

# Assessment Fit for Purpose

## Keynote Opening Presentation

Assessment in an Era of Rapid Change:  
Innovations and Best Practices  
32<sup>nd</sup> Annual Conference  
International Association for Educational Assessment

Barry McGaw

Managing Director  
McGaw Group Pty Ltd  
[barry.mcgaw@mcgawgroup.org](mailto:barry.mcgaw@mcgawgroup.org)

Director  
Melbourne Education Research Institute  
The University of Melbourne  
[bmcgaw@unimelb.edu.au](mailto:bmcgaw@unimelb.edu.au)

Assessment is a powerful educational tool. It influences the judgements of students and teachers about what is of most importance in the curriculum. The effects can be positive and can justify expectations of assessment-led reform. The effects can also be negative, for students, teachers, schools and the curriculum. It is important, therefore, for those responsible for assessment, particularly high-stakes assessment, to pay attention to the 'consequential validity' of their assessment systems. Assessment cannot be seen only as a technical task.

Assessment needs to be fit for purpose. The well-established distinction between formative and summative assessment is helpful in clarifying purpose and informing choice about method. The longstanding distinction between norm-referenced and criterion-referenced assessment is also helpful in clarifying purpose but has become less relevant for choice about method because of the capacity of modern psychometric methods to dissolve the methodological distinction between them.

## Potential power of assessment

Assessment is a powerful educational tool. It can provide information on student learning with which students can see their own progress. It can enable teachers to monitor the progress of individual students and also to obtain evidence about the effectiveness of their own teaching.

Assessment programs for whole education systems specify what the system takes to be important for students to learn. That specification can provide a very salient indication of where schools and teachers should direct major effort. In fact, if the system-level assessment is high-stakes, it can exert a powerful influence on what schools actually do choose to emphasise. If the assessment is appropriately focused, it can effectively drive reform.

Teachers often complain, however, that system-level assessment, far from driving positive reform, can lead to misdirected effort. They assert that limitations of techniques result in an assessment focus on what can be measured rather than what is important and that this, in turn, results in a focus in teaching on the outcomes that can be measured rather than those that are important. The critics do not deny that some important things can be, and are, measured but they concentrate on important objectives that cannot so readily be measured and typically are not. Their concern includes cognitive outcomes but extends to non-cognitive ones as well, such as capacity to work with others and a range of attitudes.

Critics of assessment systems also point to the long-term nature of educational objectives and claim that attention to these will also be jeopardised if assessment practices result in a concentration on short-term effects of schooling.

Whether the assessment is high or low-stakes does, of course, make a difference. Assessment can be high-stakes for either students or teachers and schools or for all of them. High-stakes outcomes for students are those on which important decisions depend, such as for access to particular streams of education in selective systems, to a higher level of education where progression is restricted, or to various positions in the labour market. High-stakes assessments for teachers and schools are those that can lead to formal judgements of their quality, based on the achievements of their students, particularly if the achievement data are made public. Even where access to the data is limited, however, its use can be high-stakes for the teacher since supervisors will typically have access and use the data to form judgements of the teacher's professional competence.

When the stakes are high, there will be the greatest risk that excessive attention will be given to those aspects of the curriculum that are assessed. That makes it all the more important that the assessment induces focus on what is important. It is not helpful simply to exhort people to pay attention to things to which the system gives little attention. Incentives influence behaviour so systems must align their incentives with the behaviours they desire. If only success is rewarded, risk-taking is likely to be suppressed. If innovation is desirable, then ways must be found for identifying and rewarding it.

Any debate about whether assessments are focused on the right things is essentially a debate about the validity of the assessments.

There are several well-known and widely accepted notions of validity. The first adopted are now typically described as content validity and criterion-related validity. Content validity is, as the name implies, about whether an assessment covers adequately the

domain it is intended to cover. It includes face validity and curricular validity, the second being of great importance for the reasons just discussed.

