Submission to the Queensland Parliament's Education and Innovation Committee's Inquiry into Assessment Methods for Senior Maths, Chemistry and Physics

This submission is made by Professor Geoff Masters and Professor Gabrielle Matters on behalf of the Australian Council for Educational Research (ACER).

1/165 Kelvin Grove Road
KELVIN GROVE
Queensland 4059
Tel: 3238 9000

Some reflections on the assessment process

Traditionally, assessment has been viewed as the process of judging how well students have learnt what they have been taught. Under this traditional view, teaching, learning and assessment occur sequentially: teachers teach, students learn, and tests and examinations are used to determine how much of what has been taught students have successfully learnt. Results are reported as percentages, which may then be converted to A to E grades.

In contrast, in modern classrooms, assessment is seen as an essential and ongoing component of effective teaching. Teachers use assessments to identify where individual students are in their learning, to diagnose errors and misunderstandings, to plan teaching, to provide feedback to guide student effort, to monitor the progress that individuals make over time, and to evaluate the effectiveness of their teaching strategies and interventions. In this sense, assessment has parallels with assessment in other professions such as medicine and psychology where the purpose is not so much to judge as to understand for the purposes of making informed decisions.

These two uses of assessment are sometimes characterised as different and competing 'purposes'. However, they need not be as different as they appear. This is because all assessment in education has the same fundamental purpose – namely, *to establish and understand where students are in their learning at the time of assessment*. A teacher who conducts an assessment to establish what an individual knows and does not know, to identify appropriate starting points for teaching, or to evaluate the progress that students have made is engaged in a process of establishing and understanding where students are in their learning at the time of assessment. Similarly, examinations conducted at the end of Year 12 are designed for the fundamental purpose of establishing students' levels of attainment (ie, their knowledge, skills and understandings) at that point in time. Different levels of diagnostic detail may be required in different contexts, but underpinning all assessment is the common purpose of identifying where students are in an aspect of their learning at the time of assessment.

It follows that the assessment process is one of gathering evidence that can be used to draw a valid and reliable conclusion (inference) about a student's current level of attainment within some specified area of learning.[1]

[1] This description of the educational assessment process is based on *Reforming Educational Assessment: Imperatives, Principles and Challenges* (Masters, 2013).

The first step in the assessment process is to establish a clear definition of the area (or 'domain') of learning within which student attainment and progress are to be assessed and monitored. At Year 12, this clarity is provided by course syllabuses which spell out the knowledge, skills and understandings that students are expected to develop. Ideally, course syllabuses are deeply grounded in discipline knowledge. They should identify knowledge and skills essential to the discipline, with a particular emphasis on the development of students' understandings of key concepts, principles and ideas in the discipline. Syllabuses usually identify sub-areas of the discipline (eg, algebra, calculus, geometry), but a syllabus should be more than a catalogue of knowledge and skills. It should be built from an empirically-based understanding of how learning occurs within the discipline, including an understanding of how the course builds on to prior and pre-requisite learning, how it lays the foundations for further learning, and how content is best sequenced within the course to promote the development of student knowledge, skills and understandings.

The second step in the assessment process is to decide on a way (or ways) of gathering evidence about where students are in their learning within the domain. The essential requirement here is that the method of assessment must be *appropriate to the domain* (ie, to the kinds of knowledge, skills and understandings that make up the domain). Domain-appropriate assessment methods in learning areas such as dance and drama include direct observations of student performances. Domain-appropriate assessment methods in learning areas such as art and technology include observations and evaluations of the products of student work. Any attempt to assess learning in domains such as these using only paper and pen assessments would lack 'construct' validity. In other words, the chosen assessment method would be inappropriate to the domain and would be incapable of providing valid information about where students are in key aspects of their learning. The requirement here is fitness for purpose, and different assessment methods often are appropriate for gathering evidence about different kinds of learning within the same area or course of learning.

Unfortunately, in educational practice, decisions about assessment methods are sometimes based not on fitness for purpose, but on generalised personal preferences. For example, some educators have developed general preferences for large, complex assessment tasks over shorter forms of assessment; for 'authentic' real-world tasks over invented tasks; for teacher-created assessments over externally-developed assessments; or for 3-hour written examinations over all other forms of assessment. Once these preferences are established, they tend to be applied across the board, ignoring the fact that they may sometimes be inappropriate (or at least, inefficient) ways of gathering evidence about particular aspects of student learning. In any assessment context, general personal preferences must to be put to one side and assessment methods chosen instead on the basis of their feasibility and capacity to provide valid and reliable evidence about the area of learning under consideration.

