the bicarbonate of soda or the egg, or both together. *

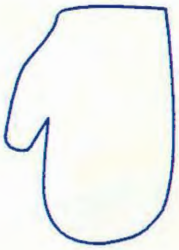
□ neither the bicarbonate of soda nor the egg.

### Question 2

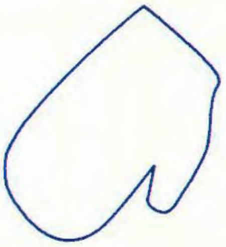
The three oven gloves shown below are lying on a table. Each of the three gloves has a pattern on the back (upper) surface only. Only the back of glove I can be seen.



I



II



III

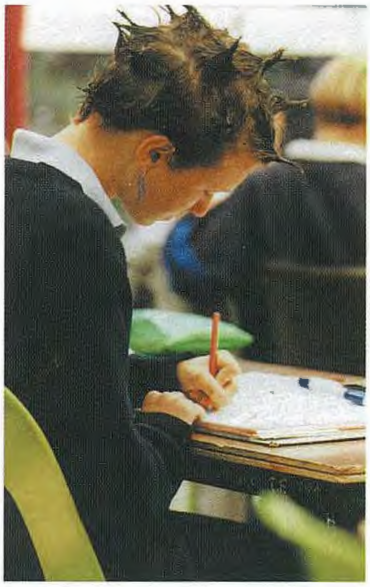
Which of the following pairs of gloves would consist of right- and left-hand gloves? [8]

□ I and II only *

□ I and III only

□ I and II, and II and III

□ I and III, and II and III



### Example 4: assessing inferential understanding

The correct answers (keys) are asterisked.

**Question 1**

The ninth and tenth lines refer to the woman's performance for Rubinstein[9]:

'Once she played
for Rubinstein, who yawned.'

This reference suggests that

□ the woman's musical talent was limited. *

□ the woman could have been a great musician.

□ Rubinstein underestimated the woman's talent.

□ the woman treasures that moment as the most fulfilling in her life.

**Question 2**

When you hear hoofbeats, think of horses before zebras.[10]

*Anonymous medical saying*

This saying warns doctors that

□ the most common occurrence is the most likely cause. *

□ symptoms of disease are not always as they appear.

□ the best diagnoses require breadth as well as depth of thinking.

□ they should be aware of all symptoms before making a diagnosis.

## Retells

Retells require students to relate, in their own words, something they have viewed, read or heard. Students completing paper and pen retells provide their response in writing.

Retells are used primarily to assess students' comprehension–the quality and organisation of ideas. Written retells also provide an opportunity to assess students' control of the surface features of language (grammar, spelling, and punctuation).

## Short answer format

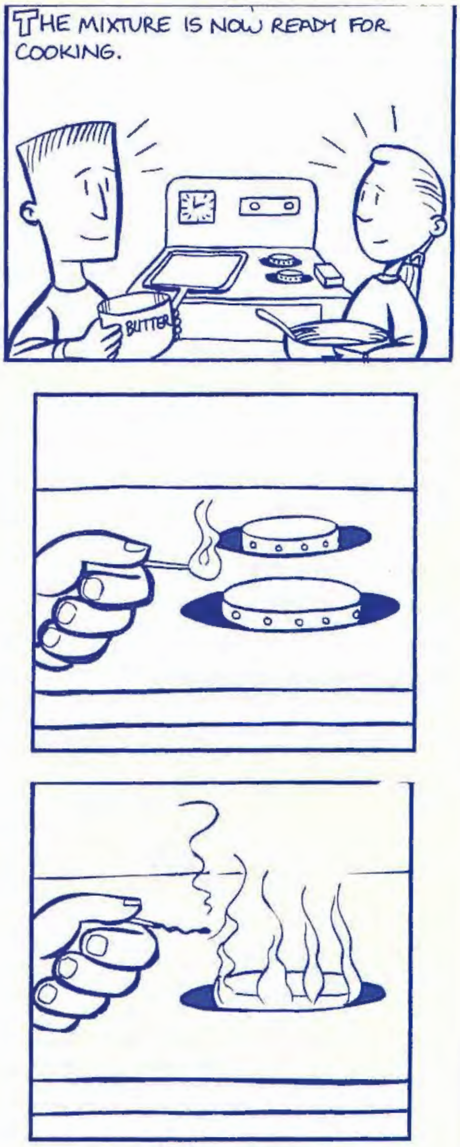
Short answer assessments, or 'constructed-response' items, require students to generate a brief answer. The answer may be to a question, or may require the completion of a sentence.

Short answer items can be 'closed' or 'open-ended'. Closed items have one acceptable answer only. Open-ended items have more than one correct answer.

Short answer items are used to assess students' conceptual understandings as well as factual knowledge. Example 1 assesses students' conceptual understandings in science. Students' responses demonstrate different levels of understanding of chemical reactions and awareness of conservation of matter. Example 2 is an open-ended item allowing for several alternative readings of the text.

### Example 1: assessing levels of conceptual understanding [11]

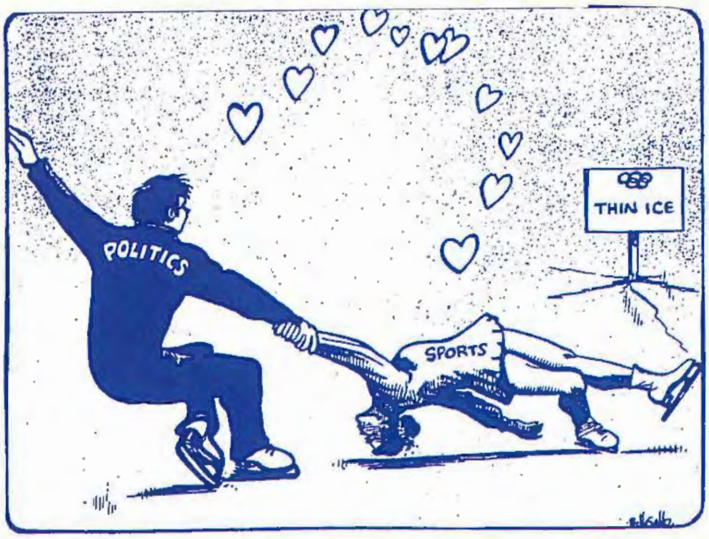
Student responses demonstrate different levels of understanding of chemical reactions and awareness of conservation of matter.



Where does the wood go when the match burns?

Does a match have the same weight before and after it burns? Explain your answer.

### Example 2: assessing multiple readings of text[12]



'At the time [1980 Moscow Olympics], this cartoon appeared in the New York Daily News. What do the various features of the cartoon convey about the mixing of sport and politics?

Make sure you reveal the central point the cartoonist is making.

# Self-reflections

Self-reflections are students' reflections on their own work and progress.

Reflections often are used as a curriculum device to assist students to negotiate specific challenges and to make them conscious of their own knowledge, skills, and thinking processes. They are also useful for self-assessment. Teachers sometimes assess students' self-reflections (see page 47 for an example of a marking guide for the assessment of students' self-assessments). An example of a guided self-reflection is shown to the right.[13]

**Name:**________________________________

**Date:** ____/____/____

Description or title of work: ____________________

About how much time did you spend on this work? __________

Where was it done?

□ mostly in school
□ mostly out of school
□ about half and half

Did anyone help you on this work? __________________

What is the one thing you did well in this work? __________

What made that part of the work so good? ______________

If you could change one thing about the work, what would you change?

________________________________________

Why would you change this? ______________________

Did you like doing this work?

□ yes
□ no
□ somewhat

1 Mossenson, L., Hill, P., and Masters, G. (1987) *Tests of Reading Comprehension* (TORCH), Camberwell: Australian Council for Educational Research.

2 Clarke, J. *Monitoring Learning in English,* Brisbane: Mount Coot–tha School Support Centre p. 83.

3 Clayton, B. (1995) *Focussing on Assessment,* Adelaide: National Centre for Vocational Education Research Ltd p. 139.

4 Holcroft, L. and Coates, R. *Monitoring Learning in Mathematics,* Queensland: Mount Coot-tha School Support Centre p. 15.

5 questionnaire developed by the Australian Council for Educational Research Camberwell Australia for John XXIII College, Western Australia.

6 Clayton, B. (1995) p. 124.

7 Victorian Board of Studies (1998) Victorian Certificate of Education General Achievement Test.

8 Victorian Board of Studies (1997) Victorian Certificate of Education General Achievement Test.

9 Queensland Core Skills Test (1997) *Core Skills Test Multiple Choice Questions II,* Queensland: Board of Senior Secondary Studies.

10 Victorian Board of Studies (1996) Victorian Certificate of Education General Achievement Test.

11 Doig, B. and Adams, R. (1993) *Tapping Students' Science Beliefs,* Camberwell: The Australian Council for Educational Research.

12 Queensland Core Skills Test (1997) *Core Skills Test Paper 3 Short Response Items,* Queensland: Board of Senior Secondary Studies.

13 adapted from Hibbard, M. et al (1996) *Performance-based Learning and Assessment,* Virginia: Association for Supervision and Curriculum Development p. 204.

# TEACHER DESIGNED PAPER AND PEN ASSESSMENTS

Teachers sometimes ask students to complete paper and pen tasks with the intention of using their responses to assess individuals' levels of knowledge, skill, understanding, or attitude. Examples of planned paper and pen tasks designed by teachers include tests consisting of multiple-choice, short answer, and essay questions, and questionnaires designed to assess values and attitudes.

This article looks at teacher designed paper and pen assessments. Assessments of this kind differ from the informal observations teachers make of students engaged in writing activities in the course of everyday teaching and learning. They differ also from paper and pen assessments designed by professional test developers for use in high-stakes contexts where special efforts are made to ensure the comparability of assessments from student-to-student, assessor-to-assessor, and school-to-school (see pages 60-61).

Teacher designed paper and pen assessments are used to collect information that teachers would be unlikely to see in passing, and to collect information efficiently. Tasks are planned in advance to provide evidence of particular outcomes in a learning area, and teachers decide in advance on an answer format and method for judging students' work.

Teachers use paper and pen assessments to collect information for many purposes, including to diagnose individual students' strengths and weaknesses, to inform the teaching process at a whole class level, and to infer students' levels of achievement for reporting to parents.

When assessing for diagnostic or formative purposes teachers sometimes focus assessment tasks on a particular aspect of knowledge or skill. For example, 'Does the student understand subtraction of two 2-digit numbers when regrouping is required?' The teacher needs to know whether the particular concept has been grasped and whether additional teaching is required. The purpose is to draw a conclusion about the student's understanding of a particular aspect of subtraction to inform the teaching process.

In contrast, when the purpose of the assessment is to report a level of achievement in an area of learning, teachers use records of student performances on tasks addressing a range of closely related skills in order to infer a level of achievement. For example, a conclusion about  a student's level of achievement in 'Number' would require the assessment of a wider range of understandings than subtraction of two 2-digit numbers.

# Planning paper and pen assessment

The usefulness of planned paper and pen tasks as a source of evidence about student achievement or attitude depends on:

- how well the tasks address the instructional goals, or the outcomes of the learning area (relevance–a validity issue);
- the amount of evidence provided (a reliability issue);
- how well the tasks allow students to show their strengths and weaknesses; and
- how fair the tasks are to students from different language, cultural and socio-economic backgrounds.

## Addressing outcomes

The purpose of paper and pen assessment determines the range of outcomes addressed. Diagnostic assessment is likely to focus on a narrow range of knowledge and understanding. Assessment for reporting levels of achievement in a broader domain will address a wider range of learning outcomes.

In developmental assessment, teachers collect evidence of students' achievement using assessments that address outcomes on a progress map. Paper and pen assessments are useful for collecting evidence about a wide range of learning outcomes in many curriculum areas. Some examples of achievement outcomes that can be addressed using paper and pen assessments are shown on page 6.

When planning paper and pen tasks, it is important first to be clear about the purpose of the assessment. When the assessment is for diagnostic or formative purposes, a narrow range of outcomes and understandings might be addressed.

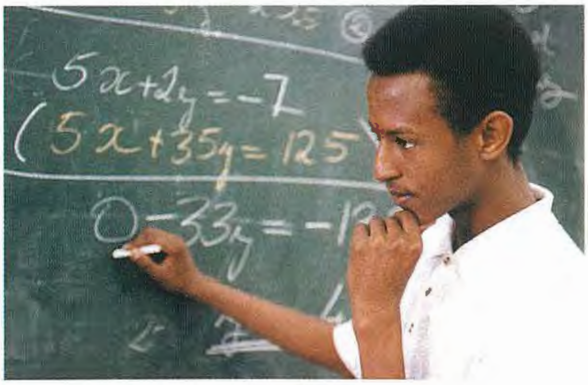
When the assessment is for summative purposes–to infer a student's level of achievement in an area of learning–it is important that the tasks address the full range of valued outcomes. That is, the tasks need to provide curriculum coverage. If the tasks provide limited coverage of outcomes, then inferences about student achievement will be limited. Assessing a limited range of outcomes also may send an unintended message to students about what is valued.