Criterion-related validity is concerned with the relationship between the assessments and those of some other measures to which it is expected or intended they be related. In an educational setting, the most obvious are likely to be subsequent measures, such as performance in a program for which the first assessments served as a selection measure. If end-of-secondary school assessments are used for university selection, then the strength of their relationship with university performance measures is an important index of their criterion-related validity. (The attenuating effects on correlation of restriction in the range of scores on the selection measure among those selected needs to be considered in this case.) Concurrent criteria can also be relevant. If assessments are limited in form, for example, to paper and pencil measures collected in a limited period of time, it can be helpful to know how strong is their relationship with a fuller set of assessments for which they might serve as a proxy. If their criterion-related validity can be established in these terms, then we can have more confidence in the more limited, formal assessment.

Cronbach and Meehl (1954) introduced the notion of construct validity which shifts attention to the underlying theoretical construct that the assessment purports to measure. Construct validity is typically evaluated using evidence that the assessment produces similar results to those of other assessments of the same construct (convergent construct validity) and different results from those of assessments of unrelated constructs (discriminant construct validity). Assessments in reading and mathematics would be expected to produce different results on the grounds that they are measuring different underlying constructs. If they do not, it could, for example, be because the items in the mathematics assessment are heavily verbal and so measure reading as well as mathematics. Alternatively, it could be that both assess general ability rather than curriculum-related reading and mathematics competence. In the former case, the mathematics assessment would have low construct validity. In the latter case, both would have low construct validity.

Messick (1989) added the concept of consequential validity which he defined as an evaluation of "the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice". At first sight, that is a much tougher validity criterion than the others. For those of you responsible for large-scale, high-stakes assessment programs, it implies that the validity of your assessments depends on their not being used in any ways that are biased, unfair or unjust. Messick however, adds that "it is not that adverse social consequences of test use render the use invalid but, rather, that adverse social consequences should not be attributable to any source of test invalidity such as construct-irrelevant variance." The latest edition of the *Standards for educational and psychological testing* (AERA, APA & NCME, 1999) similarly distances test developers from responsibility for uses of their assessments that are beyond the validity domain of the assessments.

That does not get you off the hook entirely, however, since much of the criticism of the consequences of your assessments is precisely in the domain of their validity. If tests designed to measure key learning in schools ignore some key areas because they are harder to measure and attention to those areas by teachers and schools is then reduced, then those responsible for the tests bear some responsibility for that.

It should be added that those who develop assessments cannot be held responsible for unreasonable claims made about the efficacy of their assessments. Test developers are often more circumspect than enthusiastic users.

It should also be added that test developers cannot be held responsible for failing to satisfy unreasonable expectations established by those whose intention is to use that failure to belittle the value of the assessments. There are two obvious examples.

One is the requirement that external assessments only tell teachers what they do not already know. External assessments are unlikely to tell teachers much that is new about differences in performance levels among students in their own classes. What they can add is information on how that class is performing in relation to others in similar schools elsewhere or in dissimilar schools.

A second is the requirement that assessment should become a mechanism for improvement. This point is often embellished with observations that, for example, just as weighing chickens will not make them fatter, so assessing children will not make them perform better. No one claims that assessment will itself improve things, only that assessment will inform judgements about how well something is working – a diet for the chickens or an educational program for the students.

### **Distinguishing purposes of assessment**

In addition to distinctions related to the properties of assessments, such as validity, there are important distinctions in the purposes of assessment. One is the distinction between summative and formative assessment. Another is the distinction in the point of comparison used in interpreting results, either a norm or a performance criterion.

#### *Summative vs formative assessment*

Summative assessment provides a summary of a student's achievements at some point at which it is relevant to take stock. This could be annual reporting to parents, based on local assessment by a teacher or school. It could be at particular stages identified by national or regional authorities as important. It could also be annual or less frequent, as in the Key Stages in England.

Summative assessment also occurs at key transition points where decisions depend on levels of student achievement at those points. The completion of secondary education is an obvious example of a point at which summative assessment provides information both for certification of completion of secondary education and for selection into a range of post-school destinations. The completion of tertiary education is another example but one at which the public appears more willing to accept assessments that are conducted only by institutions and not in some comparable way across institutions. One exception to that public tolerance of institution-based assessment occurred in Brazil during a period in which the Minister of Education introduced national examinations in a range of subjects for the completion of university degrees.

