The Assessments We Need

Geoff N Masters & Margaret Forster

ACER

The Assessments We Need

Geoff N Masters
Margaret Forster
Australian Council for Educational Research

Large-scale assessment programs have an important role to play in providing dependable information for educational decision making by policy makers, system managers, school leaders, teachers and parents. But these programs—which include international achievement studies, national surveys and system-wide tests—also convey powerful messages about the kinds of learning valued by education authorities and can have a profound impact on teaching and learning, particularly if results are reported and compared publicly. For this reason it is essential that large-scale programs are designed to reflect and reinforce learning priorities. This paper argues that large-scale programs are most likely to support the kinds of learning now considered important for successful functioning in society if they are designed with the primary purpose of collecting reliable information about students' current levels of achievement, if they incorporate assessments of higher-order skills and thinking, if they include a variety of assessment methods and procedures capable of providing information about a range of valued learning outcomes, and if results are reported in ways that recognise and encourage high achievement.

In education, as in other areas of professional practice, good decision making is facilitated by access to relevant, reliable and timely information. Dependable information is required at all levels of educational decision making—from student, to parent, to classroom teacher, to school principal, to system manager—to identify areas of deficiency and special need, to monitor progress towards goals, and to evaluate the effectiveness of special interventions and initiatives.

Many indicators provide useful input to educational decision making. But the most important indicators in education are those which address the central concern of the educational enterprise: the promotion of student learning. Because the ultimate goal of all education is to encourage and facilitate personal development in the interests of both individuals and societies, almost all decisions and courses of action in education ultimately must be judged on the extent to which they result in improved student learning. For this reason, measures of learning outcomes have a special significance as indicators of quality in education, as a basis for setting targets for improvement, and in monitoring progress towards educational goals.

Informed Decision-Making

Dependable information on educational outputs is required by *system managers* if they are to exercise their responsibilities for the delivery of quality education to all students in a system. Effective management depends on an ability to monitor systemwide performances over time, to gauge the effectiveness of special programs and targeted resource allocations, to monitor the impact of systemwide policies, and to evaluate the success of initiatives aimed at traditionally disadvantaged and underachieving sections of the student population. Accurate, reliable data allow system managers to measure the progress of a system against past performances, to identify areas requiring special attention, and to set goals for future improvement.

Equally, reliable data on educational performances are required for effective *school management*. Research into factors underlying school effectiveness highlights the importance of the school manager's role in establishing an environment in which student learning is accorded a central focus, and goals for improved performance are developed collaboratively by staff with a commitment to achieving them. School managers require dependable pictures of how well students in a school are performing, both with respect to school goals for improvement and with respect to past achievements and achievements in other, comparable schools.

Feedback to *parents* on student achievement is required if they are to become active partners in their children's learning. Valid and reliable information empowers parents to evaluate the quality of the education their children are receiving and to make informed decisions in the best interests of individual learners. If parents are to make informed decisions and take an active role in their children's learning, then they require dependable information about the progress individuals have made (the knowledge, skills and understandings developed through instruction) and about teachers' plans for future learning. There is considerable evidence that parents also want information about how their children are performing in comparison with other children of the same age.

From the point of view of *classroom teachers,* systematically-collected information on student learning can be useful both for classroom decision making and for wider reporting purposes. All teachers, and especially beginning teachers, are capable of benefiting from reliable information about what is being expected of, and achieved by, students of a given age —information not contained in any one teacher's classroom assessments. When teachers have access to explicit frames of reference in the form of achievement targets at particular Year levels, are given assistance in assessing student progress in relation to those targets, and receive feedback on how their classes are performing in comparison with students in other schools, they have a powerful basis for structuring and evaluating individual and classroom learning.

Students, too, benefit from feedback that enables them to monitor their own progress and to set goals for further learning. Among the skills students are likely to require in a workforce responsive to future change are the skills of self-management and independent, ongoing learning: the abilities to organise one's own learning, reflect on progress, identify areas requiring special attention, and to plan future work. Students are more likely to become successful, independent learners when they are encouraged to appreciate learning as a lifelong process of individual growth through the development of new skills, deeper understandings, and more positive attitudes and values.

Large-scale assessment programs have a significant contribution to make in providing information that can be used by each of these stakeholder groups to make decisions in the interests of improving student learning.

Communicating Values

As well as providing useful information for educational decision making, large-scale assessment programs play an important role in communicating values and expectations. Whatever might be espoused in policy documents, syllabus statements and classroom teaching, procedures for assessing school learning send powerful messages about what societies value and consider worthy of recognition and reward.

Because assessment procedures can be a potent influence on what teachers teach and students learn, the design of assessment programs and their impact on teaching and learning must be kept under constant review. A well-designed assessment system can be an effective means of focusing students' attention on valued learning outcomes, encouraging higher-order thinking and reflection, reinforcing curriculum intentions, and setting learners' sights on still higher levels of attainment. Well-designed systems, as well as operationalising and communicating the kinds of thinking and learning we wish to encourage in students, can provide a basis for valuable conversations among teachers about learning and its assessment, and between teachers, students and parents about individuals' current levels of progress, their strengths and weaknesses, and the kinds of learning experiences likely to be effective in supporting further learning. A poorly-designed system, on the other hand, may provide little support to learning and, at worst, may distort and undermine curriculum intentions, encourage superficial learning, and lower students' sights on satisfying minimal requirements.