Whatever the purpose of the assessment, it is essential that there is a sufficient number of tasks from which to draw an inference about a student's knowledge and understanding. The more opportunity a teacher has to observe what a student knows and can do, the more reliable will be their judgement of that student's underlying ability. This principle applies whether the underlying ability is a narrowly defined domain such as 'subtraction' or a more broadly defined domain such as 'number' ability. Observations of student achievement need to occur on a number of occasions.

When planning paper and pen tasks it is important to be clear about the outcomes being addressed. Tasks then can be focused to provide information, and appropriate answer formats can be selected. Below are some examples of outcomes and examples of answer formats that might be used in their assessment.

| Outcome | Answer format |
|---|---|
| Writes multi-paragraph informational pieces developing a problem and a solution when appropriate to the topic.<br>(Pennsylvania, Reading, Grade 5) [1] | essay |
| Reflects critically in meanings and values associated with particular visual art works.<br>(Australia, Visual Arts, Level 8) [2] | long response, essay |
| Identifies decimal parts of a whole<br>(Ontario, Mathematics, Grade 3) [3] | multiple-choice |
| Lists the ways materials are used for different Purposes<br>(Australia, Science, Level 2) [4] | short response |
| Participates in the political process and contributes to community services consistent with good citizenship<br>(Western Australia, Values) [5] | questionnaire |

It is essential also to consider the validity of the inference. Teachers draw an inference about more general skills and understandings from students' responses to specific tasks. When an inference is drawn about a level of achievement across a set of outcomes in a domain, teachers need to consider the validity of this aggregation. For example, it may be valid to draw an inference about a student's level of achievement in writing from a number a tasks which address a range of writing outcomes. However, it may not be valid to draw an inference about a student's level of achievement in English by aggregating across outcomes in reading, viewing, speaking, listening, and writing. Achievement in these learning areas my need to be reported separately.



# Selecting an answer format

Once the purpose of the assessment is clear and the outcomes to be addressed are identified, the next step is to consider the answer format. Some answer formats are better suited to the assessment of particular kinds of outcomes than others. For example, if a learning outcome requires students to demonstrate understanding of the relationship between historical events, or to structure a logical argument, then an essay format might be most appropriate. If the outcome requires students to demonstrate knowledge of dates, then a multiple-choice format might be used.

When selecting among possible answer formats, it is sometimes helpful to focus on the verb in the outcome statement being addressed. For example, an outcome that requires students to 'identify', or 'recall' could be addressed using multiple-choice answer format. An outcome that requires students to 'compare' or 'evaluate' might be better addressed by an answer format that requires students to generate a response–essay or short answer format.

When selecting among possible answer formats, it is sometimes helpful to focus on the verb in the outcome statement being addressed. Below are some examples of verbs commonly used in outcome statements and examples of answer formats that might be used in their assessment.

| **Outcomes which require students to** | | **Possible answer format** |
|---|---|---|
| calculate<br>distinguish<br>identify<br>determine | recall<br>select<br>define | multiple-choice |
| state<br>define<br>determine<br>classify | identify<br>calculate<br>describe<br>name | short answer<br>cloze |
| compare<br>evaluate<br>contrast<br>translate<br>develop | analyse<br>interpret<br>discuss<br>plan | essay<br>investigation |

# Planning fair tasks

**Considerations in planning fair tasks include ensuring:**

- clear focus
- accessibility of language
- inclusivity of tasks
- no tricks
- review of tasks

It is important that paper and pen tasks are designed to be fair. Individual tasks or questions must allow students to demonstrate what they know and are able to do. In designing paper and pen tasks, a number of considerations are important:

## Clear focus

Assessment of knowledge and skills in one area should not be unduly influenced by students' knowledge and skills in another area. For example, mathematics problems that require a large amount of reading may prevent students with low reading ability from demonstrating their mathematical understandings. If tests are administered on computer, students with poor keyboard skills may be unfairly disadvantaged.

## Accessibility of language

The language used in instructions and task introductions should be as simple and clear as possible. The language used in the question should not exceed the language required by the task. For example, a multiple-choice comprehension question should not be more difficult to read than the stimulus passage.

Unless knowledge of a particular term is being assessed, questions should be written in simple language. Students can fail to answer a question correctly because they fail to understand the language of the question. Changing a single word in a multiple-choice question (eg 'non-aqueous solvent' to a 'liquid other than water') can reduce the difficulty of an item.

## Inclusivity

Tasks need to be set in contexts which are accessible to students of different gender, cultural and language backgrounds. For example, a series of mathematics problems set in the context of motorbike and car racing may not engage girls well enough to allow them to demonstrate their mathematical understandings.

## No tricks

It is important that students are not unfairly tricked by tasks. For example, ambiguously worded essay prompts should be avoided unless the ambiguity is intentional. A student should not be penalised for a divergent reading of an ambiguous prompt. Negatively worded multiple-choice items should be avoided: 'Which of the following is not an accurate representation of the problem?' Students often miss the 'not' unless it is highlighted or underlined.

## Review

The development of fair tasks usually takes time, and requires feedback and refinement. Ideally, tasks are developed well in advance of the time they are to be used. This allows time for review before administration. Even the most experienced test developers benefit from having their items reviewed by a panel of peers. The task of the panel is to review the objectives the tasks are addressing, to look for errors or ambiguities in the tasks and to suggest improvements. An example of a rating checklist for peer review of assessment designs is shown below.

When tests are used over a period of time with different groups of students they also can be refined on the basis of students' responses.

**Considerations in planning valid, reliable and fair tests include:**

- selecting range of items
- selecting tasks of different levels of difficulty
- ordering items carefully
- ensuring item independence
- clear delivery
- adequate time for completion

## Design rating checklist for peer review

This extract comes from a checklist developed for peer review of assessment designs, by the Center on Learning, Assessment and School Structure. Peers evaluate the design against three criteria: credibility of design, instructional worth of design, and user friendliness of design.[6] The indicators for the 'instructionally worthy' criterion are shown here. A yes/no judgement is first made for each indicator. On the basis of these judgements, an on-balance decision about achievement of the criterion is made on a scale of 1 to 4, where 4 is the highest level of achievement.

Assessment design rating checklist for peer review

□ yes □ no Does the task require learnings at the heart of the curriculum?

□ yes □ no Is the task worthy of the time and energy required to complete it well?

□ yes □ no Is the task challenging—an apt 'stretch' for students?

□ yes □ no Will the feedback to students en route enable them to self-assess and self-adjust?

□ yes □ no Will the students likely be able to verify resultant scores and use feedback to improve later performance?

**INSTRUCTIONALLY WORTHY OVERALL RATING** ① ② ③ ④

# Valid, reliable and fair tests

In planning how to group a series of tasks into a valid, reliable and fair test, a number of considerations are important:

## Selecting the range of items

To be valid, a test needs to include a range of items addressing the relevant range of outcomes. One way to check the coverage of a test is to draw up a grid of the kind shown below. This grid shows items developed for a statistics test. To check outcome coverage, items are categorised according to the kind of outcome addressed: 'recall of facts', 'computation skills', and 'understanding'.

## Including tasks of different levels of difficulty

Most tests need to include tasks which are appropriate for students of different levels of ability. That is, they need to include tasks of different levels of difficulty. Where long tasks are included, they need to be open enough to be accessible to students of different levels of ability to allow them to demonstrate their knowledge, skills and understandings.

## Ordering items

As far as possible, the easiest questions should appear first in a test so that weaker students do not become disheartened by difficult items at the beginning of a test and have the opportunity to demonstrate what they know and can do.

## Ensuring independence of items

All items in a test need to be independent. The possibility of answering one question correctly should not depend on an understanding which has been formed in a previous question. Failure on one item should not influence a person's success or failure on any other item.

## Range and balance of items

One way to check the outcome coverage of a test is to draw up a grid. This grid shows items developed for a statistics test. Items are categorised according to the kind of outcome addressed: 'recall of facts', 'computation skills', and 'understanding'. Most of the items in this test address 'understanding'. [7]

| | | **Objectives** | | |
|---|---|---|---|---|
| **Content** | **Recall of facts** | **Computation skills** | **Understanding** | **Totals** |
| Frequency distribution | 1 item | | 2 items | 3 |
| Means | 1 item | 2 items | 1 item | 4 |
| Variances | 1 item | 2 items | 1 item | 4 |
| Correlation | 2 items | 2 items | 6 items | 10 |
| Relative standing | 2 items | | 4 items | 6 |
| **totals** | 7 | 6 | 14 | 27 |

### Delivery

Paper and pen assessments need to be delivered in a format that does not unfairly disadvantage particular students. Typed tests are preferable to hand written tests, which are sometimes difficult for students to read.

### Time

The time allowed should permit 90 to 95% of students to complete the set of tasks.

## Constructing assessments—some examples

Different paper and pen answer formats provide for different approaches to assessing what students know, can do, and value. As well as the general considerations in developing paper and pen assessments discussed above, there are considerations particular to each answer format.

Guidelines for constructing assessment tasks using eight common paper and pen answer formats are described below.

Formats are presented in alphabetical order:

- cloze procedure
- concept maps
- essay format
- investigations
- Likert scales
- multiple-choice
- short answer format

# Cloze procedure

Decide on the outcomes to be addressed. For example, cloze format may be used as a vehicle for assessing reading comprehension, the understanding of scientific concepts, or knowledge of cohesive ties in the construction of text. Select an appropriate text and leave sufficient introductory text untouched to set the passage context.

Omit the words or phrases which address the outcomes to be assessed. Consider each omission carefully to see how text cohesion is disrupted and to see how much forward or backward referencing is required to maintain meaning. Use only one or two gaps per sentence. Avoid specific clues like 'an' before a gap.

Provide sufficient answer space and make sure that all spaces are equal in length.

Ensure that the instructions to students include reading the whole passage before beginning to fill in the spaces, reading past the spaces to get context clues from surrounding text and rereading the passage after completion to check that the passage makes sense.

Remind students that more than one word may have been omitted if that is the case.

An example of a cloze test is shown below. This task has been designed to assess students' understanding of cohesion through the use of pronouns to refer to established participants in a text.

## Constructing assessments

This extract comes from a cloze task designed to assess students' understanding of cohesion through the use of pronouns to refer to established participants in a text.[8]

There was once a frog who dreamed of being someone special, someone brave and noble. ________________ wanted to be king, but ________________ knew that such a thing could never happen to ________________ while ________________ lived in a lily pond. So ________________ ventured forth into the world to find a princess. For, like all frogs, ________________ had heard the story of the princess whose kiss of love had changed a frog into a handsome prince.

# Concept maps

Decide on the area of knowledge to be mapped.

Ensure that students understand the purpose of concept mapping: to produce a diagrammatic representation in which related concepts are linked by times, arrows, and structural organisation.

Have students complete the map in four steps:

- brainstorm to identify vocabulary related to the given concept;
- classify the words hierarchically from general to specific;
- draw lines or arrows between related or interrelated concepts; and
- add labels to clarify and link the concepts.

# Essay format

Decide on the outcomes to be addressed. For example, if the subject area is English, will you use the essay format to provide evidence of students' control of text features (the quality of ideas and sense of audience and purpose), and/or language features (control of spelling, punctuation and grammar)?

Make the prompt as clear as possible, ensuring that students are directed towards demonstrating the outcomes you are addressing. For example, if you are assessing students' control of the text features of narrative form, make sure the prompt directs students to develop a narrative and not a recount or an exposition.

Ensure that the prompt does not demand knowledge or skills that are not important to your purpose. For example, if you are assessing students' abilities to construct a logical argument, make sure the prompt does not require knowledge outside their experience. Year 6 students may be able to construct an argument addressing the question 'Should pet cats be kept inside at night?' but not be able to construct an argument about the merits of private health cover. If the outcomes to be addressed do require students to demonstrate a knowledge base, make sure the task requires students to show command of essential knowledge.

Ensure also that the prompt is open enough to allow students of different cultural and language backgrounds and different levels of ability to engage with the task.

Make sure students understand what is being assessed. For example, if you tell students that you are assessing only their understanding of scientific concepts, then it is not ethical to deduct marks for poor spelling, punctuation, and handwriting.