Formative assessment, on the other hand, is intended to identify learning needs and shape teaching. It can be frequent, either formal or informal and should lead to informed discussion between student and teacher about what the student should next do. Black and Wiliam's (1998) review of the use of formative assessment showed that it has a powerful impact on student performance. They report that experiments comparing systematic use of formative assessment with normal classroom practice produce an effect size of between .40 and .70 in favour of the use of formative assessment. That is, the mean performance of students receiving formative assessment is between 0.4 and 0.7 standard deviations higher than the mean performance of those not receiving this treatment. As Black and Wiliam note, this is a larger effect than is found with most educational interventions. Furthermore, their

review revealed that many studies show that the use of formative assessment helps low achievers the most.

Given the power of formative assessment to improve learning, it is important to consider what barriers there might be to its more widespread use. Recent OECD (2005) cases studies on formative assessment suggest that the barriers include tension between classroom-based formative assessment and high-visibility summative assessments and lack of connection between system, school and classroom approaches to assessment and evaluation.

The OECD (2005) work also identified successful strategies for achieving more widespread use of formative assessment. They include legislation giving priority to formative assessment, encouragement of the formative use of summative data, guidelines embedded in curriculum materials, provision of tools and exemplars for teachers, investment in initiatives incorporating formative assessment and investment in teacher professional development in the use of formative assessment.

The use of formative assessment can also be facilitated by aligning formative and summative assessment. Strategies for achieving this include ensuring summative assessments measure key skills on which development is expected to occur, convincing teachers that the use of formative assessment will lead to better summative assessment results, encouraging risk-taking by teachers as they explore better ways of assessing and teaching, broadening the basis for judging teachers to include, for example, students' capacity to judge their own progress (OECD, 2005).

There are some important lessons here for those of you whose role is in examination and testing agencies that are responsible for major summative assessment programs. There are things that you can do yourselves, as indicated in the OECD report, but there are also things that you could do in collaboration with others that could advance the productive use of formative assessment alongside your own programs.

#### *Norm-referenced vs criterion-referenced assessment*

On the issue of how to interpret the performances of individuals, a strong, initial norm-referenced tradition in educational (and psychological) assessment was an almost inevitable consequence of the origins of the work.

Some of the earliest work on the assessment of human skills lay in the domain of psychophysics (see, for example, Torgerson, 1958). In this work, scales of human judgement were constructed for phenomena for which external measures of the relevant physical property were also available, such as weight of objects, brightness of lights and pitch of notes. The human judges were typically not required to make absolute judgements of a property but rather to compare pairs of instances and to judge which was the greater, that is heavier, brighter or higher in the examples above. The pairs would be brought closer and closer together until no difference could be detected. The smallest detectable difference, the so-called 'just-noticeable difference', was a key element in the statistical analyses that generated the scales of human judgement.

In this work, the interest lay in the nature of human judgement not in the scales themselves, since physical measures of the phenomena were available. In other work, where the interest lay in educational performance or psychological phenomena for which there are no parallel physical measures, there was no external frame of reference for calibrating a scale of the human characteristic. In these cases, the interpretive strategy used was to locate individuals in relation to one another. The

results were expressed using comparisons to the mean performance (or norm), given as the distance of an individual from the norm in units such as standard deviations (z-scores, or stanines), or as a location in the distribution of scores, expressed in percentiles or quartiles.

The fundamental deficiency of norm-referenced assessment from an educational perspective is that it cannot readily measure growth or improvement in an individual. So long as the point of reference is the performance of others, the only way in which an individual can be seen to improve is relative to the performance of others, which means effectively at the expense of others. In the extreme case of all individuals improving at the same rate, none would be judged to have improved by norm-referenced assessment since none would have improved relative to the others. That deficiency was well understood but there was seen to be no alternative.

There was, in fact, one early exception. To measure attitudes, Thurstone first had judges rate items in terms of the intensity of the attitudes they expressed and used these ratings to scale, or 'calibrate', the items. To measure attitudes, he presented the items to individuals located them on the scale by the proximity of their attitude to particular items already located on the scale.

Glaser (1963) introduced the idea of using explicit performance criteria as the basis for judging the performance of individuals. With this criterion-referenced approach, the learning requirements are specified and the performances of individuals are judged against them rather than against the performances of other individuals.

In the initial phases, numerous specific criteria were nominated and the notion of a scale of performance was rather lost. The summary measure of performance was typically the percent of criteria satisfied, with no obvious way to take account of some criteria being more difficult to satisfy than others.