The design of an assessment system also can influence teacher behaviour, particularly when assessment results are used to draw conclusions about the performances of individual teachers or schools, or to allocate resources. An example of this influence was the impact of 'minimum competency' testing on teacher behaviour in the United States. Minimum competency tests were introduced in the 1970s and 1980s to establish whether students were achieving the minimum levels of knowledge and skill expected of students in particular grades (eg, end of high school). As many commentators have observed, a common response by American teachers to minimum competency tests was to focus their teaching efforts on the foundational skills assessed by these tests and to concentrate their attention on students who had not yet achieved these skills—sometimes at the expense of extending the knowledge and skills of higher achieving students. According to some writers, these tests not only constrained classroom teaching, but also had dubious benefits for the students they were designed to serve:

> *Minimum competency tests are often used as a policy tool to require that students meet some basic level of achievement, usually in reading, writing and computation, with the intention that the use of these tests will lead to the elimination of these educational problems… [However,] the empirical findings show that the negative consequences far outweigh the few positive results… For example, Griffin and Heidorn (1996) showed that minimum competency tests do not help those they are most intended to help—students at the lowest end of the achievement distribution… There have been several studies focusing on the effects of minimum competency testing on curriculum, teaching, and learning. Most of these studies have been critical of their negative effects on curriculum and instruction.*
>
> *(Marion and Sheinker, 1999)*

Other writers are less damning, pointing to evidence of a relationship between minimum competency testing and improved performances among lower-achieving students (Frederiksen, 1994). Nevertheless, as Mehrens (1998) notes, because minimum competency tests generally were perceived not to have been effective in raising educational standards, by the 1990s there was a trend away from large-scale tests focused on the achievement of minimally acceptable standards to tests focused on newly-valued 'world class' standards.

Over recent decades, we have learnt a great deal about the ways in which large-scale assessment programs convey values and impact on practice, and have become more aware of the unforeseen and unintended consequences of particular approaches to assessment. For example, we have discovered the tendency of criterion-referenced (and more recently competency-based) assessment systems to fragment curricula into unhelpful checklists of narrow skills and behaviours. We have seen the tendency of minimum competency tests to restrict the attention of teachers and learners to the achievement of fundamental knowledge and skills. And we better understand the difficulties of using complex performance assessments in high stakes settings, particularly as they relate to teacher workload, the authentication of student work, and the achievement of adequate levels of reliability and comparability.

Through trial and error and a considerable body of systematic research in a number of countries, we have developed a deeper understanding of factors that must be taken into consideration in the design of large-scale assessment programs. These factors include considerations of curriculum fidelity, reliability and comparability, fairness and access, feedback to instruction, and practicability. Importantly, we have learnt more about the interplay between these considerations: for example, how attempts to maximise reliability can distort curricula, and how efforts to maximise useful feedback to instruction can render large-scale assessment programs unmanageable (Dietel *et al*, 1991).

Because large-scale programs can communicate powerful messages to students and teachers about valued forms of learning, and because they have the potential to influence what schools and teachers do, it is important that these programs are designed to reflect and reinforce, rather than distort and undermine, curriculum intentions.

Some Design Principles

In this paper we consider some general requirements of large-scale assessment programs if they are to provide useful feedback to decision making and be consistent with—and so reinforce—current curriculum priorities. There are many design features of assessment programs; our focus here is on just a few macro features and the principles that underlie them. In particular, we focus on the kinds of learning addressed, the range of assessment methods used, and the ways in which student achievements are summarised and reported. We argue for:

1. designing assessment procedures primarily to establish where all students are in their learning;
2. incorporating assessments of higher-order skills and thinking;
3. including a variety of assessment methods and procedures to provide information about a range of valued learning outcomes; and
4. reporting results in ways that encourage high achievement.

Establishing where all students are in their learning

Most large-scale programs are designed with the intention of estimating students' current levels of achievement in particular areas of learning. These programs recognise that, within any given Year level, students have very different levels of attainment, and assessments are designed to provide reliable information about the achievement levels of individual students or, in sample-based programs, reliable estimates of the achievement levels of the student population.

Underlying such an approach is the notion that, at any given time, students are distributed along a continuum of achievement in the learning area under consideration. In reading, for example, students are assumed to vary in their levels of reading development. The purpose of these programs is to provide reliable estimates of students' current levels of reading achievement, thus enabling the distribution of students' reading achievements to be mapped and studied.

The National School English Literacy Survey (Masters and Forster, 1997) was designed with this intention. In that study, we constructed achievement continua in reading, writing, speaking, listening and viewing, and developed descriptions and provided examples of students' work and performances at various levels along each continuum. The distributions of national samples of Year 3 and Year 5 students were then mapped and reported on these continua.

In contrast to assessment systems designed primarily to provide reliable estimates of all students' current levels of achievement in an area of learning are systems designed with the main purpose of establishing whether or not students have satisfied some minimum requirement. Under this alternative design, the main question in relation to individuals is: has the student passed or failed the minimum requirement? The main question in relation to student populations is: what percentage of students has met the minimum requirement? The interest under this alternative approach tends to be less in establishing where students are on a *continuum* of achievement, and more in establishing whether they are above or below some specified *point*.

The difference between a procedure designed primarily to establish individuals' locations on a continuum and a procedure designed solely to make a pass/fail decision is illustrated in Figure 1. In this picture, the instrument on the right is ideally suited to measuring individuals' heights; the instrument on the left is ideally suited to making pass/fail decisions in relation to a minimum height requirement.

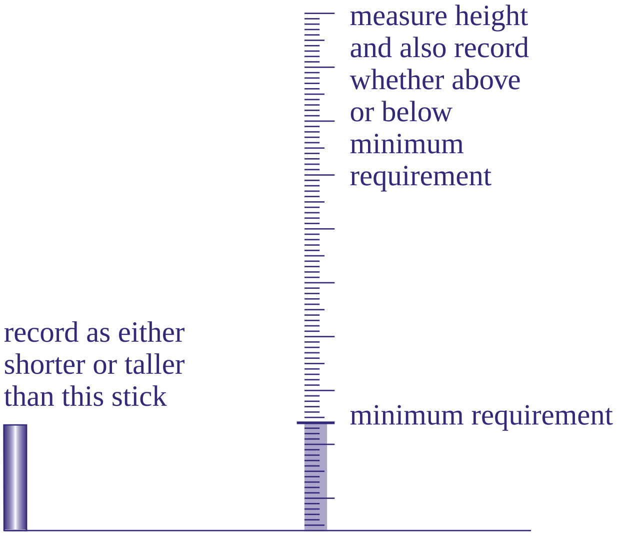

measure height and also record whether above or below minimum requirement
record as either shorter or taller than this stick
minimum requirement

Figure 1. two approaches to assessing whether a minimum standard is met

Procedures for assessing educational achievement similarly can be designed either with the primary purpose of measuring students' current levels of achievement in an area of learning, or with the primary purpose of establishing whether or not a student has met a minimum requirement. Although not as marked as the extreme situation portrayed in Figure 1, tests designed for these two purposes tend to have different characteristics. A test designed primarily for the first purpose provides tasks across a range of difficulty levels, deliberately including tasks that challenge and provide reliable information about high achieving students. In a test designed primarily for the second purpose, most items are located close to the minimum requirement (because, statistically, the way to maximise the reliability of pass/fail decisions is to locate all items at this point).

