CARI CENTRE FOR ASSESSMENT REFORM AND INNOVATION

Designing the Future | Issue 1 | August 2014

Assessment: Getting to the essence

Geoff N Masters

Introduction

A lifetime working in the field has convinced me that assessment in education has become over-conceptualised and over-complicated. Assessment concepts and terminology introduced over the past half century sometimes now function as impediments to clear thinking and good practice; and, worse, the field itself is a mess.

Fault lines fragment the field into dichotomies:

- formative *vs* summative;
- norm-referenced *vs* criterion-referenced or standards-referenced;
- qualitative *vs* quantitative;
- assessment of learning *vs* assessment for learning;
- diagnostic *vs* achievement;
- continuous *vs* terminal; and
- school-based *vs* external ...

... with academic camps often forming around these supposedly different forms and purposes of assessment. Distinctions of these kinds have been enshrined in introductory textbooks and are now passed to each new generation of educators as part of the assessment canon.

A large part of the problem originates in the unhelpful belief that there are multiple 'purposes' of assessment in education. This starting point opens the way for unlimited ways of thinking about assessment, unending concepts and terminology, and unbounded complication – all of which make for an impressively complex academic field, but are not very helpful to clarity or practice.

In reality, there is only *one* fundamental purpose of assessment in education. When this single purpose is recognised and taken as the starting point for thinking about assessment, it becomes a unifying rather than fragmenting influence in the field. I would state this fundamental purpose as follows:

> *The fundamental purpose of assessment in education is to establish and understand where learners are in an aspect of their learning at the time of assessment.*

There is no other purpose. Establishing where learners are in their learning usually means establishing what they know, understand and can do. When this single purpose is appreciated, many invented distinctions become less conceptually fundamental and some concepts can be approached in new and more useful ways.

Assessments can be undertaken at varying degrees of diagnostic detail

Consider, for example, the concept of 'diagnostic' assessment. Attempts often are made to treat diagnostic assessments as a *class* of instruments or methods, leading to debates about whether particular tests belong to this class and are correctly described as 'diagnostic'.

An alternative is to recognise that the question of where learners are in their learning can be answered at differing levels of detail. The question can be answered at a very general level – for example, by establishing a student's overall level of proficiency in a school subject. It can be answered at a more detailed level – for example, by establishing a student's levels of proficiency in a number of different areas of learning within a subject. Or it can be answered at a still finer level – for example, by investigating a student's mastery of specific skills or concepts, and by analysing errors and exploring the misunderstandings that produce them.

Australian Council *for* Educational Research

Because educational assessments are designed to provide information about knowledge, skills and understandings at differing levels of detail, 'diagnosis' is not so much a matter of kind as it is of *degree*. Assessment instruments differ in their diagnostic power in much the same way that microscopes and telescopes differ in the level of detail that they are able to reveal.

In addition, just as assessments can be designed to establish where learners are within an area of learning, in varying degrees of detail, so they can be designed to provide varying degrees of detail about student populations. For example, international sample surveys such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA) are designed to establish where entire national populations of students are in their learning. Depending on sampling designs, these surveys also provide information at a finer level of detail about the performances of subgroups of students (for example, Indigenous students). At a still finer level, other assessments are capable of providing information about how well students are performing in a particular school or a particular classroom; and, zooming in still further, many assessment instruments and methods can be used to establish where individual learners are in their learning.

The point is that, regardless of grain size, the fundamental purpose of assessment in all these contexts is the same: to establish and understand where learners (either as individuals or groups) are in an aspect of their learning at the time of assessment. This can be done in varying degrees of 'diagnostic' detail. An international achievement survey can provide diagnostic information at a high level of generality – for example, by identifying a curriculum area in which students in a particular country are performing relatively poorly. On the other hand, an assessment based on a teacher quizzing a student about how they arrived at a particular answer can provide diagnostic information at a very fine level of detail – for example, by identifying a specific misconception that an individual has developed.

Assessment results can be used in different ways

Although there is only one fundamental purpose of assessment in education, there are many different *uses* to which the results of an assessment process can be put. Intended uses often determine the required degrees of diagnostic detail.