The development of new psychometric models resolved that problem and, more significantly, reduced the distinction between norm-referenced and criterion-referenced assessment to one of purpose and not one of psychometric method. As in Thurstone's approach, the new methods involve the separate calibration of a scale and its use in the measurement of individuals, but they permit simultaneous calibration and measurement. Variations in the difficulties of tasks are reflected in the differences in their location on the scale. Variations in the performances of individuals are reflected in the differences of their location on the same scale and that permits interpretation of individual performances in terms of the tasks. This is illustrated in Figure 1.

(Insert Figure 1 about here.)

The figure illustrates how the performances of students at various levels can be interpreted in terms of tasks on which they are most likely to perform successfully, those that will be close to their limit of successful performance and those that are most likely to be too difficult for them.

Calibration of tasks is further illustrated in Figure 2 which shows three particular mathematics questions from the tests in the Programme for International Student Assessment (PISA) 2003 (OECD, 2004, p.75). All three questions relate to the exchange of currency between Singapore dollars and South African rand. Knowing the number of rand per dollar, it is easier to convert dollars to rand (the question with a difficulty level of 406 on the scale and within band 1) than to convert rand to dollars (the question with a difficulty level of 439 on the scale and within band 2). A question requiring determination of whether a shift in exchange rates between the two occasions

is advantageous or disadvantageous for an individual traveller is considerably more difficult, with a level of 586 on the scale and in band 4.

(Insert Figure 2 about here.)

The three questions used in the illustration in Figure 2 are all simple right/wrong questions that would be marked 1 or 0. In examinations and many other educational tests, the questions are more complex and it would be appropriate to give some credit for answers that are partially correct. On a right/wrong question, the scale location for the item is the 0/1 boundary, usually defined as the point at which a person performing at that level has a 0.50 probability of answering the item correctly. For an item scored on a five point scale (0 to 4), the 0/1, 1/2, 2/3 and 3/4 boundaries. The use of this approach is illustrated with selected mathematics tasks from the PISA 2003 tests in Figure 3. In this figure, both right/wrong and partial-credit tasks are mapped onto the performance scale.

(Insert Figure 3 about here.)

Among the most difficult items (in band 6 at the top of the scale) is one which is a right/wrong question: question 1 which deals with a carpenter which is scaled at 687. On a question about walking (length of pace and speed) a response sufficient to gain the full available marks of three is very difficult to achieve, being scaled at 723 towards the top of band 6. A response sufficient to gain a score of 2 rather than 1 is also quite difficult to achieve, being scaled at 666 and at the top of band 5. Even achieving 1 rather than 0 on this question is relatively difficult, being scaled at 605 and at the top of band 4. On an item dealing with a skateboard, there were two right/wrong items (question 13 at 570 and question 14 at 554) that were more difficult to answer correctly than it was to obtain the full two marks available on question 12 (496).

Measuring individuals with scales constructed by calibrating tasks by difficulty enables performance to be interpreted not only with respect to particular tasks but to a scale of increasing task difficulty. Repeated use of such scales also permits improvements in the performance of individuals over time to be registered as successively higher points on scale.

Normative comparisons among individuals can also be made using the results when that is necessary, as in selection of some number of high performers for particular opportunities or the selection of some number of low performers for supplementary or remedial instruction.

### **Using criterion-referencing in public examinations**

An important question is whether the advances in psychometrics that permit calibration of scales and measurement of individuals that allows interpretation of performance in terms of the scales can be applied in public examinations.

These examinations do need to provide normative information, particularly at the end of secondary education where selection for university courses requires the identification of the relevant number of the best performing students required to fill the available places. The examinations also need to provide criterion-referenced information in certifying the level of performance of a student. At least it is presumed that this is the kind of information provided by the certification. The endless public debates in some jurisdictions about the comparability of grades over time (whether an 'A' now is what an 'A' was then) reflect a desire for the criterion to remain constant not the percentage of candidates receiving an award of A, B, etc.