Marion and Sheinker (1999) refer to an assessment system designed to provide information about the current achievement levels of students along an achievement continuum as a 'standards-based' system and distinguish such a system from a 'minimum competency' test designed solely to establish whether students are above or below some minimum requirement:

> *The design of a standards-based assessment system will include goals that citizens and policymakers would eventually like all students to reach while providing meaningful information about the current achievement levels of students. This will include information about students all along the achievement continuum, so that the identification of students at risk of academic failure [ie, below a minimum requirement] would be accommodated by a standards-based testing system. Using the standards-based system to identify low-achieving students will fulfil the goals of a minimum competency test, but will do so in a much more meaningful and less distracting way.*

'Standards-based' systems include tasks of varying difficulties to provide useful information about students across a wide range of achievement levels. When these tasks are calibrated along an achievement continuum, it is not necessary for all students to attempt exactly the same tasks. Easier tasks can be assigned to lower achieving students, and more difficult tasks can be assigned to higher achieving students without advantaging one group of students over the other, and without reducing the comparability of their results. Provided that all tasks are appropriately calibrated, teachers can choose assessment tasks suited to individuals or groups of students, or a preliminary 'locator' test can be used to determine the most appropriate set of assessment tasks to administer to each individual. In the most flexible version of this system—a 'computer adaptive' test—items are selected one at a time based on a student's performance on all preceding items.

Open-ended tasks which permit different levels of response also can be useful for estimating students' achievement levels along a continuum. For example, the same essay prompt usually can be administered to students with very different levels of writing ability, and performances on several prompts can then be used to locate students along a continuum of increasing writing competence.

An assessment system designed originally with the intention of estimating and reporting all students' levels of achievement in an area of learning can begin to take on the characteristics of a minimum competency test if a decision is made to set a minimum performance standard and to report publicly the percentage of students achieving that standard. Under these circumstances, a decision sometimes is made to increase the reliability of pass/fail decisions by increasing the number of test items at and around the minimum standard. If the stakes are raised (eg, by comparing schools or education systems only on the basis of the percentage of students achieving the minimum standard), then there may be a reduced incentive for maintaining an assessment system that challenges and provides reliable information about the achievements of students performing well above the minimum.

We believe that large-scale programs should maintain as their primary focus the estimation of students' levels of attainment along a continuum of achievement. Such systems are capable of also providing information about whether or not individual students have met a minimum required level of performance. However, to the extent that the items in a test begin to be targeted on only a minimum performance standard, they become less effective in defining a range of achievement levels and in providing reliable information about all students' achievements.

Incorporating assessments of higher-order skills and thinking

In today's world it is necessary, but not sufficient, for students to achieve minimal competence in areas such as reading, writing and numeracy. Beyond the achievement of minimal competence, students also need to develop critical literacy and numeracy skills of the kind required for effective functioning in everyday life. Skills of this kind are now widely advocated by school systems and are adopted as a starting point in most international assessment programs:

International Adult Literacy Survey

Reading literacy is no longer defined merely in terms of a basic threshold of reading ability which everyone growing up in developed countries is expected to attain. Rather, literacy is now equated with an individual's ability to use written information to function in society. *(Kirsch and Murray, 1998)*

Programme for International Student Assessment

Reading literacy: Understanding, using and reflecting on written texts, in order to achieve one's goals, to develop one's potential, and to participate in society. *(OECD, 1999a)*

International Life Skills Survey

Numeracy involves abilities, behaviours and processes which include identifying, interpreting, acting upon and communicating about mathematical information flexibly in real contexts and situations, to enable full, critical and effective participation in a wide range of life roles. *(OECD, 1999b)*

Programme for International Student Assessment

Mathematical literacy: Identifying, understanding and engaging in mathematics and making well-founded judgements about the role that mathematics plays, as needed for an individual's current and future life as a constructive, concerned and reflective citizen.

(OECD, 1999a)

Minimal reading proficiency involves an ability to decode text, to interpret word meanings and grammatical structures, and to understand meaning at least at a superficial level. But reading literacy for effective functioning in modern society requires much more than this: it also depends on an ability to read between the lines and to reflect on the purposes and intended audiences of texts, to recognise devices used by writers to convey messages and to influence readers, and the ability to interpret meaning from the structures and features of written materials. Reading literacy depends on an ability to understand and interpret a wide variety of text types, and to make sense of texts by relating them to the situations in which they appear.

In the context of the US National Assessment of Educational Progress, Applebee et al (1987) argue that, in today's world, 'literacy' consists of much more than the ability to extract surface meaning from texts and to express basic ideas in writing:

> *There are two important components of literacy: (1) the ability to derive surface understanding from written materials and to express similar understanding in writing; and (2) the ability to reason effectively about what one reads and writes in order to extend one's understanding of the ideas expressed.* *(Applebee et al, 1987)*

These authors go on to point out that in national assessments of reading and writing in the United States, most children and young adults can understand what they read and can express their thoughts in writing at a surface level. However, only a small percentage can reason effectively about what they read and write—in other words, function as 'literate' consumers of text, including being able to recognise the techniques through which writers seek to influence and occasionally manipulate readers.