Informing and guiding future action

At all levels of educational decision making, reliable information is required about current levels of student achievement. An understanding of current achievement levels informs starting points for action. Reliable information about the status quo is required across the range of decision makers, including

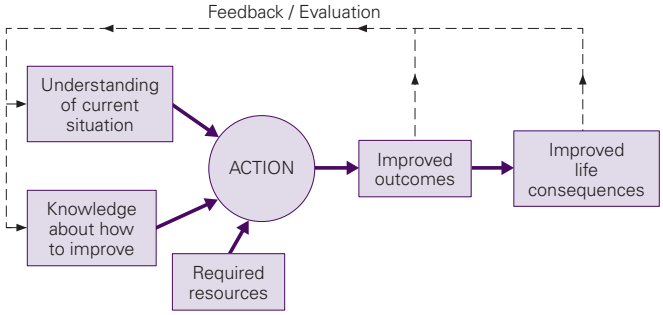
- governments;
- educational policy makers;
- system managers;
- school leaders;
- classroom teachers;
- parents; and
- students themselves.

Reliable information from national and international assessment programs can provide governments and education systems with better understandings of current levels of student achievement, including by identifying areas of underperformance and achievement gaps (for example, between males and females, or between Indigenous and non-Indigenous students).

Reliable information about current levels of achievement also is necessary for assessing past learning progress and evaluating the effectiveness of teaching strategies, policies and initiatives to improve outcomes. Establishing where students have been, and where they are in their learning at the time of assessment, enables the study of growth and trends over time.

Assessments focused on establishing where students are in their learning thus form part of an ongoing decision-making process. They can be used retrospectively to evaluate past progress and prospectively to plan future action. At the level of the classroom, teachers' finer-grained and more diagnostic assessments can provide valuable guidance on appropriate next steps in teaching and learning.

The crucial role of assessment in educational decision-making is illustrated in Figure 1. This diagram is referred to as an educational decision-making 'loop' because it represents an iterative process through which feedback on past decisions and actions informs future practice.



Source: Adapted from Masters, GN (2013, p 10)

Figure 1. Educational decision-making loop

The ultimate purpose of using assessments to guide decision making is to enhance learning and so improve levels of achievement. In other words, this use of

assessment information is *for* improved student learning. This is true whether the decision maker is a national government, system manager, school leader, classroom teacher, parent or student. Assessment *for* learning is not a different form or class of assessments – it is simply the use of assessment information to guide decision making to improve learning outcomes.

Evaluating progress

A second general use of assessment results is to evaluate progress. Once information is available about where learners are in their learning at the time of assessment, this information can be used to evaluate progress since some earlier time. Once again, good information about trends and progress is required by decision makers at all levels, from governments and system managers to parents and students. It is essential to evaluating the impact and effectiveness of policies, programs, interventions and teaching strategies.

The monitoring of progress over time might be described as the monitoring *of* learning. After all, there is no more direct way of evaluating learning success than by monitoring change over time. For this reason, the assessment *of* learning progress is an integral and essential element of effective teaching; but it also is essential to learning itself. Feedback that enables learners to see the progress they are making is crucial to building individuals' self-efficacy as learners, as well as their appreciation of the relationship between effort and success.

Importantly, the assessment *of* learning does not imply a different class of assessments – it is simply the use of assessment information to draw conclusions about progress, whether that progress is at the level of groups (for example, improving performance levels of 15-year-olds) or at the level of individual growth.

Similarly, the concepts of 'formative' and 'summative' assessment are often treated as fundamentally different kinds or classes of assessments. However, the assessment literature is in disagreement about this distinction. Some writers describe the distinction primarily in terms of timing: formative assessments are undertaken at various points during a course of instruction while summative assessments are undertaken at the end of a course. Others describe the distinction primarily in terms of the method of assessment: formative assessments are based on detailed day-to-day classroom observations made by teachers, while summative assessments are based on more formal, often externally developed, tests and examinations. Still others describe the distinction primarily in terms of intended use: formative assessments are used prospectively to identify starting points for teaching and learning, while summative assessments are used retrospectively to determine and report on past learning success.