The third step in the assessment process is to decide how student responses or performances are to be evaluated and recorded. This is a crucial step in the process. The selected assessment method/s must be capable of eliciting useful information about learning within the domain, but it is through *marking guides* (sometimes called 'rubrics') that the direct connection is built back to the learning intentions. For example, it may be decided that a domain-appropriate method for assessing writing ability is to assign a writing task (or perhaps to have students assemble a portfolio of their writing). But the next question is how pieces of student writing should be evaluated. What will be looked for as evidence of quality? This decision needs to be guided by the specification of the learning domain. Typically, several different aspects of students' writing are considered and evaluated (eg, mastery of the conventions of language such as spelling, punctuation

and grammar; and the structuring and presentation of ideas). In assessing complex tasks, these aspects of student work usually are referred to as assessment 'criteria'.

Marking guides, or rubrics, are an essential element of all forms of assessment. An assessment task is not an assessment task until a decision has been made about how responses to, or performances on, that task are to be evaluated. In general, a rubric consists of two or more ordered levels of response to, or performance on, an assessment task. For complex tasks, rubrics may be developed for several assessment criteria. At its simplest, a rubric defines just two levels of response: right and wrong. Alternatively, a rubric may define more than two levels of performance on a task, for example, by recognising the partially complete solution of a problem or varying levels of quality in a response or performance. The distinctions made in developing and using a rubric are always based on qualitative judgements. Even in the case of test questions scored right or wrong, judgements must be made about which responses will be marked right and which will be marked wrong (eg, will 2+2=ϟ be marked right or wrong? What about 2+2=☰?).

When teachers develop assessment tasks, the design of a rubric for evaluating students' performances on that task is an essential element of task design and a crucial component of professional work. The design of task rubrics depends heavily on teachers' expert knowledge and professional judgement.

The fourth step in the assessment process is to bring together records of a student's responses to, or performances on, a number of assessment tasks to draw a *conclusion* about the student's overall level of attainment in the learning area being assessed. It is usual to base overall conclusions of this kind on multiple assessment tasks because individual tasks (unless they are particularly large and complex and generate significant information about a learning area) provide limited information and thus relatively unreliable conclusions.

The simplest way to bring together records of student performance is to sum marks (eg, to count the number of items answered correctly on a test or to sum marks across questions on an examination). Because tests and examinations vary in length, marks often are converted to percentages. But there are several shortcomings of percentages as a way of summarising students' levels of attainment. One shortcoming is that percentages do not represent absolute standards of achievement. What it means to have 85 per cent of questions right on a test or examination depends on the difficulty of the particular questions asked. In practice, it is possible to write both easy questions and hard questions that address the same syllabus, and it is almost impossible to develop two tests with identical difficulties. This means that there is no guarantee that a score of 85% on one test represents the same level of achievement as 85% on another test addressing the same syllabus. If tests become progressively easier year after year, the distribution of students' percentage scores can be maintained over time while absolute levels of achievement steadily decline. This is a problem with examination systems that report student results only as percentages: they have no straightforward way of measuring changes in standards over time.

A second shortcoming of percentages is that they generally do not provide substantive information about what exactly a student has achieved. For example, a score of 85% is difficult to interpret substantively because a score of 85% on an easy test represents a lower level of knowledge, skills and understandings than a score of 85% on a harder test.

In an attempt to circumvent these shortcomings of marks and percentages, the Queensland Studies Authority has introduced instrument-specific 'criteria' and 'standards' for evaluating student responses. This approach can be illustrated using the

sample student assessment booklet for Year 7 science published on the QSA website.[2] The assessment booklet contains fifteen tasks that student are to complete relating to food webs. Students' responses to the fifteen tasks are evaluated in terms of four criteria and five standards labelled A to E (Figure 1). For each criterion (row in Figure 1), teachers decide on the standard demonstrated in students' responses. For example, on the basis of a student's responses to the fifteen tasks, the teacher decides whether the student 'consistently drew well-justified conclusions' (A), 'consistently drew considered conclusions' (B), 'generally drew plausible conclusions' (C), 'occasionally drew valid conclusions' (D), or 'rarely drew valid conclusions' (E).