Numeracy similarly depends on a familiarity with a body of mathematical knowledge and skills. Basic number facts and operations, working with money, and fundamental ideas about space and shape, including working with measurements, form part of this essential body of knowledge and skills. But numeracy for effective functioning in modern society requires much more than this: it also depends on an ability to think and work mathematically, including modelling and problem solving. These competencies include knowing the extent and limits of mathematical concepts, following and evaluating mathematical arguments, posing mathematical problems, choosing ways of representing mathematical situations, and expressing oneself on matters with a mathematical content. Numeracy depends on an ability to apply these skills, knowledge and understandings in a variety of personal and social contexts.

In the National School English Literacy Survey (NSELS) we developed and administered assessment tasks with a wide range of difficulties appropriate to Year 3 and Year 5 students. These tasks included a significant number of tasks designed to assess students' abilities to infer meaning, to reflect on writers' purposes and stances, and to interpret writers' uses of linguistic structures and features. The intention of these tasks was to assess not only literal comprehension, but also students' abilities to reason about text and to reflect on the ways in which meaning is conveyed through text—the kinds of skills that Applebee *et al* (1989) argue are essential to modern definitions of 'literacy'.

Radio Telescopes

One way of exploring the Universe from the Earth's surface is by radio telescope. Many objects in space, exploding galaxies for example, give out energy in the form of radio waves. These waves penetrate the Earth's atmosphere and can be picked up by radio dishes like the one at Parkes, NSW. In this way astronomers can 'see' what is happening in space.

QUESTION *The article on radio telescopes says in this way astronomers can 'see' what is happening in space. Why is the word 'see' in inverted commas?*

Lovely Mosquito

Lovely mosquito, attacking my arm
As quiet and still as a statue,
Stay right where you are! I'll do you no harm—
I simply desire to pat you.

Just puncture my veins and swallow your fill
For nobody's going to swat you.
Now, lovely mosquito, stay perfectly still—
A SWIPE! and a SPLAT! and I GOT YOU!

Doug MacLeod

QUESTION *Does the writer think the mosquito is lovely?*
Explain your answer.

Odd Spot

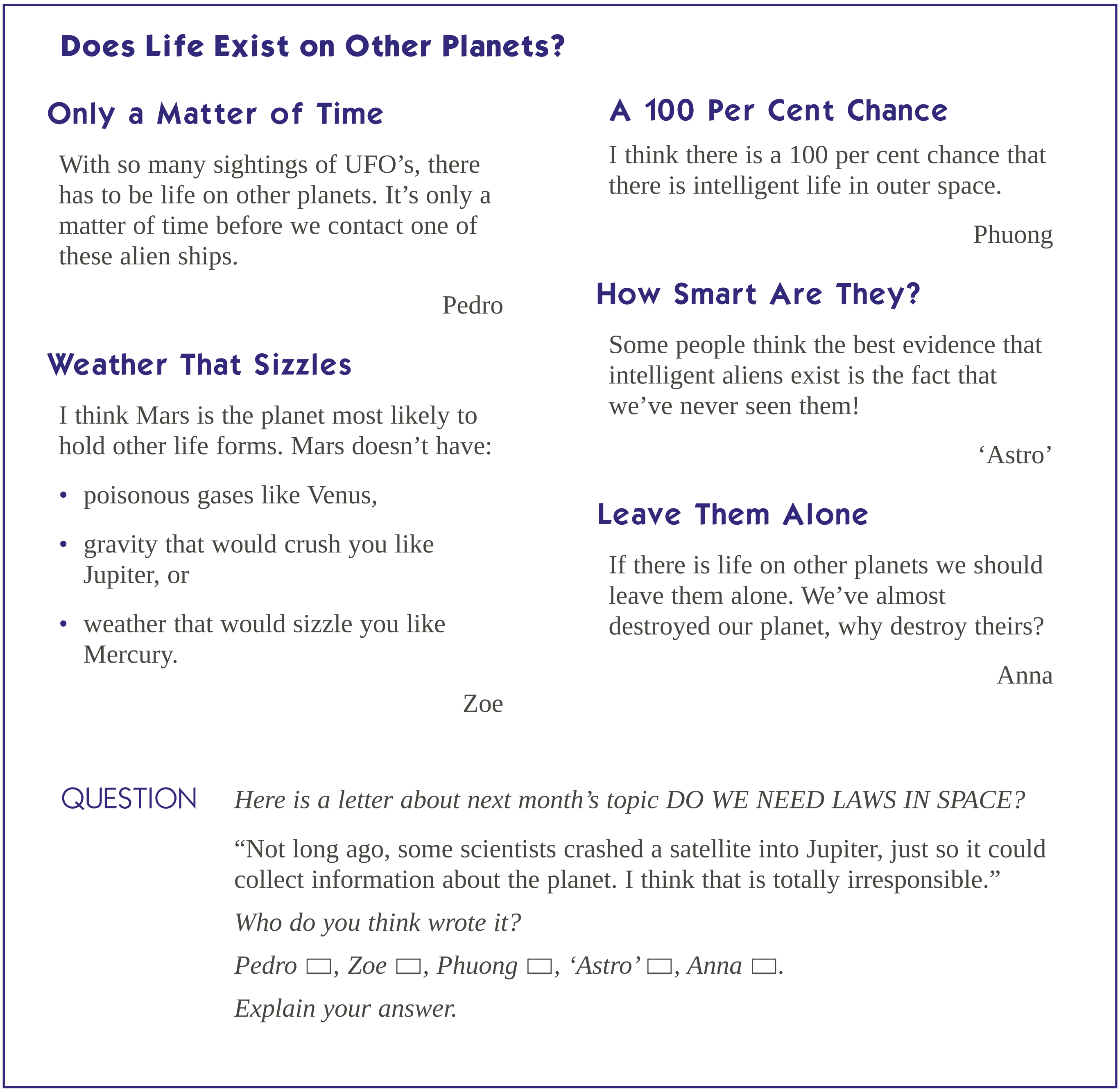
A scientist scanning the Universe with the radio telescope at Parkes NSW, picked up the same strange signal, at the same time, every night for four months. Convinced he was being signalled by an intelligent alien life form, he began an in-depth investigation— only to find he was picking up signals from the microwave in the canteen downstairs!!