When assessment is conceptualised as the process of establishing and understanding where learners are in an aspect of their learning at the time of assessment, the formative/summative distinction becomes less fundamental. Information about where learners are in their learning can be used prospectively to identify starting points for future teaching and learning. Such information is generally most useful when it includes fine-grained learning detail, but the same information also can be used retrospectively. By comparing current information about where students are in their learning with previous assessments, it is possible to evaluate past learning progress. Assessments of constructs in other disciplines do not differ depending on whether they are to be used prospectively to plan future action or retrospectively to evaluate past progress. In the same way, the formative/summative distinction is more usefully understood in terms of intended use – recognising that the results of an assessment process can be used either 'formatively' or 'summatively' – rather than as different classes of assessments.

There are many other uses to which assessment results can be put, including to

- allocate scarce resources, such as scholarships and places in competitive educational institutions;
- assign students to courses and remedial programs;
- award credentials; and
- evaluate the effectiveness of educational initiatives.

All these and other uses depend on reliable information about the points that learners have reached in their learning at the time of assessment.

Assessment results can be interpreted in different ways

Other common distinctions are seen to be less fundamental when it is recognised that different frames of reference can be used to *interpret* assessment results. For example, information about where students are in their learning can be interpreted by reference to the performances of other students (comparing with age norms or benchmarking against performances in other countries); by reference to year-level curriculum expectations; or by reference to past levels of performance.

Particular frames of reference are sometimes mistakenly believed to require particular methods of assessment (for example, norm-referencing is sometimes thought to be possible only with multiple-choice tests). However, the interpretation of assessment results generally follows and is independent of the assessment process itself.

The learning domain

Because the fundamental purpose of assessment in education is to establish and understand where learners are in an aspect (or 'domain') of learning at the time of assessment, the primary frame of reference for interpreting assessment results is the learning domain itself. The processes of establishing and describing where learners are in their learning depend on a deep understanding of the domain through which they are progressing. A well-constructed and richly described domain map is essential to the entire assessment process.

Most learning domains have both a horizontal and a vertical structure (for an example of such a learning domain, see Box 1.) Both need to be mapped and understood. The horizontal structure is made up of sub-areas of learning. These may be different content areas (topics or sub-areas of knowledge and understanding) or different skills. The vertical structure, on the other hand, describes how knowledge, skills and understandings develop and change with increasing proficiency. In other words, the vertical structure describes what it means to improve, grow or make progress within the domain, typically over several – and sometimes many – years of learning. Ideally, these descriptions are accompanied by examples of performances and responses that illustrate increasing levels of proficiency.

The mapping of a learning domain is based on the empirical study of how learning occurs within that domain, including by identifying typical sequences and paths of development and the role of prerequisites (such as pre-reading and early reading skills) in successful subsequent learning. A complete mapping of a domain includes the mapping of pathologies – for example, by identifying common difficulties, errors and misunderstandings; and, ideally, the mapping process results in deeper theoretical understandings of the domain.

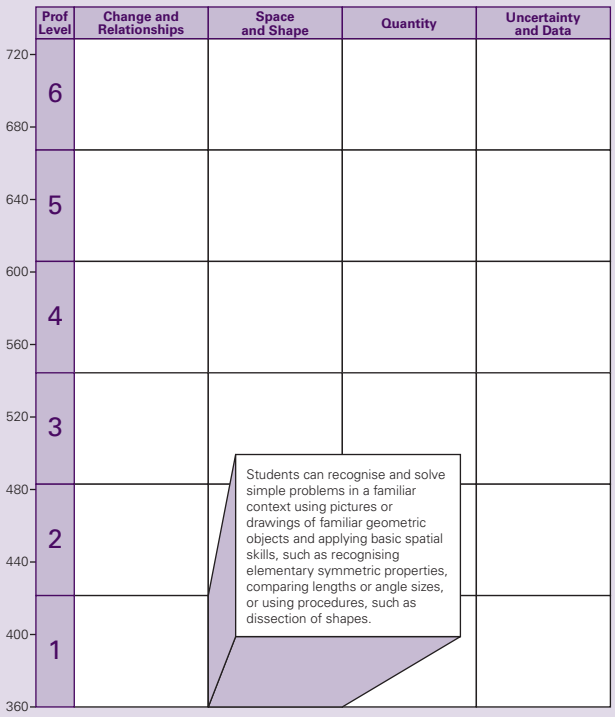
When assessments are made against an empirically-mapped domain, the outcomes of the assessment process can be interpreted by reference to this map. What points have learners reached in their learning, and what does this mean for the kinds of knowledge, skills and understandings that they now demonstrate?