Test the question by writing an ideal answer to it.

An example of an essay prompt is shown on the left This task is designed to address science outcomes in the 'working scientifically' and 'life and living' strands of the Australian profiles.

## Constructing assessments

The following task is designed to address science outcomes in the 'working scientifically' and 'life and living' strands of the Australian profiles:

'Explains how living things obtain, store and transport nutrients, transform energy and manage wastes' ('life and living', level 6);

'Selects ways to present information that clarify patterns and assist in making generalisations'.[9]

**Plant nutrition**

People often say that plants get their food from the soil. Do you agree with this or not? Explain carefully what you think, paying special attention to what you mean by food.

# Investigations

Decide on the outcomes to be addressed.

Ensure that the topic is structured sufficiently to direct students to address the outcomes you want to assess and that the task is appropriate for the time allowed. Ensure that the context is familiar, interesting, and engaging.

Good topics, while providing structure, are open enough to allow students to carry out the investigation in a variety of ways and to allow students to provide multiple solutions.

# Likert scales

Decide on the outcomes–opinions, attitudes, values or interests–to be addressed.

Decide on the set of ordered categories to be used, for example: Strongly Disagree, Disagree, Agree, Strongly Agree. Try to have an even number of categories to prevent 'fence sitters' from selecting the centre category.

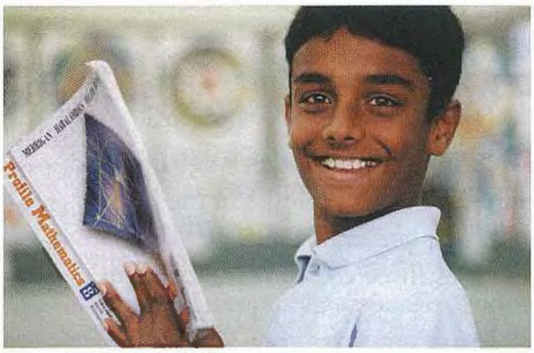
Word statements as clearly and specifically as possible. Statements that are very general encourage the response 'it depends'.

If the purpose of the questionnaire is to gather classroom level data, have students respond anonymously. They are then more likely to respond honestly.

An example of a student survey designed to provide teachers with information about some major aspects of the classroom experience from the students' point of view is shown on page 37. A teacher using this questionnaire would consider responses to each item separately.

This student survey is designed to provide teachers with information about aspects of the classroom experience from the students' point of view.[10] Students could respond anonymously to a survey of this kind. Responses to each item are considered separately and not combined to obtain a measure of students' attitudes.

| **In this class I feel** | **1** | **2** | **3** | **4** | **5** | |
|---|---|---|---|---|---|---|
| Important | □ | □ | □ | □ | □ | Ignored |
| Comfortable | □ | □ | □ | □ | □ | Uncomfortable |
| Involved in lessons | □ | □ | □ | □ | □ | Restless, bored |
| Part of a team | □ | □ | □ | □ | □ | Alone |
| Good about work | □ | □ | □ | □ | □ | Bad about work |
| Sure where I stand | □ | □ | □ | □ | □ | Not sure where I stand |
| **The teacher has been** | | | | | | |
| Prepared | □ | □ | □ | □ | □ | Unprepared |
| Fair | □ | □ | □ | □ | □ | Unfair |
| Helpful | □ | □ | □ | □ | □ | Unhelpful |
| Well organized | □ | □ | □ | □ | □ | Locking organization |
| Clear about what's expected | □ | □ | □ | □ | □ | Unclear about what's expected |
| Sensitive to my needs | □ | □ | □ | □ | □ | Insensitive to my needs |
| Fully engaged and excited | □ | □ | □ | □ | □ | Seemingly bored |
| Knowledgable | □ | □ | □ | □ | □ | Not on top of the subject |
| Able to make difficult ideas accessible and interesting | □ | □ | □ | □ | □ | Over our heads |
| **Our work has generally been** | | | | | | |
| Thought provoking | □ | □ | □ | □ | □ | Dull |
| Effective in helping me learn | □ | □ | □ | □ | □ | Ineffective in helping me learn |
| Too fast | □ | □ | □ | □ | □ | Too slow |
| Too easy | □ | □ | □ | □ | □ | Too hard |
| Too much the same | □ | □ | □ | □ | □ | Too unpredictable |
| Too abstract | □ | □ | □ | □ | □ | Too simplistic |
| Too little | □ | □ | □ | □ | □ | Too much |



# Multiple-choice

Decide on the outcomes to be addressed and design each item to address a learning outcome.

## Stem

The stem (question or incomplete statement) should focus students on what is being asked. Students should have some idea about what is being asked before they read the options.

✗

Next Sally
□ had dinner.
□ went home.
□ went to bed.
□ read a book.

✓

What did Sally do next?
□ had dinner.
□ went home.
□ went to bed.
□ read a book.

To minimise the amount of reading required, as many words as possible should be put in the stem.

✗

Jim went to
□ the river with his friends.
□ the park with his friends.
□ the beach with his friends.
□ the movies with his friends.

✓

Jim and his friends went to the
□ river.
□ park.
□ beach.
□ movies.

Avoid using superfluous language in the stem. 'One of the colours in the painting is...' is preferable to 'The painting uses three colours. One of them is...'

State the stem in a positive form wherever possible.

## The alternatives

If the alternatives complete the stem to form a sentence, then each alternative should be concluded with a full stop. If the alternatives form complete sentences on their own, they should be punctuated accordingly. If the alternatives are single words or phrases, no punctuation is required.

For example:
The sign in the picture is showing the way to the
□ café.
□ shop.
□ toilets.
□ creche.

What did the father say to his son when he became angry?
□ 'How could you do that?'
□ 'Go to your room at once!'
□ 'When you behave like that I feel really angry.'
□ That's the last time I invite you to the football.'

How will the winner be notified?
□ by letter
□ by phone
□ on television
□ in the newspaper

The answer must be correct and not simply the best of the alternatives presented. The answer must not stand out among the alternatives because of its length, wording, or other superficial quality. When putting together a series of multiple-choice questions for a test, check to see that the correct answer is not always in the same position in the set of alternatives.

## Constructing assessments

This multiple-choice item addresses the outcome: 'Identifies processes of energy transfer and conditions that affect them'.[11] Notice that the distractors are designed to diagnose students' conceptual misunderstandings.

The amount of energy dissipated each second in a circuit depends on

□ the voltage only.<br>
□ the current only.<br>
□ the current and the voltage.<br>
□ the resistance and the voltage.

Distractors (incorrect alternatives) must be indisputably incorrect, while being reasonable and plausible. Any distractor which is absurdly wrong reduces the number of real choices available to the student, and contributes nothing to the item.

Distractors should have different meanings from each other.

The meaning of one alternative should never be subsumed by another as in 'She took: an apple, a jumper, some sunscreen, some food'.

Alternatives should be arranged from shortest to longest, longest to shortest, or in some logical order.

Use four or five alternatives to reduce the chance that a student will select the right answer by guessing.

Avoid using 'all of the above' or 'none of the above'.

Sometimes the distractors can be designed to gather information about students' conceptual misunderstandings, as in the example above. This item has been designed to address the outcome 'Identifies processes of energy transfer and conditions that affect them'.

## Short answer format

Decide which outcomes are to be addressed. Ensure that each question addresses an outcome. Decide, as the questions are being developed, whether there is only one correct answer or multiple acceptable responses. If there are multiple acceptable responses, decide whether a partial credit scoring guide is to be used.

## Judging and recording

Considerations in developing criteria for judging the quality of students' responses include

- addressing outcomes
- clarity
- fairness
- usability

Teachers use a variety of methods to judge and record the quality of students' responses to paper and pen tasks. Some paper and pen answer formats (eg multiple-choice items and true/false items) do not require teachers to use their judgement when they assess students' answers. A judgement about the correct answer has been made during the construction of the question. Other formats, however, require teachers to judge the quality of the answer.

In developing criteria to judge the quality of students' responses, a number of considerations are important. The criteria need to address the outcomes being measured. They need to be clear to students, parents and other teachers. They need to be fair and not reflect variables over which students and teachers have no control. And they need to be usable. That is, for any one task, the number of criteria on which the work is to be judged needs to be limited or the marking will be too time consuming and onerous.

Methods for judging and recording the quality of students' written responses include:

- scoring guides;
- diagnostic categories;
- analytic ratings; and
- holistic ratings.

## Scoring guides

Methods for judging and recording the quality of students' written responses include:

- scoring guides;
- diagnostic categories;
- onolytic ratings; and
- holistic ratings.

Scoring guides are used for judging the quality of students' responses to short answer questions. Sometimes responses are scored dichotomously (right/wrong). Sometimes responses are scored in a way that allows teachers to recognise and record students' varying levels of partial success–their partially correct understandings and strategies. Scoring of this kind is called 'partial credit' scoring.

The first step in developing scoring guides it is to be clear about the outcomes being addressed by the task. Assessment criteria used to judge students' responses then can be developed. For example, if students' conceptual understandings in science or their reading comprehension are being assessed, then assessment criteria probably will not include correct spelling.

Scoring guides are most useful when the criteria for assessment are supported by examples of students' correct and incorrect responses. An example of a dichotomous scoring guide for a mathematics question and an example of a partial credit scoring guide for a viewing comprehension question are shown on the next page.

## Judging and recording — Dichotomous scoring

Below is an example of a dichotomous scoring guide for a mathematics question. The question comes from a unit designed around a map of a zoo. In this question, students' use of the language of position and direction is being assessed. Note the examples of students' responses used to support the scoring guide.[12]

**Question:**

Explain how to get from the Rail Gate to the Seal enclosure.

**Scoring guide:**

**1 adequate and correct detail given—must be from RAIL gate, and direction of turn must be correct**

eg:
- Go straight and turn right at the cockatoos.
- Go straight down past the telephone to the cockatoo and turn right.
- Go past the toilet and the picnic pavilions. Keep going past the lions, bandstand, small cats and then turn at the first left.

NB Mark as correct an answer that uses the convention of north at the top of the page ie 'go south and then turn west'.

**0 incorrect directions**

eg:
- Go past the zoo school and turn left.

## Judging and recording — Partial credit scoring

The following example of a partial credit scoring guide is for a viewing comprehension question. Students view a short video and then answer a series of questions. This question assesses students' abilities to recognise the central thematic significance of an event. Students who recognise the narrative position of the scene but not its central importance are credited with partial understanding only. Note the examples of students' responses used to support the scoring guide.[13]

**Question:**

Why does the story start with the sun rising?

**Scoring guide:**

**2 explicitly states the importance of owl's job OR the sun's central role in the story**

eg:
- Because it's all about how the sun doesn't rise later on.
- So you know what the owl does.
- It's the main thing.

**1 refers to use as a general 'beginning' OR narrative answer (owl hooted)**

eg:
- Because it's the beginning of the story.
- It's the start of the day.
- Because owl hooted.

**0 other reason**

eg:
- Because it's daytime.
- To give the story interesting pictures.
- So the animals wake up.

## Diagnostic categories

Another method for judging the quality of students' answers is to allocate answers to pre-defined diagnostic categories of response to a task. These categories can be constructed to represent increasingly sophisticated understandings of a task, or as in the example below, to provide information about different kinds of understandings and strategies.

Jan Clarke suggests response categories for recording the words students use in completing a cloze writing task.[14]

**NAME** ________________________________________ **DATE** __/__/__

| Words deleted in cloze | Word used | Correct word? | Semantically acceptable | Syntactically correct | Spelling analysis | Comments |
|---|---|---|---|---|---|---|
| fell | sliped | No | yes | Yes | Double consonant and add ed | Tutorial on short vowel and doubling consonant |
| swam | swam | Yes | | | | |
| shirt | shert | Yes | Yes | | -er, -ir confusion | Add to spelling list |

# Rating scales

Rating scales usually are used to assist markers to make judgements of extended responses. In using rating scales, markers judge the quality of student work against specified criteria. Markers make analytic judgements of work when they rate different aspects of the work. They make holistic judgements when they make a single rating based on an overall impression of the work.

In developing analytic and holistic rating scales, the first step is to consider the outcomes being addressed. Assessment criteria can then be developed to reflect those outcomes. For example, an essay answer format might be used to gather evidence about students' skills in structuring a particular kind of text and in controlling the surface features of writing. The criteria in an analytic rating scale would then include control of the particular text form (for example, narrative or exposition) as well as control over the surface features of the writing (spelling, punctuation, and grammar).

An example of an analytic rating scale for non-fiction writing is shown below, and for a mathematics investigation, and a concept map on page 44. An example of a holistic rating scale for the assessment of open-ended mathematics investigations is shown on page 45.