QUESTION *Why are there exclamation marks at the end of the Odd Spot piece of writing?*

Figure 2. reading questions from the National School English Literacy Survey

Some of the reading tasks we used in NSELS are reproduced in Figures 2 and 3. These tasks go well beyond surface meaning and explicitly-stated detail, and require students to read between the lines and to reason about text.

When assessment programs include tasks that challenge students to think and to apply their learning, these programs set high rather than low expectations of performance. They also provide teachers and students with examples of the kinds of understanding and thinking to which all students should be aspiring.



Does Life Exist on Other Planets?

Only a Matter of Time

With so many sightings of UFO's, there has to be life on other planets. It's only a matter of time before we contact one of these alien ships.

Pedro

Weather That Sizzles

I think Mars is the planet most likely to hold other life forms. Mars doesn't have:

- poisonous gases like Venus,
- gravity that would crush you like Jupiter, or
- weather that would sizzle you like Mercury.

Zoe

A 100 Per Cent Chance

I think there is a 100 per cent chance that there is intelligent life in outer space.

Phuong

How Smart Are They?

Some people think the best evidence that intelligent aliens exist is the fact that we've never seen them!

'Astro'

Leave Them Alone

If there is life on other planets we should leave them alone. We've almost destroyed our planet, why destroy theirs?

Anna

QUESTION *Here is a letter about next month's topic DO WE NEED LAWS IN SPACE?*

"Not long ago, some scientists crashed a satellite into Jupiter, just so it could collect information about the planet. I think that is totally irresponsible."

Who do you think wrote it?

Pedro □, *Zoe* □, *Phuong* □, *'Astro'* □, *Anna* □.

Explain your answer.

Figure 3. a reading question from the National School English Literacy Survey

This is not to say that tasks of the kind shown in Figures 2 and 3 will be appropriate for all primary school students. Clearly they will not. Some students will be at much lower levels of reading ability, and for those students, significantly easier reading tasks will provide better information about their levels of achievement and progress. A few of these students will require additional diagnostic testing to identify the nature of obstacles to their reading development. Our point here is that large-scale programs, if they are to communicate high expectations of student achievement, must be designed to assess not only foundational knowledge and skills, but also higher order-skills, including students' abilities to apply their learning and to reason about the material they encounter.

Using a variety of assessment methods and procedures

It is common in large-scale programs to limit assessment tasks to test items that can be scored quickly and automatically by machine. The advantages of machine-scored tests are that they can be scored reliably, require minimal class time, and enable a quick turn-around of results.

Tests of this kind are well suited to the collection of information about students' factual and procedural knowledge, their ability to extract and infer detail from text, and their ability to apply basic concepts. Machine-scored tests also are capable of providing information about some higher-order skills.

A number of large-scale programs also include open-ended questions and writing tasks to assess students' abilities to write under test conditions, to construct arguments, describe procedures, develop narratives, and critically reflect on provided material. Students' written responses usually are collected and marked centrally to ensure consistency of marking and comparability of results.

In these programs, classroom teacher involvement typically is limited to the administrative role of proctor: distributing papers, supervising testing, and collecting and despatching answer sheets to a central agency for marking. Teachers may receive feedback in the form of student scores and tables of results that allow them to compare school and/or class performances with performances in other schools and to identify areas in which their students are performing unusually poorly or well.

Although existing large-scale programs provide valuable information about a range of learning outcomes, they are less well suited to many skills now considered important to effective functioning in society—skills such as collecting, sifting, analysing and evaluating information, and communicating and working with others to apply knowledge and skills to the solution of complex real-world problems. Skills of these kinds present very considerable challenges to large-scale assessment programs. Some require direct observation and professional judgement of student work and behaviour, perhaps over a period of time. Some may be difficult if not impossible to assess in large-scale programs. But, in our view, these are not reasons for giving up on their assessment, or for restricting assessments only to skills that can be measured with conventional tests. The risks of focusing the efforts of teachers and students on only a limited range of outcomes are too great to not attempt to assess these more challenging learning outcomes.

The involvement of classroom teachers in the observation and judgement of student work provides a way of assessing outcomes not readily assessed with paper and pen tests. And the direct involvement of teachers in the assessment process may have other advantages. For example, the involvement of teachers in the assessment of their own students' performances and work may give them insights that they are denied when their involvement is limited to the distribution, collection and despatch of test papers for central marking.

Beyond this, classroom teachers may benefit professionally from their exposure to models of good assessment practice, from their efforts to apply professionally constructed marking guides, and from conversations with colleagues about their judgements of student work. Indeed, there are probably few professional development activities more powerful in promoting teacher learning than the peer analysis and discussion of student work. Dialogue around students' responses, performances and work has the potential to clarify intended learning outcomes, develop improved understandings of the ways in which individual achievement can be recognised and monitored, raise awareness of the very different stages that students are at in their learning, and encourage teachers to analyse and reflect on their own instructional practices. Assessment programs that provide classroom teachers with no opportunity to analyse and discuss student work provide little or no opportunity for teachers to make a professional contribution or to develop their own professional skills in the process.

The involvement of classroom teachers in the analysis and discussion of their own students' work was a feature of the National School English Literacy Survey. In that survey, national samples of Year 3 and Year 5 teachers used provided marking guides to assess students' performances on professionally constructed tasks in reading, writing, listening, speaking and viewing. All participating teachers attended training workshops where they were introduced to the survey tasks and marking guides and had opportunities to discuss and assess provided samples of student work. Back in their classrooms, teachers administered the survey materials and made assessments of their students' responses to and performances on the provided tasks. In each State and Territory a number of specially trained assessment 'experts' visited classrooms to observe and assist in making judgements of student work. Finally, all student work was returned to the Australian Council for Educational Research where samples of work were re-marked as a check on national consistency.