Although public examination results are often used only in a normative fashion, there is certainly criterion-referenced information available in them. Standards of learning expected of students (criteria) are typically well-specified in the curricula and in turn provide the basis on which the examinations are built. The typically careful processes of examination setting and review provide a good example of how a strong link is built between the curriculum and the examination. Students' responses to the examination questions then provide information on student performance in relation to the standards or criteria.

What happens in many public examination systems is that the criterion information is ignored and the results are used only normatively to rank students for the award of grades and for selection into post-examination study options. It would be much better if both norm and criterion-referenced uses of the results could be supported.

The procedures used by some examination authorities attempt this marrying of criterion and norm-referenced assessment. Criteria are defined for some of the grade boundaries and marked student scripts at those grade boundaries in previous years are inspected in an effort to ensure that the definition and application of the criteria are consistent over years. Normative information is also introduced with an analysis of the distributions of grades that would be awarded by locating the grade boundaries at different marks in a range around the initially proposed cut and a comparison with the distributions in previous years. If the criteria-based judgements would result in a significantly different distribution of grades from those of previous years, evidence of any change in the student cohort taking the course is sought to see if a marked shift in grade distributions would be justified. In the end, the distribution actually awarded, and the grade boundaries that create it, are set on the basis of both criterion and normative considerations.

In the state of New South Wales in Australia a stronger attempt is made with the end-of-secondary school Higher School Certificate to maintain consistent criteria over time and then to report explicitly on performance in relation to the criteria while also providing normative information. The move to this form of measurement and reporting on a well-defined scale was recommended by McGaw (1997) and the detailed strategy for achieving it was developed within the New South Wales Board of Studies by Bennett (2001).

To develop grade descriptors, Bennett used past examinations. Experienced examiners for each subject reviewed examination papers and students' marked scripts to develop descriptions of students' performance for Bands 6 to 2. A low Band 1, representing inadequate (failing) performance, was not described.

Use of these band descriptors in subsequent years involves several stages. First, examiners independently form an 'image of each band' and then, marking an initial sample of scripts, set cut marks for each band boundary on each question. Secondly, the examiners work together to reach agreement on boundary locations for bands on each question. Boundary locations for total scores also established. Thirdly, student work at the boundaries on total scores are inspected and cut points reviewed and finally determined. The band boundaries are then located on a 0-100 mark scale that is used for reporting results. The 5/6 boundary set to 90, the 4/5 boundary to 80 and so on down to the 1/2 boundary which is set to 50 making 50 essentially the pass mark below which performance is declared to be inadequate.

A student receives a Course Report for each subject taken in the form shown in Figure 4. The student receives an overall mark as well as separate components from the examination and school-based assessment (called the 'assessment mark' in New



South Wales). There are descriptions of the performance bands to permit a criterion-referenced interpretation of the student's overall mark. Descriptions of bands above the one in which the student is located report what the student does not know and is not able to do. Descriptions of the bands below the band in which the student is located report what the student does know and can do. The description of the band in which the student is located reports the student's current level of development. The student will satisfy some of the criteria in this band, with the extent indicated by the position of the student's overall mark in the mark range for the band.

(Insert Figure 4 about here.)

Students' results can be described in terms of the band they have reached in the same way that grades of A, B, C, etc indicate the performance band in which the student falls. In New South Wales, as Figure 4 makes clear, the underlying mark that determines the band location is also published and not suppressed as it is in the English A levels.

Higher School Certificate students in New South Wales also receive a summary Record of Achievement of the type shown in Figure 5. For each course (subject) studied, this document gives the student's (school-based) assessment mark, the examination mark, the (overall) HSC (Higher School Certificate) mark and the performance band in which it lies. There is also a listing of preliminary courses taken (typically in the penultimate year of secondary education).

(Insert Figure 5 about here.)

The New South Wales Higher School Certificate results provide both norm-referenced and criterion-referenced information, with the latter being reported on well-defined scales that are consistent over time, at least in the short-run.

### **Measuring status or change**

In all of the measures discussed so far, formative or summative, norm-referenced or criterion-referenced, the purpose is to determine the current status of individuals. In many settings we are more interested in change in status than in current status so the question becomes how best to measure change.