Box 1. Learning domain: An example

The OECD's Programme for International Student Assessment (PISA) uses assessment tasks to establish where students are in their Mathematical Literacy development. The PISA Mathematical Literacy domain has a *horizontal* structure that takes into consideration areas of mathematical content knowledge, differing contexts for applying mathematics, and fundamental mathematical processes. The domain also has a *vertical* structure that describes increasing levels of mathematical proficiency.

Figure 2 shows the structure of this learning domain. Mathematical Literacy is assessed and reported in each of four content areas (Change and Relationships; Space and Shape; Quantity; and Uncertainty and Data). Increasing proficiency is described and illustrated through six Proficiency Levels labelled 1 to 6. Part of the description of the lowest level of proficiency in Space and Shape is shown.

This map of the PISA learning domain is empirically based. In other words, it is derived from an analysis of how students performed on assessment tasks constructed to address the domain. The scale on the far left is used to estimate and report – in a finer level of detail – where students are in their progress through this domain.



Source: Thomson, De Bortoli and Buckley (2013, p 57–8)

Figure 2. The structure of the PISA Mathematical Literacy domain

A specified minimum standard

A second frame of reference for interpreting the results of an assessment process is a specified minimum standard of proficiency. A minimum standard may be an expected level of proficiency, such as the reading level expected of students by the end of Year 5, or a requirement, such as the level of proficiency required to fly an aeroplane or to practise surgery. The setting of a minimum standard is always a matter of professional judgement. Usually, interest then focuses on whether or not learners have reached this point in their learning, with the conclusion being recorded as a yes/no.

Past performance

A third frame of reference is past performance. When assessments are referenced to past performances, they can be used to evaluate the progress made since earlier assessments. How much progress has an individual made in her/his learning? Have achievement levels improved over the past decade? Is there a clear trend? Usually, interest focuses on gains, growth trajectories, rates of progress and trends over time.

The performance of others

A fourth frame of reference is the performance of other learners. Once information is available about where learners (either as individuals or groups) are in their learning, this information can be compared with the performances of other relevant groups of learners.

- How does a student's level of mathematics achievement compare with the mathematics achievements of other students of the same age or year level (that is, age-level or year-level 'norms')?
- How do the performances of students in a given school compare with performances in 'like schools' with similar student intakes?
- How do national levels of science achievement at 15 years of age compare with achievement levels in other countries?

Usually, interest focuses on where learners stand in comparison with a relevant reference population. For example, a student's progress through a learning domain may put her/him among the most advanced ten per cent of her/his age group (that is, above the 90th percentile); achievement levels in a school may place that school in the bottom third of like schools; or average achievement levels in a nation may place that country among the top five nations in the world.

Provided that the relevant frame of reference is available, each of these four ways of interpreting assessment results can be used with any assessment method – from classroom assessments and standardised tests to national and international surveys. They also can be used to interpret assessments before, during or upon completion of a course of instruction or, in most areas of learning, without reference to a course of instruction at all.

Assessments can be based on a variety of observation methods

Finally, when assessment is understood as the process of establishing where learners are within a learning domain at the time of assessment, the role of assessment tasks is clarified. Assessment tasks provide *observations* for drawing conclusions (or inferences) about the points that learners have reached in their learning.