## Judging and recording
## Analytic rating scale for non-fiction writing

Teachers using the following marking guide judge student work against ten criteria using a 3-point rating scale. The criteria address writing outcomes specific to non-fiction wiriting as well as outcomes relating to general writing skills such as control of the mechanics of writing.[15]

| | **Criteria for assessment** | **Low** | **Medium** | **High** |
|---|---|---|---|---|
| **1** | The writing demonstrates an ability to interpret ideas meaningfully in context. | □ | □ | □ |
| **2** | The 'big idea' of the paper is interesting and clear. | □ | □ | □ |
| **3** | All of the main ideas are clearly related to the 'big idea'. | □ | □ | □ |
| **4** | The main ideas are organized into a logical sequence. | □ | □ | □ |
| **5** | The transitions from one main idea to the next are smooth. | □ | □ | □ |
| **6** | There are enough appropriate and accurate details to support each main idea. | □ | □ | □ |
| **7** | The choice of words is appropriate, varied, and creates a natural voice. | □ | □ | □ |
| **8** | The mechanics and grammar are integral to the meaning and effect of the writing. | □ | □ | □ |
| **9** | Stylistic variety and choice are evident in elements of rhetoric such as diction, syntax, structure, and figurative language. | □ | □ | □ |
| **10** | The paper is neat and presentable. | □ | □ | □ |

## Judging and recording
## Analytic rating scale for mathematics investigation

Teachers using this marking guide at The Gap State School assess mathematics investigations against six criteria using a 4-point rating scale. Two of the criteria ('group skills' and 'attitude and effort') require teacher observation of students while the students work.[16]

| | **Very high standard** | **High standard** | **Satisfactory** | **Experiencing difficulties** |
|---|---|---|---|---|
| | **Well established** | **Established** | **Developing** | **Beginning** |
| **Mathematical language** | Extensive and relevant use | Competent use | Appropriate use of basic terms | Incorrect or insufficient use of terms |
| **Communication** | Ideas clear, concise, informative, well organised | Ideas explained with understanding | Basic understanding | Limited ability to explain ideas logically |
| **Group skills** | Active involvement | Frequent involvement | Occasional involvement | Little or no involvement |
| **Investigative strategies** | Approach is consistently logical and systematic. | Approach is structured and systematic. | Attempts to use a structured and systematic approach. | Little evidence of any systematic plan of attack |
| **Data interpretation and representation** | Demonstrates very accurate and suitable data presentation and a wide range of manipulative and computational skills | Structured links in data presentation. Demonstrates a majority of manipulative and computational skills | Demonstrates basic manipulative and computational skills. Data presentation not always complete. | Incomplete and inaccurate data presentation with little manipulative and computational skills |
| **Attitude and effort** | Consistently enthusiastic, confident, persistent. Creative and innovative. | Enthusiastic, and generally confident and persistent. | Shows an interest but lacks confidence and persistence. | Shows little effort, interest or enthusiasm. |

## Judging and recording
## Analytic rating scale for concept maps

Teachers who use this analytic rating scale developed by Jan Clarke assess the quality of students' concept maps against three criteria using a 4-point rating scale.[17]

| | **With assistance** | **Satisfactorily** | **Capably** | **Efficiently** |
|---|---|---|---|---|
| **1** Identifies main ideas | □ | □ | □ | □ |
| **2** Classifies and categorises information | □ | □ | □ | □ |
| **3** Makes inferences to link information | □ | □ | □ | □ |

# Judging and recording
# Holistic rating scale for open-ended mathematical problems

This generic marking guide for open-ended mathematical problems was developed for use in the California Assessment Program.[18] The holistic ratings could be used to assess responses to a range of mathematical problems of different levels of difficulty.

## Rubric for open-ended mathematical problems

## *Demonstrated competence*

**Exemplary response** **Rating 6**

Gives a complete response with a clear, coherent, unambiguous, and elegant explanation; includes a clear and simplified diagram; communicates effectively to the identified audience; shows understanding of the problem's mathematical ideas and processes; identifies all the important elements of the problem; may include examples and counter examples; presents strong supporting arguments.

**Competent response** **Rating 5**

Gives a fairly complete response with reasonably clear explanations; may include an appropriate diagram; communicates effectively to the identified audience; shows understanding of the problem's ideas and processes; identifies most important elements of the problem; presents solid supporting arguments.

## *Satisfactory response*

**Minor flaws but satisfactory** **Rating 4**

Completes the problem satisfactorily, but the explanation may be muddled; argumentation may be incomplete; diagram may be inappropriate or unclear; understands the underlying mathematical ideas; uses ideas effectively.

**Serious flaws but nearly satisfactory** **Rating 3**

Begins the problem appropriately but may fail to complete or may omit significant parts of the problem; may fail to show full understanding of the mathematical ideas and processes, may make major computational errors; may misuse or fail to use mathematical terms; response may reflect an inappropriate strategy for solving problems.

## *Inadequate response*

**Begins but fails to complete problem** **Rating 2**

Explanation is not understandable; diagram may be unclear; shows no understanding of the problem situation; may make major computational errors.

**Unable to begin effectively** **Rating 1**

Words used do not reflect the problem; drawings misrepresent the problem situation; fails to indicate which information is appropriate.

**No attempt** **Rating 0**

# Different assessors

Although responses to paper and pen assessments often are judged by teachers, they also can be assessed by students' peers, and students themselves. Peer and self-assessment can help students reflect on what they know and how they learn and can encourage responsibility for learning.

Two examples of marking guides developed for use by students are shown below and on page 49. The first example assists very young students to assess their descriptions of the life cycle of a butterfly. The second example, developed by the University of Wollongong Faculty of Law, is used to collect evidence of students' abilities to assess their own performance–a stated learning objective of the faculty. An example of an analytic rating scale used by teachers to assess students' self-reflections is shown opposite.

# The need for comparability of judgements

In the classroom context, whether another teacher would make the same judgement of a student's piece of work using the same marking guide usually is not of great concern. In high-stakes contexts, however, inter-marker reliability is crucial.

Usually, the greater the requirement for comparability the more tightly the assessment criteria are specified. Markers are trained and the marking process is carefully monitored to ensure a high level of inter-marker agreement. In some high-stakes settings written work is assessed by several judges.

## Judging and recording
## Self-assessment using an analytic rating scale

This self-assessment guide was developed for grade 1 students. Students rated their descriptions of the life cycle of a butterfly against eight criteria using a 3-point rating scale.[19]

| | | Needs work | OK | Terrific |
|---|---|---|---|---|
| 1 | Did I draw the four parts of the butterfly cycle? | 😐 | 🙂 | 🙂 |
| 2 | Did I show details? | 😐 | 🙂 | 🙂 |
| 3 | Did I label the parts of each picture? | 😐 | 🙂 | 🙂 |
| 4 | Did I write a sentence for each picture? | 😐 | 🙂 | 🙂 |
| 5 | Do my sentences tell what the pictures show? | 😐 | 🙂 | 🙂 |
| 6 | Did I start each sentence with a capital letter? | 😐 | 🙂 | 🙂 |
| 7 | Did I end each sentence with a period? | 😐 | 🙂 | 🙂 |
| 8 | Is my work neat? | 😐 | 🙂 | 🙂 |

## Judging and recording
## Assessing students' self-assessments using an analytic rating scale

The following is an extract of an analytic rating scale used by teachers to assess students' self-assessments.[20] Teachers rate the self-assessments against criteria using a 5-point rating scale. Four of the eleven criteria are shown here.

| | No evidence | Little evidence | Some evidence | Significant evidence | Extensive evidence |
|---|---|---|---|---|---|
| **1** Has the student described his/her use of the writing process accurately? | □ | □ | □ | □ | □ |
| **2** Has the student described his/her writing strengths accurately? | □ | □ | □ | □ | □ |
| **3** Has the student described his/her writing weaknesses accurately? | □ | □ | □ | □ | □ |
| **4** Has the student conveyed an overall sense of who he/she is as a writer? | □ | □ | □ | □ | □ |

# Self-assessment

The following example of a student self-assessment guide was developed by the University of Wollongong Faculty of Law.[21] The analytic rating scale appears opposite. The explanation provided for students is reproduced below. The self-assessment sheet is used as evidence of students' abilities to assess their own performance—a stated learning objective of the faculty.

**A note on self-assessment**

You will be required to assess your own participation in class. This is not a cop-out by academic staff, as we will be making an evaluation of your participation. But one of the objectives of the law subjects at Wollongong is to enable you to develop your own learning skills. You will need to assess your own learning performance at various stages of your career, so in this subject, and in other subjects, you will receive assistance in developing your ability to assess your own performance.

Academic assessment is supposed to be a process by which the achievement of specified objectives is measured. The objectives of class participation are:

**1** to provide an opportunity to get you to develop your learning skills:
- to reinforce your own learning
- to give an incentive for you to plan, allocate and manage your own time including the development of techniques of selective reading
- and making judgements about the priority to be given to different tasks

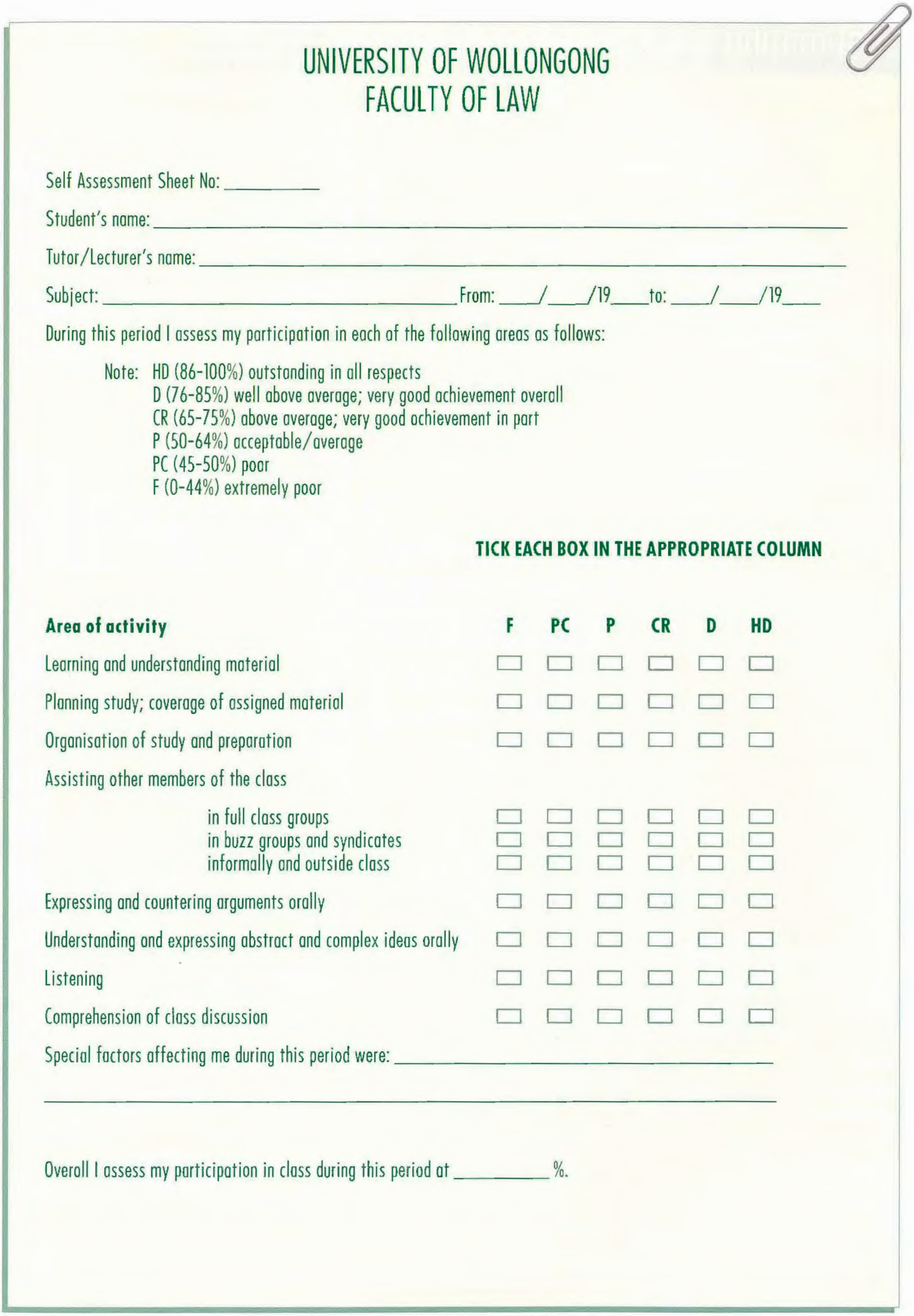
**2** to assist other members of the class to give and provide feedback on their learning

**3** to indicate to academic staff that you either have understood the material or that you need assistance

**4** to assist you to develop skills of
- expressing arguments orally
- expressing abstract and complex ideas orally
- listening
- comprehension

In this context, we have stated what we think the objectives of the process are. In real life you will need to work out your own objectives and the way in which you measure them. For this purpose you will have to work out the criteria by which you assess whether or not you achieve the objectives.