It turns out that this is not as simple as it seems. First there is the question of what kind of change to measure. One could use the simple difference between measures of status at two different times but that would be likely to give an advantage to those whose initial status is higher. If they have already achieved more by the start of the period, then that indicates that they are on a higher growth trajectory than others and could be expected to grow more during the period than others starting at a lower level. An alternative approach would be to use as the measure of growth that part of the final status that could not be predicted from the initial status – effectively a residual measure of growth. That would remove the advantage of higher initial status but it may unfairly disadvantage those who have already done well in their earlier education prior to the period under review. The residual measure of growth has the further disadvantage that it is normative, since 'growing more or less than expected' involves a comparison with how others have grown.

Both the absolute and residual measures of change have a further problem of relatively low reliability. Their reliability is lower than the reliabilities of the two status measures from which they are calculated.

Without a measure of prior status, an absolute measure of change cannot be calculated. In some cases, 'residual' measures of change are derived using some proxy for prior status. Measures of social background are used in some cases. In Victoria in Australia at the end of secondary education, a General Achievement Test, given during the final months of Grade 12, is taken to be a proxy for what students were like at the beginning of Grade 11 and used to estimate residual growth. In this case, there is a further problem. The General Achievement Test is also used to identify schools for which internal assessment results are 'out-of-line'. The results are not adjusted on the basis of scores on the General Achievement Test but the inconsistency triggers a review of the internal assessments by external reviewers and may result in adjustments to them. So, in this case, a lack of alignment of the internal assessment component and the General Achievement Test can trigger 'corrections' to the internal assessments before they are combined without further adjustment with external examination results to produce overall results. At that point, a lack of alignment between the General Achievement Test and the overall final results in Grade 12 is used to compute 'residual' gain. The trick for schools then is to keep the lack of alignment 'under the bar' in order to avoid adjustment in the first phase and so to preserve it for recognition as gain in the second phase.

There is considerable interest currently in measures of gain as a means of understanding the value that schools add. It is a key element in the implementation of the *No Child Left Behind Act* in the US. The French Ministry of Education publishes the examination results for each school, but also a result predicted on the basis of the school's student intake, and the difference between two as an estimate of what the school has added (<http://indicateurs.education.gouv.fr/brochure.html>). The UK Department for Education and Skills publishes similar school performance tables on its website ([www.dfes.gov.uk/performancetables](http://www.dfes.gov.uk/performancetables)). These tables give the percentage of students in each school achieving at or above particular levels in English, Mathematics and Science. For Key Stage 3, there is also an estimate of the value that schools have added, given the point their students had reached in Key Stage 2.

A helpful review of various techniques for estimating 'value-added' is provided by Braun (2005).