<table>
<tr><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td></tr>
<tr><td colspan="3">Identifies 4 living things that function as a producer, herbivore or carnivore</td><td>Identifies 2 living things that function as a producer, herbivore or carnivore</td><td>Identifies 1 living thing that functions as a producer, herbivore or carnivore</td></tr>
<tr><td colspan="2">Identifies 2 appropriate food chains from the food web.</td><td>Identifies the producer for both food chains and completes 1 chain.</td><td>Identifies 1 food chain and sections of the other.</td><td>Identifies sections of both food chains.</td></tr>
<tr><td>Consistently draws well-justified conclusions.</td><td>Consistently draws considered conclusions.</td><td>Generally draws plausible conclusions.</td><td>Occasionally draws valid conclusions.</td><td>Rarely draws valid conclusions.</td></tr>
<tr><td>Analysis leads to skilful representation of the data with a complete food web structured according to the feeding hierarchy.</td><td>Analysis leads to accurate representation of the data with a complete food web and an easily discerned feeding hierarchy.</td><td>Analysis leads to proficient representation of the data with a food web that contains isolated pieces of data and/or partially completed food chains.</td><td>Analysis leads to representation of obvious feeding relationships.</td><td>Analysis leads to isolated representations of some feeding relationships.</td></tr>
</table>

Figure 1. Guide to making A-E judgements, Year 7 Sample Student Booklet (food webs)

Each of the fifteen assessment tasks in this student booklet was presumably designed to elicit information about students' knowledge and understandings of food webs. Ideally, each of the fifteen tasks would have had a carefully-designed rubric for evaluating and recording students' responses to that task.

The use of this table of criteria and standards (rather that the qualitative distinctions captured in each of the fifteen task rubrics) as the basis for evaluating students' responses to the fifteen tasks and drawing overall conclusions about achievement in this aspect of science learning introduces a number of complications. First, it is likely that full use will not be made of all of the distinctions available through the application of the fifteen task rubrics and that valuable assessment information will be lost in the process. Second, because there is no explicit relationship between the fifteen task rubrics and the criteria and standards, teachers must make this connection themselves, meaning that judgements may differ from teacher to teacher, introducing unreliability into the system and the possibility of students with the same pattern of task responses being assessed differently. Third, the possibility of this occurring is greatly increased by the fact that the described

[2] We have deliberately chosen a non-senior secondary school example to illustrate this general approach to criteria and standards.

standards are open to interpretation. Fourth, this whole process greatly increases teacher workload by introducing an unusual (and arguably unnecessary) additional demand on teachers.

Best Practice

International best practice in educational assessment proceeds through the set of steps outlined above, beginning with a clearly defined learning domain grounded in discipline knowledge and evidence about how learning occurs within that domain. Assessment methods are chosen on the basis of their domain-relevance (construct validity) rather than personal preference. Students' task responses/performances are recorded using rubrics (or marking guides) that are informed by, and aligned with, the learning domain and learning intentions. Conclusions about where students are in their learning within the area being assessed are then based on evidence provided by (usually multiple) assessment tasks.

When it comes to combining evidence across assessment tasks, research in the field of psychometrics suggests that the simplest, most reliable and fairest method is simply to *sum marks across tasks*. If an assessment booklet consists of twenty tasks all scored right or wrong, then the best way of combing responses to those tasks is simply to assign a score of 0 for a wrong response and a score of 1 for a right response and to sum over the twenty tasks to obtain overall student scores in the range 0 to 20. If an assessment booklet contains some tasks with a 3-level rubric (eg, by recognising the partially complete solution of a task), then the three levels on those task rubrics are best scored 0, 1 and 2, with scores again being summed across all tasks. If a student performance (eg, in dance) or a product of student work (eg, a research project or piece of artwork) is judged on, say three, separate criteria, each with a rubric that defines five levels of quality, then those levels are best scored 0 to 4 and – if an overall assessment is to be made – summed across the three criteria to obtain student scores in the range 0 to 12.

When task responses/performances are combined in this way, the conclusion reached about a student's overall level of achievement in the domain being assessed is an 'on-balance' conclusion. This is because low performances on some tasks (or criteria) can be compensated for by high performances on others. And when all students attempt the same sets of tasks, their total scores on this set of tasks can be compared directly, without attempting to take into account differences in task difficulty. Again, there is good psychometric evidence to support this practice.