Assessment tasks are never important in themselves. They are transient and interchangeable. Students may never again encounter the specific problems on a mathematics test, or the passages and questions asked in a reading assessment. Such tasks are simply convenient opportunities to gather evidence about what is really of interest – a student's underlying mathematics knowledge or level of reading comprehension. Individual tasks are important only to the extent that they elicit observations helpful in inferring where learners are in their learning. They provide concrete observations for inferring the unobservable.

In practice, assessment activities sometimes stop short of their essential purpose. They stop at the point of recording how students perform on a particular task or set of tasks. However, no matter how large and complex a task, and no matter how impressive the rubric for recording responses to that task, it is still only one of an unlimited number of possible domain-relevant tasks. In addition, when a test score is calculated by counting correct answers, that score is nothing more than a record of how students performed on that particular set of questions. The assessment process must go beyond recording how students perform on specific tasks to its central purpose of *inferring* from task-specific observations where learners are within the relevant learning domain.

It is also important to recognise that no one observation method is inherently superior to any other. Complex assessment tasks, set in real-world ('authentic') contexts, may provide more valid evidence about some kinds of learning, than simpler tasks may provide. Teachers' classroom observations, similarly, may provide more valid evidence about some kinds of learning than externally developed tests may provide. Open-ended ('constructed response') tasks may provide more valid evidence about some kinds of learning than multiple-choice tests may provide. However, all of these, and many other, observation methods are capable of providing valuable information about specific kinds of learning.

Too often, advocates of particular assessment methods (for example, school-based assessment, standardised testing, performance/authentic assessment) fail to acknowledge the ability of other forms of assessment to provide valid and reliable information about specific kinds of learning. Assessment methods must be chosen not on the basis of philosophical positions or personal preferences, but on their demonstrated capacity to provide domain-relevant observations.

In summary

Current conceptualisations of educational assessment revolve around distinctions based on

- varying 'purposes' (for example, diagnostic, formative, summative, assessment of learning, assessment for learning); or
- 'methods' (for example, standardised tests, classroom observations, performance assessments, written examinations, authentic tasks).

These distinctions often are presented as dichotomies. Often the result is unhelpful fragmentation of the field, with proponents championing one assessment purpose or method while denigrating others.

Advances in assessment theory and practice require a more unified conceptualisation. The starting point is to recognise that there is only one fundamental purpose of assessment in education: to establish and understand where learners are in an aspect of their learning at the time of assessment. This question can be answered in varying degrees of diagnostic detail.

To establish the points that learners have reached in an area of learning, an empirically – and, ideally, theoretically – derived map of the learning domain is required. This map has both a horizontal structure (for example, sub-areas of knowledge and skills) and a vertical structure (descriptions of long-term learning progress). It describes and illustrates learning within the domain.

The assessment process involves making observations that can be used to infer where learners are in their learning progress within a domain. The essential question of any assessment method is whether it is capable of providing valid observations about the domain of interest. No method is inherently superior to any other; methods capable of providing valid information for some aspects of learning will be invalid for others. Whichever method is used, the result is always a set of task-specific observations. The next and crucial step in the assessment process is to infer from those observations where learners are in their progress within the relevant domain.

When assessment is conceptualised in this way, many supposedly important distinctions become less significant. Information about the points that learners have reached in their learning can be used both prospectively, to identify starting points for future teaching and learning, and retrospectively, to evaluate past learning progress (the assessment of learning). The results of an assessment process also can be interpreted against different frames of reference: the domain itself (criterion referencing), minimum standards, past performances, and the performances of others (norm referencing).

References

Masters, G N (2013) *Reforming Educational Assessment: Imperatives, Principles and Challenges*, Australian Education Review No.57, Australian Council for Educational Research (ACER), Melbourne. Editor's note: In CSE Occasional Paper 135, Professor Masters has synthesised themes and arguments that he expounded in more detail in Australian Education Review No.57, which may be accessed at http://research.acer.edu.au/cgi/viewcontent.cgi?article=1021&context=aer

Thomson, S, De Bortoli, L and Buckley, S (2013) *PISA 2012: How Australia Measures Up*, Australian Council for Educational Research (ACER), Melbourne.