Each four weeks you will be asked to complete a form which will ask you to rate your achievement of each of the objectives listed above on a six-point scale, and to give yourself an overall score. This form will also give you the opportunity to indicate any abnormal factors affecting your work. We will check your self-assessment against our assessment of your participation, and will discuss with you any different perception we may have of your participation. You may be aware of factors affecting your participation of which we are unaware. The figure ultimately recorded will be agreed, though academic staff must reserve the final power of decision. When we have used this method previously, we have found that, initially, students tend to underestimate their own participation but, by the end of the session, there is a good measure of agreement between staff and students.



# UNIVERSITY OF WOLLONGONG
# FACULTY OF LAW

Self Assessment Sheet No: __________

Student's name: ______________________________________________

Tutor/Lecturer's name: ______________________________________________

Subject: ______________________ From: ___/___/19___ to: ___/___/19___

During this period I assess my participation in each of the following areas as follows:

Note: HD (86-100%) outstanding in all respects
D (76-85%) well above average; very good achievement overall
CR (65-75%) above average; very good achievement in part
P (50-64%) acceptable/average
PC (45-50%) poor
F (0-44%) extremely poor

**TICK EACH BOX IN THE APPROPRIATE COLUMN**

| **Area of activity** | **F** | **PC** | **P** | **CR** | **D** | **HD** |
|---|---|---|---|---|---|---|
| Learning and understanding material | □ | □ | □ | □ | □ | □ |
| Planning study; coverage of assigned material | □ | □ | □ | □ | □ | □ |
| Organisation of study and preparation | □ | □ | □ | □ | □ | □ |
| Assisting other members of the class | | | | | | |
| in full class groups | □ | □ | □ | □ | □ | □ |
| in buzz groups and syndicates | □ | □ | □ | □ | □ | □ |
| informally and outside class | □ | □ | □ | □ | □ | □ |
| Expressing and countering arguments orally | □ | □ | □ | □ | □ | □ |
| Understanding and expressing abstract and complex ideas orally | □ | □ | □ | □ | □ | □ |
| Listening | □ | □ | □ | □ | □ | □ |
| Comprehension of class discussion | □ | □ | □ | □ | □ | □ |

Special factors affecting me during this period were: ______________________________

______________________________________________

Overall I assess my participation in class during this period at __________%.

# Estimating and reporting achievement

In order to estimate and report student achievement, teachers need to make on-balance judgements from records of student work.

Paper and pen assessments can provide teachers with evidence that can be used to draw an inference about, and report, an individual's level of achievement in an aspect of their learning. In order to estimate and report this achievement, teachers make on-balance judgements from records of student work. Judgements of this kind are made on the basis of assessments that address one area of learning only. For example, it is usual to assess a student's achievement in numerical reasoning separately from the other dimensions along which that student might be progressing (such as reading, writing, and spoken language), even though those aspects of development may be related.

Teachers use paper and pen assessments to collect information for many purposes including to gather information about students' attitudes and reflections on their own progress. Sometimes the intention in gathering this kind of information is to report responses to each question separately, and not to combine responses across questions. For example, a teacher might construct a questionnaire made up of questions such as: Do you like mathematics classes? and Which areas of mathematics do you find most difficult? with the intention of analysing students' responses to each question separately.

The validity, reliability and objectivity of estimates of student achievement will depend on the quality of the evidence on which the estimates are based.

But in developmental assessment, students' responses to questions or tasks are combined to estimate individuals' locations on a map of developing knowledge, skills, understandings, attitudes or values. These questions might be statements on an attitude questionnaire (eg designed to measure attitude to mathematics) or items on a test (eg designed to measure achievement in an aspect of mathematics).

A progress map describes the path of typical development so there is rarely a perfect match between what is described on the map and evidence gathered for particular individuals. A best estimate must be made from the available evidence. The validity, reliability and objectivity of the estimate will depend on the quality of the evidence on which the estimate is based.

The *validity* of the estimate depends on the relevance of the evidence. When planning paper and pen assessments teachers need to ensure that the tasks provide evidence about relevant learning outcomes.

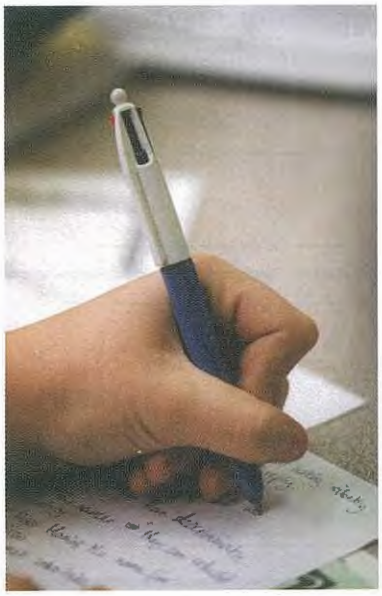
The *reliability* of the estimate depends on the amount of evidence on which it is based. Generally, the more evidence used to make the estimate, the more reliable the estimate.

The *objectivity* of the estimate depends on the extent to which it is unaffected by choice of task or choice of assessor.

The process of estimating a student's level of achievement on a progress map is relatively straightforward when judgements address outcomes and levels on a progress map directly. For example, rating scales can be constructed so that criteria address outcomes on a progress map, so that points on the scale correspond exactly with levels of a progress map.

When teachers base their reporting procedures on the principles of developmental assessment, reports are built around the concept of a progress map. These reports are likely to provide estimates of students' levels of achievement, and descriptions of the kinds of knowledge, skills, attitudes or values at each level. Descriptions usually are in the language of progress map outcomes but without the jargon. Parent reports also sometimes include information about the achievements of other students in the same grade.

Two examples of English reports are shown on this page. The first shows a student's achievement in relation to 80% of students in the class. The second shows the student's achievement in relation to grade percentiles (in this case the middle 80% and lower and upper 10% of students in the grade) and the grade average. Both reports are accompanied by descriptions of the kinds of knowledge, skills and understandings typically demonstrated at each achievement level.



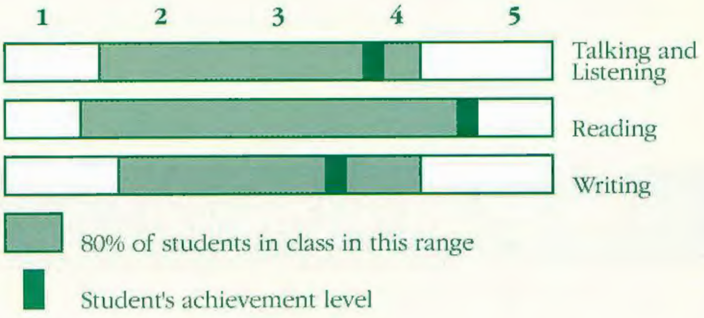
**Example 1:**

This report provides information about a student's level of achievement on a progress map and normative information about that achievement in relation to the class.[22]

Student: ____________________

Teacher: ____________________

Class: ____________ Date: ____________



General comments: ____________________

***This sort of reporting provides normative information about a student's achievement in relation to the class***

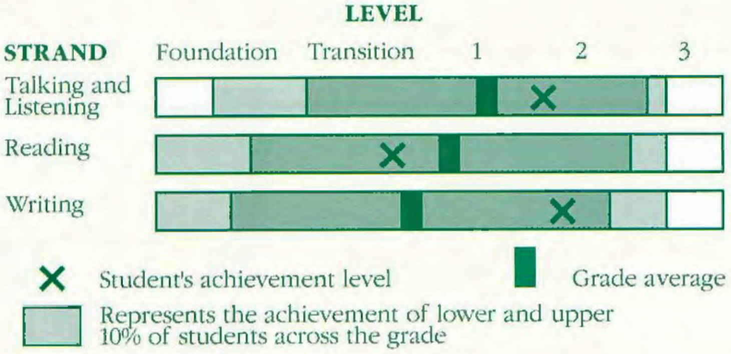
**Example 2:**

This report provides information about a student's level of achievement on a progress map and normative information about that achievement in relation to grade percentiles and average.[23]

Student: ____________________

Teacher: ____________________

Class: ____________ Date: ____________



Comments: ____________________

Teacher's signature: ____________________

Principal's signature: ____________________

Parent comments: ____________________

***This report provides normative information about the student in relation to grade percentiles and average.***

# Evaluating assessment and reporting strategies

Some teachers systematically evaluate their assessment and reporting strategies in order to improve the quality of their practices. An example of a self-evaluation questionnaire for teachers is shown on this page. This questionnaire focuses reflection on the effectiveness of a particular assessment task.

An example of a questionnaire for learners is shown opposite. This questionnaire is designed to provide feedback on the effectiveness of the assessment and reporting for an entire course.

## Evaluating assessment strategies — Self-evaluation

Regular evaluations of assessment strategies can be used to refine the assessment process. This self-evaluation questionnaire can be used to evaluate a range of paper and pen assessments.[24]

**Assessment event:** ______________________________

**Group:** ______________________________ **Date:** ___/___/___

| **Aspect of assessment** | ✔ | **Comments** |
|---|---|---|
| Did your assessment go according to plan? | □ | ______________________ |
| Was the assessment valid?<br>(eg appropriateness of level, balance and coverage of skills) | □ | ______________________ |
| Was the assessment reliable? | □ | ______________________ |
| Was the assessment fair to all learners? | □ | ______________________ |
| What changes would you make to the assessment before using it again? | □ | ______________________ |
| Did you get any feedback from learners on the assessment? If so, what was it? | □ | ______________________ |

# Evaluating assessment strategies
# Learner questionnaire

Regular evaluations of assessment strategies can be used to refine the assessment process. This learner questionnaire developed by the National Centre for Vocational Education Research can be used to evaluate a range of paper and pen assessments.[25]

| | Strongly disgree | Disgree | Agree | Strongly agree |
|---|---|---|---|---|
| I would have liked more information about this assessment at the beginning of the module. | □ | □ | □ | □ |
| There was too much assessment in this module. | □ | □ | □ | □ |
| The assessment in the module was just what I expected. | □ | □ | □ | □ |
| I was not able to demonstrate my skills fully during the assessment. | □ | □ | □ | □ |
| My assessor provided encouraging feedback during/after the assessment. | □ | □ | □ | □ |
| The assessment was a waste of time. | □ | □ | □ | □ |
| My assessment result was fair. | □ | □ | □ | □ |

1 Pennsylvania Department of Education *Academic Standards for Reading, Writing, Speaking and Listening,* p. 9.

2 Curriculum Corporation (1994) *The Arts—A Curriculum Profile for Australian Schools,* Carlton: Curriculum Corporation, p. 155.

3 Ministry of Education and Training, *Ontario Provincial Standards: Mathematics*

4 Curriculum Corporation (1994) *Science—A Curriculum Profile for Australian Schools,* Carlton: Curriculum Corporation, p. 35.

5 Curriculum Council Western Australia (1998) *Curriculum Framework,* p. 324.

6 Wiggins, G. (1998) *Educative Assessment,* San Franciso: Jossey-Bass Publishers, p. 151.

7 Izard, J.F. (1977) *Construction and Analysis of Classroom Tests,* Hawthorn: Australian Council for Educational Research, p. 8.

8 Clarke, J. *Monitoring Learning in English,* Mount Coot-tha School Support Centre p.59.

9 Curriculum Corporation (1994) *Science—A Curriculum Profile for Australian Schools,* Carlton: Curriculum Corporation, p. 93.

10 Wiggins, G. (1998) *Educative Assessment,* San Franciso: Jossey-Bass Publishers p. 294.

11 Curriculum Corporation (1994) *Science—A Curriculum Profile for Australian Schools,* Carlton: Curriculum Corporation, p. 7.

12 Recht, E., Forster, M. & Masters, G. (1998) *DART Mathematics Upper Primary,* Camberwell: Australian Council for Educational Research.

13 Bodey, W., Darkin, L., Forster, M. & Masters, G. (1997) *DART English Middle Primary,* Camberwell: Australian Council for Educational Research.