The best assessment programs internationally go one step further and, in so doing, address the shortcomings of percentages described above. These programs convert student marks to a numerical scale that can be used to report results on *different* assessment instruments. (The conversion process maintains the rank order of students.) The results of this process are illustrated in Figure 2 for the OECD's Programme for International Student Assessment (PISA). Students' results in this particular aspect of PISA Science ('identifying scientific issues') are reported on the numerical scale on the left of Figure 2.[3] This same scale is used in each cycle of PISA, enabling levels of achievement in this aspect of science learning to be compared across PISA cycles (regardless of unintended fluctuations in test difficulty from cycle to cycle)[4].

[3] 'Identifying scientific issues' is one of a number of aspects of science achievement reported in PISA assessments.

[4] The conversion of marks to the PISA scale is based on a statistical process known as 'item response modelling'. The process makes automatic adjustments to students' marks to take account of differences in test difficulties from cycle to cycle.

Also illustrated in Figure 2 are qualitative descriptions of six levels of achievement along this numerical proficiency scale. By estimating the difficulties of PISA science tasks, it is possible to provide *substantive interpretations* of students' overall marks (sometimes referred to as attaching meaning to marks). Notice that the assessment process itself does not involve making judgements against these described levels. Rather, the levels provide a substantive interpretation of students' overall marks once they have been calculated. The conversion of marks to a general numerical reporting scale and the substantive interpretation of levels of achievement along this scale are features of most major international and national assessment programs, including the most advanced Year 12 assessment systems.

	Proficiency level	General proficiencies students should have at each level	Tasks a student should be able to do
	6	Students at this level demonstrate an ability to understand and articulate the complex modelling inherent in the design of an investigation.	◗ Articulate the aspects of a given experimental design that meet the intent of the scientific question being addressed. ◗ Design an investigation to adequately meet the demands of a specific scientific question. ◗ Identify variables that need to be controlled in an investigation and articulate methods to achieve that control.
708	5	Students at this level understand the essential elements of a scientific investigation and thus can determine if scientific methods can be applied in a variety of quite complex, and often abstract contexts. Alternatively, by analysing a given experiment, can identify the question being investigated and explain how the methodology relates to that question.	◗ Identify the variables to be changed and measured in an investigation of a wide variety of contexts. ◗ Understand the need to control all variables extraneous to an investigation but impinging on it. ◗ Ask a scientific question relevant to a given issue.
633	4	Students at this level can identify the change and measured variables in an investigation and at least one variable that is being controlled. They can suggest appropriate ways of controlling that variable. The question being investigated in straightforward investigations can be articulated.	◗ Distinguish the control against which experimental results are to be compared. ◗ Design investigations in which the elements involve straightforward relationships and lack appreciable abstractness. ◗ Show an awareness of the effects of uncontrolled variables and attempt to take this into account in investigations.
559	3	Students at this level are able to make judgements about whether an issue is open to scientific measurement and, consequently, to scientific investigation. Given a description of an investigation can identify the change and measured variables.	◗ Identify the quantities able to be scientifically measured in an investigation. ◗ Distinguish between the change and measured variables in simple experiments. ◗ Recognise when comparisons are being made between two tests (but are unable to articulate the purpose of a control).
484	2	Students at this level can determine if scientific measurement can be applied to a given variable in an investigation. They can recognise the variable being manipulated (changed) by the investigator. Students can appreciate the relationship between a simple model and the phenomenon it is modelling. In researching topics students can select appropriate key words for a search.	◗ Identify a relevant feature being modelled in an investigation. ◗ Show an understanding of what can and cannot be measured by scientific instruments. ◗ Select the most appropriate stated aims for an experiment from a given selection. ◗ Recognise what is being changed (the cause) in an experiment. ◗ Select a best set of internet search words on a topic from several given sets.
410	1	Students at this level can suggest appropriate sources of information on scientific topics. They can identify a quantity that is undergoing variation in an experiment. In specific contexts they can recognise whether that variable can be measured using familiar measuring tools or not.	◗ Select some appropriate sources from a given number of sources of potential information on a scientific topic. ◗ Identify a quantity that is undergoing change, given a specific but simple scenario. ◗ Recognise when a device can be used to measure a variable (within the scope of the student´s familiarity with measuring devices).
335			

Figure 2. PISA scores and levels for reporting student proficiency in *identifying scientific issues*