Figure 4 provides a brief description of the speaking assessment activities used in the National School English Literacy Survey. The speaking assessments had two components: a set of provided speaking tasks (small group discussions and presentations) administered and assessed by teachers, and teachers' judgements of individuals' speaking achievements based on observations made in day-to-day teaching. Feedback from teachers participating in the survey indicated that, while these tasks were demanding in terms of the time they required, teachers greatly valued the professional opportunity they provided to analyse and discuss this aspect of student learning with colleagues.

The difficulties of incorporating assessment tasks of this kind into large-scale assessment programs are not to be underestimated. And yet, for the reasons we outline in this paper, we believe it is exactly this challenge that the developers of international and national surveys and system-wide assessment programs should be taking up. The risks in not addressing this challenge are that the sights of students and teachers are not raised above the kinds of skills and learning that can be assessed using conventional paper and pen tests. Pellegrino *et al* (1998) make a similar observation in the context of the US National Assessment of Educational Progress:

> *[Current national assessments] do not fully capitalize on current research and theory about what it means to understand concepts and procedures, and they are not structured to capture critical differences in students' levels of understanding. Thus, they do not lead to portrayals of student performance that deeply and accurately reflect student achievement. The development of such portrayals will require the use of multiple methods for measuring achievement that go beyond current large-scale assessment formats … It is clear that [these formats] are not adequate for assessing complex aspects of achievement described in current frameworks. Nor are they adequate for assessing broader conceptualizations of achievement that are consonant with the more comprehensive goals for schooling that will be prominent in the 21st century.*

Speaking Assessment Activities

Students' levels of speaking achievement were assessed by having them complete a set of speaking tasks under controlled conditions ('common tasks') and by collecting records of speaking performances in the classroom within specified categories ('best work'). Classroom teachers made judgements about the quality of their students' speaking using provided assessment guides.

COMMON TASKS

The speaking common tasks completed by students included:

- narrative presentation (telling a story or poem to entertain); and
- argument/opinion presentation (offering an opinion to convince a listener).

Year 3 students retold their favourite narrative, and reviewed a character from the provided videotape. Year 5 students talked about their favourite TV show and discussed a poem in small groups in preparation for individual presentation and commentary.

Individual presentations required students to consider the ways in which spoken text is used to communicate meaning through:

- content of presentation (quality of ideas and ability to justify opinions); and
- performance elements (awareness of, and ability to engage, the audience).

Teachers made on-the-spot judgements of students' common task performances. Both Year 3 and Year 5 students completed two common tasks in speaking.

BEST WORK

Students' best work in speaking was assembled by teachers in three specified categories. Teachers were asked to base their assessments on two speaking performances/presentations:

- a reflective/discursive piece (a performance /presentation of a personal narrative, or a response to an issue eg, morning talk or debate); and either
- an imaginative piece (eg, a performance/presentation of a narrative, poem or play); or
- a piece from a subject area other than English (eg, a science report, an individual project, a report on a group activity in mathematics).

Figure 4. speaking assessment activities used in the National School English Literacy Survey

Reporting results in ways that encourage high achievement

Finally, the ways in which results from large-scale programs are reported also can have a significant impact on educational practice. In this section we summarise three approaches to reporting results from large-scale programs. All three of these approaches were used in the National School English Literacy Survey.

1. including reports against minimum standards

The first of these three approaches identifies a minimum standard that all students are expected to achieve by a particular stage of their schooling and reports performances against this minimum expectation. An example of such a standard would be the level of scientific literacy that all students are expected to reach by the end of compulsory schooling. In an international achievement study, results might be reported as the percentage of students in each country reaching this standard.

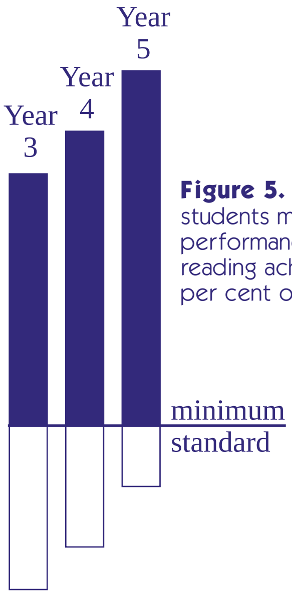


Figure 5. percentage of students meeting a performance standard in reading achieved by 85 per cent of Year 5 students

Minimum performance standards are established through a process known as ‘standard setting’. This process, which depends on experts making judgements about a minimally acceptable level of performance, is often controversial in practice. But the attempt to identify minimum standards that all students should be expected to reach by the end of particular years of school—and that are essential for success in future learning—seems to us to be eminently sensible. The achievement of minimum standards is a stepping stone to high achievement. Teachers need to be aware of the level of achievement that all students must reach if they are not to fall further behind in their learning. Rigorous assessment of student progress against this expectation, and diagnostic assessments to identify the reasons for individuals’ slow progress, should be a central part of the professional work of every teacher. And it may be important for teachers to spend more time and effort working with low-achieving students to ensure that they do not complete the year without meeting this minimum standard.

Although the achievement of minimum standards can present a significant challenge for some students and some schools, for other students, the demonstration of minimum competence does not pose a major challenge. In fact, many students achieve the minimum standard for a Year level much earlier in their schooling. For example, a performance standard in reading or writing achieved by 85 per cent of all Australian Year 5 students, also is achieved by more than 70 per cent of Year 4 students and by 60 per cent of Year 3 students (Figure 5).

The challenge posed by the specification of minimum standards is the challenge of better understanding the difficulties being experienced by low-achieving students and of identifying effective strategies for lifting the achievements of these students to at least the standard expected of students in that grade. In the National School English Literacy Survey we set a minimum performance standard using as a guide drafts of the Year 3 and Year 5 reading and writing 'benchmarks' and investigated various characteristics of students performing below that standard. The results of some of our analyses are shown in Figure 6. From Figure 6 it can be seen that, while almost 90 per cent of Year 3 and Year 5 students in the highest socioeconomic groups (children of parents in professional and managerial occupations) achieved our minimum standards for those Year levels, only 62 per cent of students in the lowest socioeconomic category met the Year 3 standard, and by Year 5 the percentage meeting the standard had declined to 47 per cent. In the separate Special Indigenous Sample (largely Indigenous students in rural and remote schools), only about 20 per cent of students met the Year 3 and Year 5 standards.