14 Clarke, J. *Monitoring Learning in English,* Mount Coot-tha School Support Centre, p. 55.

15 adapted from Hibbard, M. et al (1996) *A Teacher's guide to Performance-Based Learning and Assessment,* Virginia: Association for Supervision and Curriculum Development, p. 186.

16 Holcroft, L. and Coates, R. *Monitoring Learning in Mathematics,* Brisbane: Mt Coot-tha School Support Centre, pp. 20-21.

17 Based on Clarke, J. *Monitoring Learning in English,* Brisbane: Mt Coot-tha School Support Centre, p. 85.

18 Wiggins, G. (1998) p. 160.

19 Hibbard, M. et al (1996) pp. 132-133.

20 Hibbard, M. et al (1996) p. 207.

21 Nightingale, P., Te Wiata, I., Toohey, S., Ryan, G., Hughes, C., & Magin, D. (1996) *Assessing Learning in Universities,* University of New South Wales: Professional Development Centre, pp. 243-244.

22 NSW Department of School Education (1994) *Preliminary Ideas on Assessment and Reporting,* p. 50.

23 NSW Department of School Education (1994) p. 51.

24 Adapted from Clayton, B. (1995) *Focussing on Assessment South Australia*: National Centre for Vocational Education Research Ltd, p. 81.

25 Clayton, B. (1995) p. 82.

# PROFESSIONALLY DEVELOPED TESTS

Paper and pen assessments often are based on tests and questionnaires constructed by professional test development agencies. These instruments sometimes are designed for classroom use to complement teacher-developed assessments, either to provide broad measures of student achievement or more focused measures for diagnostic purposes. Professionally constructed paper and pen tests also are commonly used in selecting applicants for entry to competitive educational courses and as a basis for the award of scholarships. Other uses of paper and pen tests include the monitoring of educational standards (through international, national, and state/territory surveys of student achievement); testing for competence for the purposes of graduation and certification; and the measurement of student outcomes for use in educational research.

In professionally developed paper and pen assessments it is possible, and usual, to implement a variety of quality control checks that are more difficult, and perhaps less necessary, to implement for teacher-developed tests. This article outlines a range of common considerations in the development and use of professionally designed paper and pen tests and questionnaires.

In professionally developed paper and pen assessments it is possible, and usual, to implement a variety of quality control checks that are more difficult to implement for teacher-developed tests.

## Planning to measure

The starting point in the construction of any test is the intention to estimate individuals' levels of attainment in some aspect of their development–for example, to estimate their levels of achievement in some area of the school curriculum such as algebra, narrative writing, or knowledge of the workings of government; to estimate levels of attainment in cross-curricular skills such as verbal reasoning, numerical reasoning, or mechanical comprehension; and to estimate levels of development of attitudes and values such as empathy, goal orientation, and attitude toward mathematics.

### Designing a test

Underlying every test is the intention to estimate students' levels of attainment on just one aspect or dimension of their development. This intention is reflected in the intention to summarise test performances in a single score so that all students can be positioned with respect to each other in a single score order along the same progress map. Some tests are designed to provide estimates of students' levels of attainment in several areas of achievement (for example, a mathematics test might be designed to provide scores in number, measurement and space). Composite tests of this kind are really three tests administered in the same sitting. In this article we refer to a 'test' as a set of items designed to provide a single score for each student in a defined area of student development.

The first and most important criterion in designing a test is that the assembled items work together to provide evidence about the one aspect of development that the test is intended

In designing a test, professional test constructors consider whether the assembled items work together to provide evidence about the one aspect of development that the test is intended to measure.

- Do the test items work together to define one dimension? (internal consistency)
- Is this the dimension the test was intended to measure? (construct validity)

In selecting a professionally-constructed test, teachers need to consider the relevance of the test to their intended curriculum.

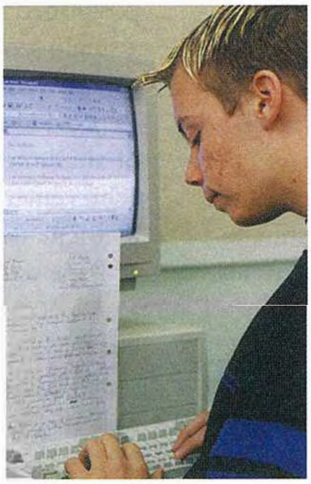
to measure. There are two separate considerations here: do the test items work together to define one dimension? (the criterion of internal consistency), and is this the dimension the test was intended to measure? (construct validity).

For example, the validity of a test designed to measure attainment in some aspect of the school curriculum (eg, writing and balancing chemical equations) will depend on the extent to which performances on the assigned tasks can be summarised meaningfully in a single score and mapped on to one dimension of achievement, and on the extent to which the tasks provide coverage of the intended curriculum in this area of chemistry learning. When the curriculum is expressed in terms of intended learning outcomes, the test should provide the broadest possible coverage of the relevant outcomes.

To ensure coverage of the range of important outcomes in an area of learning, it is common for test developers to use a test specification 'grid'. The purpose of the grid is to plan and then monitor the extent to which drafted test questions provide coverage of the outcomes within an aspect of learning.

## Selecting professionally developed tests

In selecting a professionally constructed test, teachers need to consider the relevance of the test to their intended curriculum. Does the test address the range of learning outcomes to which they have been teaching? Is it of appropriate difficulty for this group of students? If the test addresses only some of the intended learning outcomes, is there a risk that the test will distort teaching and students' understandings of what is important? Are there other tests that are better aligned with the curriculum intentions?



## Developing/Selecting Appropriate Tests [1]

**test developers should:**

1. Define what each test measures and what the test should be used for. Describe the population/s for which the test is appropriate.
2. Accurately represent the characteristics, usefulness, and limitations of tests for their intended audience/s.
3. Explain relevant measurement concepts as necessary for clarity at the level of detail that is appropriate for the intended audience/s.
4. Describe the process of test development. Explain how the content and skills to be tested were selected.
5. Provide evidence that the test meets its intended purpose/s.
6. Provide either representative samples or complete copies of test questions, directions, answer sheets, manuals, and score reports to qualified users.
7. Indicate the nature of the evidence obtained concerning the appropriateness of each test for groups of different racial, ethnic, or linguistic backgrounds who are likely to be tested.
8. Identify and publish any specialised skills needed to administer each test and to interpret scores correctly.

**test users should:**

1. First define the purpose for testing and the population to be tested. Then, select a test for that purpose and that population based on a thorough review of the available information.
2. Investigate potentially useful sources of information, in addition to test scores,to corroborate the information provided by tests.
3. Read the materials provided by test developers and avoid using tests for which unclear or incomplete information is provided.
4. Become familiar with how and when the test was developed and tried out.
5. Read independent evaluations of a test and of possible alternative measures. Look for evidence required to support the claims of test developers.
6. Examine specimen sets, disclosed tests or samples of questions, directions, answer sheets, manuals, and score reports before selecting a test.
7. Ascertain whether the test content and norm group/s or comparison group/s are appropriate for the intended test takers.
8. Select and use only those tests for which the skills needed to administer the test and interpret the scores correctly are available.

The most common answer format in professionally developed tests has traditionally been the multiple-choice response. The reason for the widespread use of this format is that it maximises the reliability of marking.

Once the aspect of development or achievement to be measured has been identified and clarified, the second step in the test development process is to decide on the kinds of questions or prompts to be used.

The most common answer format in professionally developed tests has traditionally been the multiple-choice response. The reason for the widespread use of this format is that it maximises the reliability of marking. In fact, most large-scale multiple-choice tests are marked by machine. Other answer formats commonly used in professionally developed tests are essays and short answer responses, both of which usually are marked by hand and involve a degree of judgement on the part of markers. With the advent of technology able to read students' handwritten responses, the machine marking of short answer responses and even essays is becoming more common.

When designing test questions, test developers face the question of whether to provide questions which can be scored either right or wrong, or more open questions which allow for a range of different kinds of responses from students. Open questions often are not marked right or wrong, but are scored in several ordered categories to recognise different levels of quality in individuals' responses. The advantage of open questions of this type is that they can be used with students at different levels of ability or achievement. Ideally, these questions allow students at low levels of achievement to respond, perhaps in relatively incomplete and unsophisticated ways, but also challenge students at higher levels of achievement. Ratings and schemes of 'partial credit' scoring are developed to capture these different levels of response.

To ensure tests are as fair as possible, professional test developers

- avoid particular topics
- minimise the use of unfamiliar terms and concepts
- check for cultural bias and gender bias
- invite particular interest groups to review test items for inclusivity and
- conduct statistical checks for item bias.

## Minimising bias

It is important that professionally developed tests are designed to be as fair as possible to all students likely to take them. Test developers attempt to eliminate bias in tests by avoiding particular topics and by minimising the use of terms and concepts likely to be unfamiliar to particular groups of test takers. Special attention is paid to the possibilities of cultural bias and gender bias.

In large testing programs it is common to invite particular interest groups to review test items for inclusivity. For example, draft tests might be referred to people familiar with the special difficulties experienced by students who do not speak English as a first language, or to representatives of minority groups to ensure that test items are not culturally biased.

It is common in the test development process to conduct a further statistical check for item bias. This check, sometimes called 'differential item functioning' (dif), is designed to identify items which are unusually difficult for a particular group of students. The statistical identification of differential item functioning can be useful in drawing attention to items that need to be examined more closely for possible bias.

# Standardising test conditions

It is usual for professionally developed tests to specify the conditions under which tests are to be taken. Standardised test conditions are used to ensure that results are comparable across different test locations and over time. The time available for testing, the materials that students are able to take into the test centre (notes, textbooks, calculators), the introductory instructions, and the test centre conditions are all specified in an attempt to ensure that some students are not unfairly advantaged, and others unfairly disadvantaged by the conditions under which tests are administered.

To ensure that results are comparable across different test locations and over time, professionally developed tests usually specify the conditions under which tests are to be taken.

## Minimising unfair advantage

It is common in professionally developed tests for students to be provided with some 'practice' questions prior to commencing the test. The purpose of these practice questions is to give all students an opportunity to familiarise themselves with the type of question to be asked and with the method for recording responses. The purpose is again to try to minimise unfair advantages some students may have because of their familiarity with test-taking procedures.

Many professional testing programs make collections of sample questions available to candidates before the date of testing. These collections may take the form of entire practice tests. The availability of practice tests provides all students with an opportunity to practice answering questions under conditions similar to those they will encounter in the test itself. In this way, students have opportunities to practice pacing themselves on a range of items similar to the range included in the test. Answers to practice tests and worked solutions (including samples of student essays) provide further opportunities to ensure that students undertake tests on a similar footing and are not disadvantaged by individual differences in test-wiseness and test-taking skills.

Special provisions often are made to ensure that students are not disadvantaged by other test-irrelevant factors such as poor eyesight or limited manual dexterity. Many large testing programs provide Braille and large-print versions of tests. Scribes may be provided for students who are unable to record their own answers, and—where reading skills are not themselves being assessed—tests may be read to students with limited reading skills.

Professionally developed tests also may make provisions for candidates who, for religious reasons, are unable to attend on the day of testing or must be absent for part of the testing period. For example, in the Graduate Australian Medical School Admissions Test, a group of candidates who, for religious reasons, were unable to sit the test on the day of testing were supervised as a group from the time of testing until the following day when they were able to sit the test.

In an attempt to standardise test conditions to ensure that all students are treated equally and fairly, professional tests also may be accompanied by instructions designed to minimise cheating (eg, rules for the kinds of electronic devices that can be admitted to the room, seating plans, arrangements for accompanying students when they are absent from the test room). Proctors also may be provided with guidelines for handling printing errors in test booklets, claims that a test item contains an error, fire-alarms, and other unpredictable events. (One Australian testing centre was confronted with the

To ensure that all students are treated equally and fairly, professional tests usually

- provide 'practice' questions
- make provisions for students with disabilities
- make provisions for students with religious needs
- provide instructions to minimise cheating
- provide guidelines for unpredictable events
- specify appeal mechanisms.

problem of a bird flying into the room and circling throughout the test!)

As a final attempt to ensure that all students are treated fairly by the conditions under which standardised tests are taken, many testing programs provide appeal mechanisms that allow students to present cases for special consideration arising from factors such as illness and personal stress on the day of testing.

### Administering professionally developed tests

Users of professionally developed tests have a responsibility to test takers to ensure that tests are administered under the conditions specified by the developers. It is possible to compare students' performances from one testing centre to another and with published test norms only if standardised directions are followed. Users also have a responsibility to ensure that individuals are not disadvantaged by factors irrelevant to the test content.

## Standardising marking

To ensure that all students' responses are judged in the same way, professionally developed tests provide

- assessment criteria
- examples of both acceptable and unacceptable responses to open-ended items, and
- examples of student responses at each point on a rating scale

In professionally developed tests, steps are taken to ensure that all students' responses are judged and recorded in the same way. Consistency of marking is easier to ensure in multiple-choice tests where the scoring rule involves the identification of a single correct alternative (the 'key'). In tests of this kind, scoring can be automated and the possibilities for unfair marking can be minimised or eliminated.

However, even in multiple-choice tests, routine checks can be built in to the scoring process. Optical mark readers can be programmed to identify items for which students' responses are ambiguous (eg, where two alternatives appear to have been marked). Once identified, these unclear responses often are checked by hand and a decision made about how they are to be scored. The automatic identification of atypical patterns of response also is used in some testing programs. An atypical pattern of responses can arise because an answer sheet was misaligned when being read, or because a student turned two pages rather than one, thus failing to see and attempt some items.

### Ensuring consistency of judgement

The marking of open-ended test items almost always involves a degree of judgement on the part of markers. To assist in the making of judgements, markers are provided with criteria and examples of both acceptable and unacceptable responses to each item. When part-marks are available for partial completion or partially correct responses to an item, criteria and examples are provided to illustrate each of the possible scores on the item. The identification of varying levels of response to an item can assist in illustrating levels along a map of increasing achievement (see example on page 62).

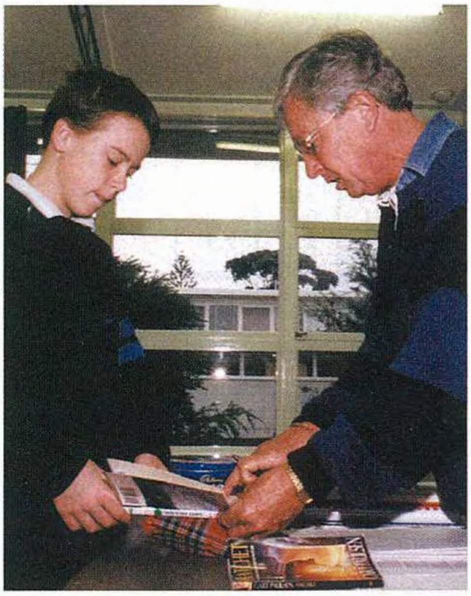
Professionally developed writing tests provide examples of the criteria to be used in assessing student scripts (see page 63) as well as annotated samples of student writing to illustrate the application of those criteria. Assessments of student writing always involve an on-balance judgement of quality in which account is taken of the piece of work as a whole. For this reason, professionally developed writing tests usually provide a number of examples of student writing at each point on a rating scale to illustrate the variety of ways in which students can achieve the same score.

In large testing programs it is usual to put in place procedures for monitoring the consistency of markers. Different markers are asked to assess the same student work (eg essay), and the assessments made of that work are compared. Through comparisons of this kind it is possible to identify 'discrepant' markers. Discrepant markers sometimes are more lenient (tend to give higher scores) than other markers; sometimes they are harsher (tend to give lower scores); and sometimes they tend to make less use of the available score range (give all work very similar marks). When large discrepancies are found, student work is commonly remarked by a third marker.

Even with the use of procedures for identifying and handling discrepancies among markers, small but consistent differences in marker harshness/leniency can remain undetected. In an attempt to ensure that all students are treated fairly, statistical adjustments to students' results are sometimes made to remove remaining marker effects (eg to increase a students' final result slightly to adjust for the fact that the student's work was marked by two relatively harsh markers).

## Using professionally developed marking guides

Users of professionally developed tests have a responsibility to test takers to ensure that provided marking guides and criteria are used in accordance with the test developers' guidelines. Only if these guidelines are followed is it possible to compare students' performances from one testing centre to another and with published test norms.



## Standardising Marking — Ensuring consistent judgement

The following open-ended question from a DART reading test for upper primary students provides a score of 1 (partial credit) for answers which provide examples only, but do not generalise. A score of 2 (full credit) is provided for generalisations showing a deeper understanding of the term 'suitable soils'. The picture shows that a score of 1 on item 7 illustrates Level 3 reading achievement; a score of 2 illustrates Level 5 reading achievement.[2]

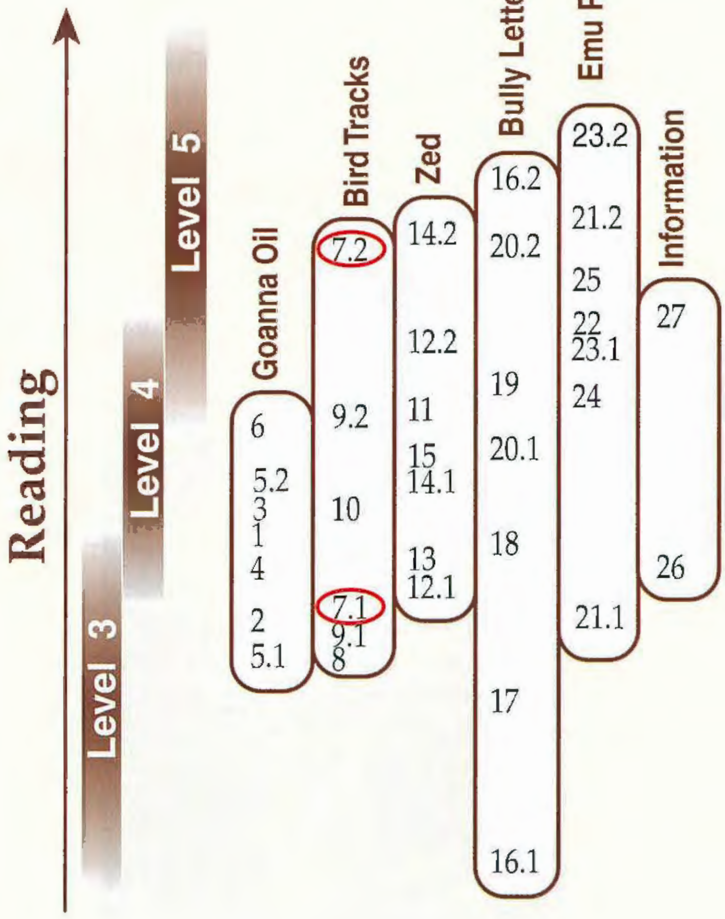
**7. In sandy regions tracks will be left in suitable soils. What are suitable soils?**

**7.2** *generalisation*

eg, "Soils that are soft so they leave a clear track."
"Soils that allow a clear impression to be made."

**7.1** *example only*
eg, "Sand"

**7.0** *incorrect response*
eg, "tracks; soils"
eg, "Sand."



## Marking Guide for Assessing Language Features in Student Writing
## Source: National School English Literacy Survey[3]

| | |
|---|---|
| **5** | Revises writing to be consistent in content and style.<br>Experiments with rearranging sentences.<br>Controls grammatical structures and punctuation in complex sentences.<br>Organises writing into coherent whole appropriate to context (eg paragraphs for a narrative, headings and sub-headings for informational text).<br>Uses precise and effective vocabulary.<br>Approximates spelling of particularly difficult words using patterns and conventions. |
| **4** | Begins to adopt organisational conventions of structured format (eg general introductory statement to a report).<br>Contains a variety of sentence forms (eg simple and complex sentences).<br>Uses appropriate punctuation most of the time.<br>Shapes writing with clear beginning and end and possibly paragraph divisions.<br>Uses appropriate vocabulary most of the time.<br>Spells most words correctly. |
| **3** | Shows some evidence of planning, revising and proof reading own writing.<br>Controls simple sentence structure and attempts more complex structures.<br>Attempts to vary sentence beginnings.<br>Attempts to shape piece structurally (eg notion of beginning and end).<br>Spells many common words correctly.<br>Writes legibly. |
| **2** | Uses simple sentences.<br>Uses repetitive sentence structure.<br>Uses simple conjunctions (eg 'and' and 'but').<br>Controls common punctuation some or all of the time (eg capital letters, full stops).<br>Spells high frequency words correctly most of the time.<br>Writes in a way that can be generally interpreted by others. |
| **1** | Uses some basic conventions (eg writes from left to right, puts spaces between words).<br>Contains some known words, or words represented by their initial letters.<br>Uses some correct initial letters and other sounds.<br>Can be read back by the child at the time of writing. |

The calibration of professionally developed test items along a progress map allows test users to interpret students' test scores in terms of the kinds of knowledge, skills, and understandings typically associated with each score.

In professionally developed tests, statistical analyses are undertaken to establish the extent to which the items that make up a test work together to support the intention to summarise each student's responses in a single score. These statistical methods are known as 'item response' modelling.

One outcome of applying item response modelling to a set of items is that the items that make up a test are 'calibrated' according to their difficulty. An example of the outcome of calibrating a set of items is shown on page 62. The picture on page 62 shows the results of using the responses of a large group of primary school students to calibrate the items on a reading test. The tasks that students most often completed correctly are shown towards the bottom of the picture. The easiest item for these students was 16.1 (in other words, to make a score of 1 on Item 16). The tasks least often answered correctly are shown towards the top of the picture. The hardest item was 23.2 (in other words, to score 2 on Item 23).

The statistical calibration of items in this way provides a 'map' of increasingly difficult items. When the items from two, three or more tests are calibrated together, a larger number of items is located at each position along the map. By examining and describing what students have to do to answer these items, an increasingly rich description of positions along the map can be developed.

## Using progress maps

The picture on page 65 shows descriptions of items calibrated along a map of increasing reading achievement.[4] In developmental assessment, we call this kind of map a 'progress map'. The easiest item on this reading test is calibrated near the bottom of the picture. This item required young children to predict what a book was about from the title and illustration on its cover. The most difficult item on this test is calibrated near the top of the picture. This item required children to demonstrate an appreciation of the appropriateness of the format of a piece of text for its purpose (namely, the appropriateness of a question and answer format for recording interview data). By reading up this page it is possible to 'see' the increasing levels of reading ability required to answer increasingly difficult questions about text. Students' performances on this test are expressed on a scale that runs from near 100 to near 600. The text alongside the scale shows the kinds of reading tasks that students at that level are likely to be mastering.

The calibration of professionally developed test items along a progress map allows test users to interpret students' test scores in terms of the kinds of knowledge, skills, and understandings typically associated with each score. Consider, for example, a student who scored 350 on this reading scale. Just below a score of 350 is the description 'finds evidence to support a statement'. This is an example of the kind of skill that a student at this level of reading proficiency has a reasonable chance of demonstrating. Skills listed below this level are more likely to be demonstrated; skills listed above are less likely to be demonstrated.

## Reading Achievement

60

Recognises the connection between presentation style and nature of information (eg question & answer format for interview data).

Infers meaning from figurative language.

Level 5

Interprets idiomatic language (eg 'last but not least').

50

Recognises how linguistic features (eg exclamation marks) support ideas implicit in a text.
Selects several pieces of information from a complex presentation of text.
Recognises probable context for a piece of writing.
Explains an author's point of view.
Recognises the tone of a simple poem.
Orders detailed events from a narrative.

Level 4

40

Recognises conventional linguistic features (eg pronunciation guides).
Interprets factual information.
Recognises the relationship between two pieces of text.
Generates research question to explore topic about which they have read.
Works out meaning of unknown word from context and picture clues.
Finds evidence to support a statement.
Orders instructions in a procedure.
Extracts information from complex presentation of text and pictures.

Level 3

30

Recognises main idea in paragraph of factual text.
Decides whether writing is fact or fiction based on described events.
Recognises text genre from book titles.
Makes connections between pieces of factual info. in simple text.
Predicts a plausible ending for an illustrated story.

20

Recognises how elements of an illustration support text in a story.
Uses title and illustration to predict story setting.
Interprets picture to predict what happens next in illustrated story.

Level 2

10

Uses book title and illustration to identify key elements of story.

Level 1

Test 'norms' allow students' test results to be interpreted in terms of the performances of other students of the same age or in the same grade. Normative information is provided in tables and/or graphically.

## Using test norms

In professionally developed tests it is common to provide test users with information about the performances of some relevant population of students on a test. These test 'norms' allow students' test results to be interpreted in terms of the performances of other students of the same age or in the same grade. For example, a grade 6 mathematics test may show how a national sample of sixth graders performed on that test.