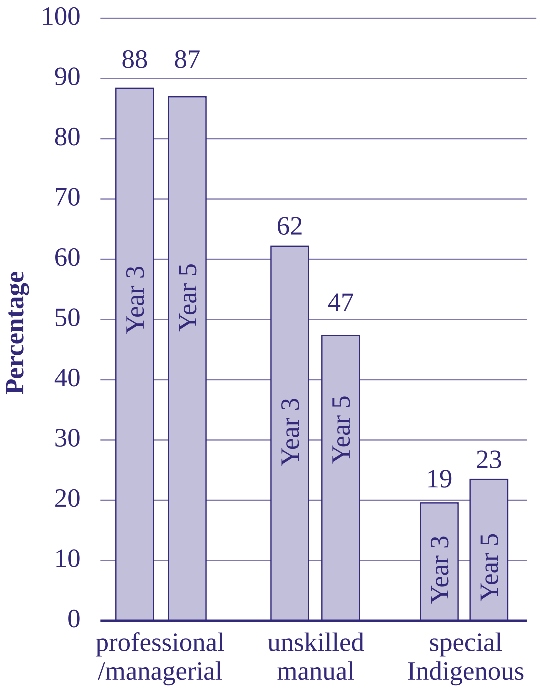


Figure 6. percentage of students meeting minimum performance standards in reading in the National School English Literacy Survey

2. reporting average achievement measures

A second method of reporting students' performances in large-scale assessment programs is to show the entire distribution of student results and/or statistics (such as the mean, median, standard deviation, or key percentile points) summarising this distribution. This method is the most common approach to reporting the results of large-scale programs. In international achievement studies and programs such as the US National Assessment of Educational Progress, it is usual to report and compare average levels of achievement nationally and cross-nationally. An advantage of this method is that it uses all students' estimated levels of achievement in the comparison of student performances.

Table 1 provides an example of the use of this method. In Table 1, the average mathematics results of students in different Australian States and Territories are shown and compared (Lokan *et al*, 1996). Because of differences in the average ages of students in the various States and Territories, the mean age of the students tested is shown. The table also allows differences in the means to be interpreted in terms of their statistical significance.

Table 1.

mathematics achievement by state (TIMSS)

State	Best estimate of mean score	Best estimate of average age	WA	ACT	SA	QLD	NSW	VIC	TAS	NT
WA	546 ± 8	14.0 ± .01		❇	❇	❇	▲	▲	▲	▲
ACT	544 ± 10	13.6 ± .02	❇		❇	❇	▲	▲	▲	▲
SA	536 ± 6	14.3 ± .02	❇	❇		❇	▲	▲	▲	▲
QLD	529 ± 8	14.0 ± .02	❇	❇	❇		❇	▲	▲	▲
NSW	509 ± 8	13.5 ± .02	▼	▼	▼	❇		❇	❇	❇
VIC	492 ± 6	13.5 ± .02	▼	▼	▼	▼	❇		❇	❇
TAS	488 ± 11	13.5 ± .04	▼	▼	▼	▼	❇	❇		❇
NT	487 ± 14	14.0 ± .20	▼	▼	▼	▼	❇	❇	❇	

❇ No statistically significant difference from comparison state
▲ Mean achievement significantly higher than comparison state
▼ Mean achievement significantly lower than comparison state

A difference between this report and reports based on the percentage of students achieving a minimum standard is that greater recognition is given to high achievement. When results are reported only as the percentage of students achieving a minimum standard, less importance is attached to the question of how far above the minimum standard some students perform (as long as they perform above the minimum, they contribute to the percentage). In contrast, when student groups are compared on the basis of their mean performance, the better the performance of students above the minimum standard, the higher the group mean.

3. reporting against a set of performance standards

A third approach to reporting student results is to define a set of performance standards and to report the percentage of students meeting each of these standards. Assessment programs using this approach usually describe and illustrate performance at each of the defined standards.

An example of this approach is the decision of the US Congress to report results of the National Assessment of Educational Progress (NAEP) not only in terms of student score distributions and means, but also in terms of three defined achievement ‘standards’: basic, proficient and advanced:

> *The congressional legislation that established the state NAEP program also mandated standards-based reporting of NAEP results; it stated that NAEP results should be presented both as overall scores and in terms of percentages of students who meet established standards for performance. Thus, in the 1990s, most NAEP assessments have reported summary scores and the percentages of students performing at or above basic, proficient, and advanced levels of performance.* *(Pellegrino et al, 1998)*

(Notice that, if the ‘basic’ level of performance represents a minimally acceptable standard, this third method of reporting actually incorporates the first.)

In the National School English Literacy Survey, we reported students' literacy achievements in terms of the levels of the national English profile. In effect, Levels 1 to 5 of the profile functioned as described and illustrated performance 'standards', and we reported the percentage of Year 3 and Year 5 students meeting each of these standards (Table 2).

Table 2.
percentage of students working in each profile level in reading
(National School English Literacy Survey)

Standard	Year 3	Year 5
Level 5	*	12
Level 4	12	39
Level 3	42	28
Level 2	42	21
Level 1	4	*

* Year 3 students were assessed as 'at or above' Level 4;
Year 5 students were assessed as 'at or below' Level 2.

An advantage of this method of reporting over simple numerical reports (eg, reports of national means) is that, by describing and illustrating each performance standard, this method makes explicit the nature of high achievement (the kinds of knowledge, skills and understandings required to perform at that level). When all achievement levels are considered together, they help to clarify the nature of development within the area of learning under consideration.

A disadvantage of this method is that the percentage of students achieving a particular performance standard is in general a less reliable statistic than the group mean for monitoring trends in performance over time.

Applying the Principles to Design an Assessment System

In this paper we have argued that large-scale assessment programs have an important role to play in informing educational decision making, but that when programs of this kind are used to report and compare publicly the performances of schools, school systems or nations, they also send influential messages about the kinds of learning valued by education authorities. For this reason, it is important that assessment programs reflect and reinforce curriculum priorities.

In practice, this observation means ensuring that assessment programs do not place undue emphasis on a limited range of easily-measured skills at the expense of students' abilities to apply their learning, to reflect on and think about what they are learning, and to engage in higher-order activities such as critical analysis and problem solving.

The goal of addressing and reinforcing valued learning outcomes may require the use of assessment methods other than paper and pen tests. The collection of useful information about some learning outcomes may require direct observations and judgements of students' work and performances, or perhaps classroom evidence assembled over a period of time through day-to-day work (eg, portfolios of student writing).

Central to our argument has been our belief that the primary goal of an assessment system should be to establish where all students are in their learning. In practice, this means ensuring that assessment exercises are designed to challenge and provide useful information about the achievements of all students. An assessment system should provide opportunities for lower-achieving students to demonstrate what they are able to do and, ideally, will provide these students with a degree of success and a sense of accomplishment. At the same time, it is important that the highest achieving students are engaged and challenged and that reliable information is provided about their levels of attainment and progress.

The goal of providing reliable information about students with very different levels of achievement may require assessment procedures rather different from fixed-length tests consisting of items scored right or wrong. These procedures may involve greater use of open-ended tasks that allow students to respond at a variety of levels, or tests that do not require all students to attempt exactly the same items (eg, tailored tests in which students take items of different difficulties).

Whatever the assessment methods and procedures they use, large-scale programs should begin with a recognition that individuals are likely to be at very different levels of attainment in an area of learning. When levels along an achievement continuum are described in terms of the kinds of knowledge, skills and understandings typical of students at those levels, and are illustrated with tasks and samples of student work, these levels provide a framework against which all students' achievements can be mapped and reported.

Once such a framework is constructed, it is possible to identify on this framework a minimum level of achievement that all students are expected to reach by a particular point in their schooling and to report whether or not students have met this standard. This approach is likely to be useful for identifying students in need of special assistance, and for whom further diagnostic information may be required.

However, the history of minimum competency testing tells us that setting a minimum performance standard and reporting the results of assessments primarily or solely as the percentage of students achieving that standard can have adverse implications for teaching and learning. This is particularly true when the stakes are raised by reporting and comparing publicly the percentages of students in different schools or school systems achieving the minimum standard. Under these conditions, two trends sometimes are observed.

First, an increasing number of test items are written at the level of the minimum performance standard. This is an entirely appropriate response: statistically, the way to improve the reliability of the decision as to whether or not a student has met the minimum standard is to increase the number of test items in the vicinity of the standard. However, because most tests are of fixed length, an increase in the number of items near the minimum standard usually means a decrease in the number of items well above the standard.

Second, in high-stakes contexts, teaching and learning can be focused on ensuring that low-achieving students are brought up to the level of the minimum standard. This is a highly desirable outcome for low-achieving students. The implications for students already performing well above the minimum may be less desirable if they are not also challenged and extended by classroom teaching and by the assessments themselves.

A way of avoiding these unintended consequences may be to ensure that public comparisons of schools and school systems are not based solely on the percentage of students achieving a minimum standard. Provided that assessment programs continue to be designed to allow more than pass/fail decisions in relation to a minimum standard, a straightforward alternative would be to also report and compare the mean scores achieved in different schools and/or school systems. Such a decision not only would encourage the design of assessment systems to provide reliable estimates of all students' levels of attainment, but also would recognise and encourage high achievement.

If large-scale programs are to provide useful input to educational decision making *and* reflect current teaching and learning priorities, then it will be important that both intended and unintended consequences of these programs are continually monitored and critically evaluated to ensure that they are operating in the best interests of all students' learning.

References

Applebee, AN, Langer, JA and Mullis, IVS (1987). *Learning to be Literate in America: Reading, Writing and Reasoning.* Princeton NJ: Educational Testing Service.

Dietel, RJ, Herman, JL, and Knuth, RA (1991). *What does research say about assessment?* Oak Brook: North Central Regional Educational Laboratory.

Frederiksen, N (1994). *The Influence of Minimum Competency Tests on Teaching and Learning.* Princeton: Educational Testing Service.

Griffin, BW and Heidorn, MH (1996). An examination of the relationship between minimum competency test performance and dropping out of high school. *Educational Evaluation and Policy Analysis*, 18, 243–52.

Kirsch, IS and Murray, TS (1998). *Adult Literacy in OECD Countries: Technical Report on the First International Adult Literacy Survey.* Washington, DC: US Department of Education, p13.

Lokan, J, Ford, P and Greenwood, L (1996). *Maths and Science on the Line: Australian Junior Secondary Students' Performance in the Third International Mathematics and Science Study.* Camberwell: ACER.

Marion, SF and Sheinker, A (1999). *Issues and consequences for State-level minimum competency testing programs.* National Centre on Educational Outcomes, January 1999.

Masters, GN and Forster, M (1997). *Mapping Literacy Achievement: Results of the 1996 National School English Literacy Survey.* Canberra: Department of Employment, Education, Training and Youth Affairs.

Mehrens, WA (1998). Consequences of assessment: what is the evidence? *Education Policy Analysis Archives*, 6(13).

Organisation for Economic Cooperation and Development (1999a). *Measuring Student Knowledge and Skills: A New Framework for Assessment.* Paris: OECD.

Organisation for Economic Cooperation and Development (1999b). *International Life Skills Survey.* Paris: OECD.

Pellegrino, JW, Jones, LR, and Mitchell, KJ (1998). *Grading the Nation's Report Card.* Washington, DC: National Academy Press.