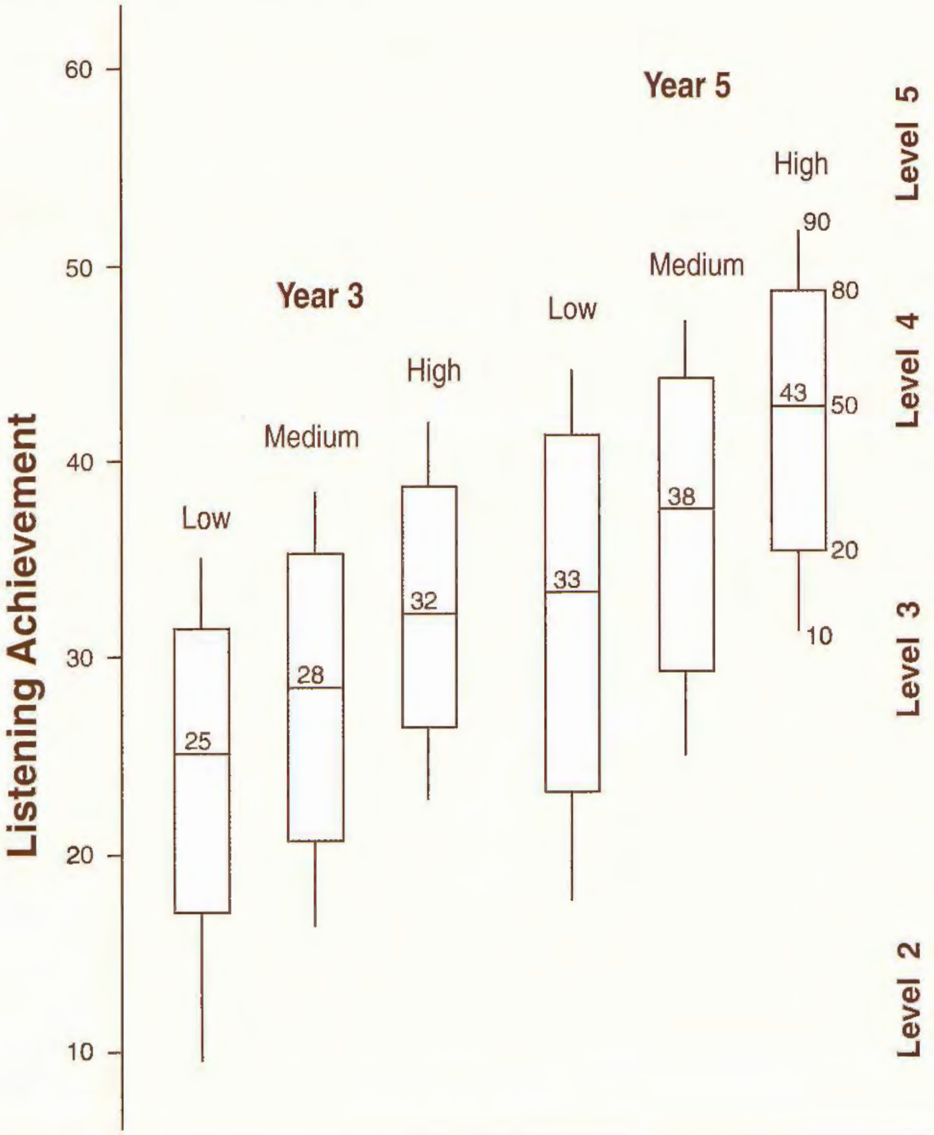
Test norms usually are reported as 'percentiles'. A student who scores at the 80th percentile scores as well as, or better than, 80 per cent of students in the relevant norm group. Manuals for standardised tests usually contain norm tables showing the percentage of students scoring at or below each possible test score. These percentiles may be shown for the entire student population as well as for relevant subgroups (eg fifth graders, sixth graders, boys, girls).

As well as providing norm tables to allow teachers to compare test results with typical performances for the age or grade, test manuals sometimes provide graphical displays of normative information. One format for displaying test norms is a 'box-and-whisker' plot (see page 67) in which particular percentile points are marked (eg, 10th, 25th, 50th, 75th and 95th percentiles).

In comparing students' test results with published test norms, it is important that test users consider the relevance of the norm population and ensure that the test is administered under the conditions and procedures recommended by the test developer.

## Using test norms

Teachers using the standardised viewing tests constructed for the National School English Literacy Survey are able to compare students' listening test results with the performances of a national sample of students on these tests.[5] The box-and-whisker plot below shows the performances of Year 3 and Year 5 students in low, medium and high socio-economic groups. For example, a Year 3 student with a test performance of 340 on the Listening Achievement scale has performed above the 50th percentile for Year 3 students in the highest socio-economic group and also above the 50th percentile for Year 5 students in the lowest socio-economic group.



[1] Excerpt from the *Code of Fair Testing Practices in Education,* prepared by a joint committee of the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (1998).

[2] Forster, M., Mendelovits, J., & Masters G. N. (1995) *Developmental Assessment Resource for Teachers DART Upper Primary,* Melbourne: Australian Council for Educational Research.

[3] Masters, G. N., & Forster, M. (1997) *Mapping Literacy Achievement,* Canberra: Department of Employment, Education, Training and Youth Affairs.

[4] as above, p. 108.

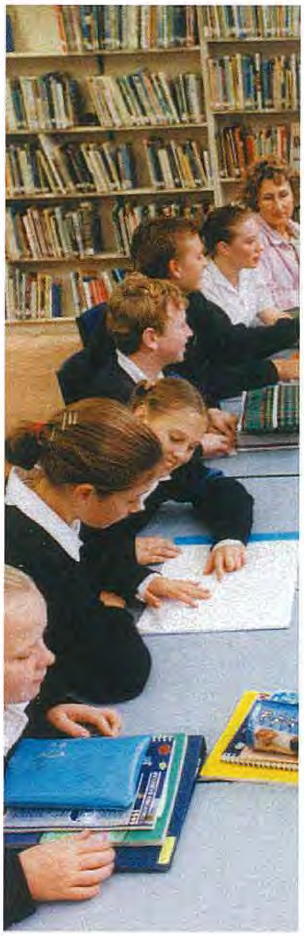
[5] as above, p. 182.

# DESIGNING PAPER AND PEN ASSESSMENTS

In developmental assessment, teachers monitor student progress against a map of developing skills, knowledge, understandings, attitudes or values. Paper and pen assessment is one method that can be used to collect evidence of student achievement. Different kinds of answer formats can provide different kinds of evidence that teachers can use to estimate students' locations on a progress map.

This article lists issues which teachers need to consider when designing paper and pen assessments. These issues include questions of assessment purpose, methods for judging and recording student performance, and ways of estimating and reporting students' levels of achievement. A 'checklist' summary of the assessment design process is included at the end of the article.

## Planning paper and pen assessments



### What is the purpose of the assessment?

For example, do you want to diagnose individuals' strengths and weaknesses in a particular area of learning, inform the teaching process at a whole class level, or infer students' levels of achievement for reporting to parents?

How you answer this question will determine the range of curriculum goals or outcomes you address (see pages 2-5 & 28-30).

For which curriculum goals or outcomes will you assemble evidence?

Is the evidence you are assembling relevant? That is, does it focus on explicit instructional outcomes or goals?

Depending on your assessment purpose, does the evidence you are assembling address the full range of outcomes (see pages 2-5 & 28-30)?

### Which answer format is best suited to the outcomes you are addressing?

For example, if the outcomes require students to demonstrate factual knowledge, then a multiple-choice format might be used. If the outcomes require students to structure a logical argument, then an essay format might be used (see pages 30 and 33-39).

### Are the planned tasks fair?

Do the tasks have a clear focus? Are the tasks inclusive of students from different gender and cultural backgrounds? Is the language used in presenting the tasks accessible? Have you had time to review the task/s before administering them (see pages 30-33)?

# Judging and recording

## How will you judge and record student work?

Which method for judging and recording student work is most appropriate to your assessment purpose and answer format? Will you use scoring guides, diagnostic categories, analytic ratings, or holistic ratings (see pages 40-45)?

## Which criteria will you use?

Do your assessment criteria provide evidence that is relevant? That is, do they address the outcomes being assessed? Are they clear, fair and useable? Do you have accompanying examples of student work (see pages 40-50)?

How great is the need for comparability of judgements?

## Who will do the assessing?

Will the tasks be assessed by the teacher, by students' peers, or self-assessed (see page 46)?

# Summarising student work

## How will you estimate students' levels of achievement on a progress map?

From which evidence will you estimate a student's level of achievement? Will you use holistic ratings or analytic ratings of student work, or scoring criteria?

Have you checked that your estimates are valid? (ie, that the estimates are based on relevant evidence?)

Have you checked that the estimates are reliable? (ie that they are based on sufficient evidence?)

Have your checked that the estimates are objective? That is, unaffected by choice of task or choice of assessor (see page 50).

# Reporting student work

## How will you report student achievement?

Will you interpret levels of achievement descriptively? Will you display achievements graphically (see page 51)?

# A SUMMARY OF THE PAPER AND PEN ASSESSMENT DESIGN PROCESS

| Design stage | Design strategies |
|---|---|
| Deciding the assessment purpose | • Describe the assessment purpose.<br>• Review these descriptions against important curriculum objectives and outcomes of the learning area. |
| Deciding the curriculum goals or outcomes to be targeted | • List the goals or outcomes. |
| Deciding on the answer format | • Check that the format is suited to the outcomes being addressed. |
| Reviewing before administration | • Check for fairness, (including clarity, inclusivity, accessible language) |
| Deciding on a procedure for judging and recording evidence | • Decide who will assess (self, peer, teacher).<br>• Develop marking guides (scoring criteria or rating scales).<br>• Review these against outcomes being assessed.<br>• Review for clarity and useability. |
| Deciding on a procedure for estimating levels of achievement on a progress map | • Describe the procedure for estimating levels of achievement.<br>• Review these descriptions against the task, purpose and audience. |
| Deciding on a procedure for reporting levels of achievement | • Describe the procedure for reporting levels of achievement.<br>• Review these descriptions against the task and purpose and audience. |

Please photocopy and fax or mail.

# ASSESSMENT RESOURCE KIT (ARK)

**ORDER FORM**

Australian Council for Educational Research
19 Prospect Hill Road (Private Bag 55)
Camberwell Vic 3124 AUSTRALIA
Telephone: (03) 9277 5656 Fax: (03) 9277 5678
International: Tel: 61 3 9277 5656 Fax: 61 3 9277 5678
email: sales@acer.edu.au

## Charge to

Name *or* Organisation ______________________

Purchase Order no. ______________________

Address ______________________

______________________

__________ Postcode __________

## Deliver to

Name ______________________

Organisation ______________________

Street address ______________________

______________ Postcode ______

Telephone ( ) ______ Fax ( ) ______

Order date __________ Date required __________ ACER account no. __________

| Title | Cat. No. | Price | Quantity | Total Price |
|---|---|---|---|---|
| Developmental Assessment | 100ARK | AUD$12.95 | | |
| Portfolios | 101ARK | AUD$12.95 | | |
| Performances | 102ARK | AUD$12.95 | | |
| Projects | 103ARK | AUD$12.95 | | |
| Progress Maps | 105ARK | AUD$12.95 | | |
| Understanding Developmental Assessment (videotape) | 700ARK | AUD$49.95 | | |
| Set of 3 Posters: Reading, Spelling, Writing | 400ARK | AUD$6.95 | | |
| Products | 104ARK | AUD$12.95 | | |
| Paper and Pen | 107ARK | AUD$12.95 | | |
| **Forthcoming titles:** | | | | |
| Assessment Methods (available late 1999) | 106ARK | AUD$12.95 | | |
| Implementing Developmental Assessment:<br>Workshop Manual & Videotape (available late 1999) | 500ARK | TBA | | |
| | | | Subtotal | |
| | | | Freight | |
| | | | TOTAL | $ |

ADD FREIGHT:
*Within Australia:* 10% of invoice value, min $5.00 max $20.00.
*Overseas(airmail):* 1 - 5 copies: AUD$30
6 - 10 copies: AUD$45
More than 10 copies: please ring or fax for a quote.

NOTE: Prepayment required for non-account orders.
Prices subject to change without notice - you are welcome to check before you order.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

**CHARGE:** □ ACER Account No. ______________ □ Cheque Enclosed (AUD$)

□ Bankcard □ Mastercard □ American Express □ Visa □ Diners Club

ACER PRESS

□□□□ □□□□ □□□□ □□□□

Name (*please print*) ______________________ Signature ______________ Expiry date ________

ACN 004 398 145

*Paper & Pen* is one in a series of magazines in the ACER Assessment Resource Kit (*ARK*).

This video and magazine resource provides information about assessment issues and methods.

For further details about other magazines, videos and the workshop manual in this series contact the
Australian Council for Educational Research,
19 Prospect Hill Road,
Camberwell, Victoria,
Australia, 3124.
Phone: +61 3 9277 5656
Facsimile: +61 3 9277 5678