### **Back to the consequences**

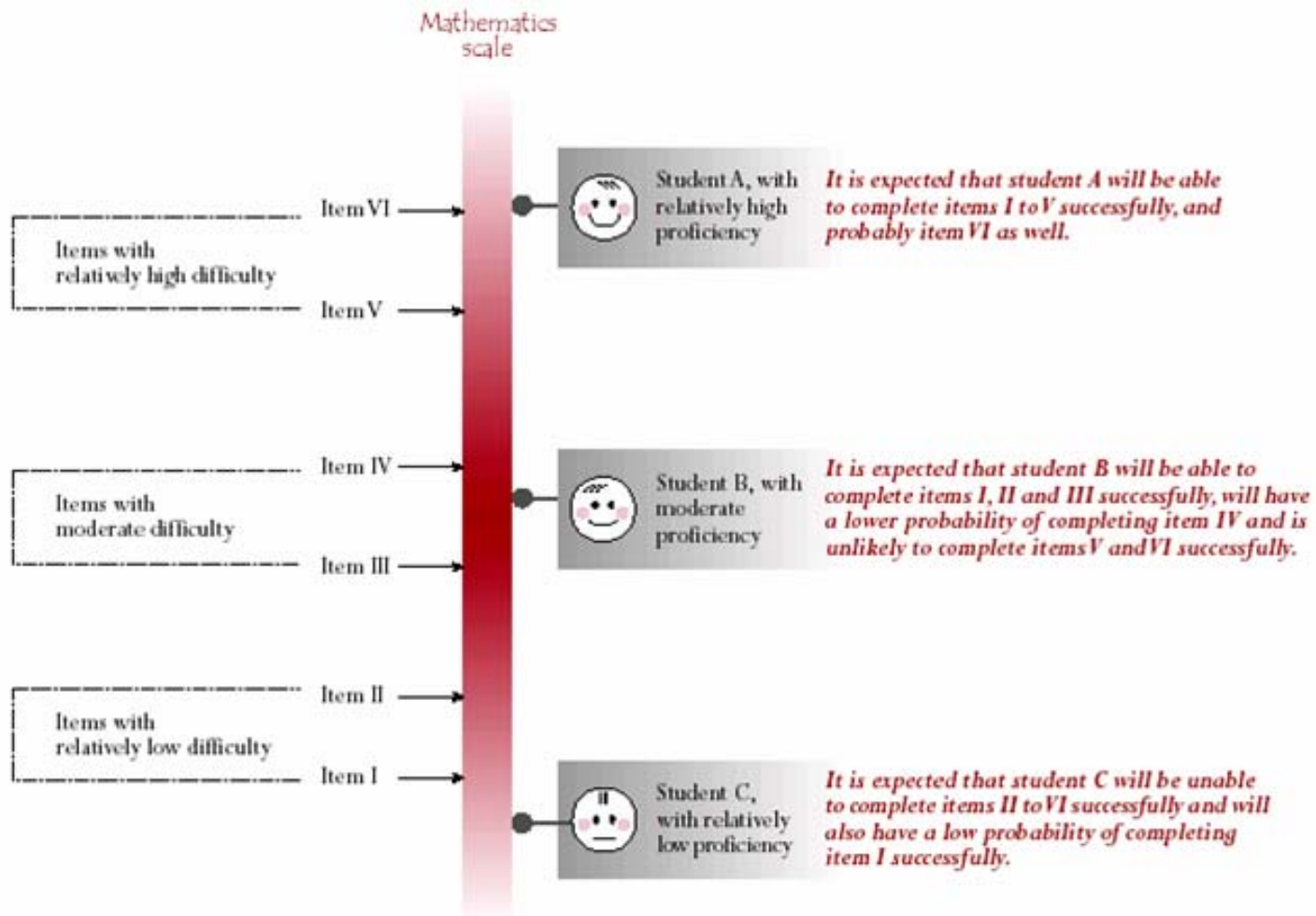
There are clear benefits of good assessment. It makes the goals of teaching and learning clear to learners; it makes improvement clear and it teaches learners how to monitor their own learning. The ability to monitor one's own learning is a key meta-cognitive capacity and one that helps to build the base for effective lifelong learning.

There are clear risks of high-stakes assessment and examination programs diverting teaching and learning from goals defined in the curriculum, and understood by the public and the profession, as being important. There are also risks that high-stakes external assessment will reduce the likelihood of productive use being made of formative assessment.

The responsibility for maximising the benefits and minimising the risks must be shared but those in examination and assessment roles have a special responsibility to play their part.

## References

- American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME). (1999). *The standards for educational and psychological testing*. Washington, DC: Author.
- Bennett, J. (2001), Standards-setting and the NSW Higher School Certificate [www.boardofstudies.nsw.edu.au/manuals/pdf\\_doc/bennett.pdf](http://www.boardofstudies.nsw.edu.au/manuals/pdf_doc/bennett.pdf).
- Black, P. & William, D, (1998), Assessment and classroom learning. *Assessment in education: Principles, policy and practice*, **5**, 7-74.
- Braun, H. I. (2005). Using student progress to evaluate teachers: A primer on value-added models. Princeton, New Jersey: Educational Testing Service.
- Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, **52**, 281-302.
- Glaser, R. (1963), Instructional technology and the measurement of learning outcomes: some questions. *American Psychologist*, **18**, 519-521.
- McGaw, (1997), *Shaping their future: Recommendations for reform of the Higher School Certificate*. Sydney: Department of Training and Education Co-ordination.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.) *Educational measurement* (3<sup>rd</sup> ed). New York: Macmillan, pp.13-103.
- OECD (2004), *Learning for tomorrow's world: first results from PISA 2003*. Paris: Author.
- OECD (2005), *Formative assessment: improving learning in secondary classrooms*. Paris: Author.
- Torgerson, W.S. (1958), *Theory and methods of scaling*. New York: Wiley.



**Figure 1: Location of persons and tasks on same scale**

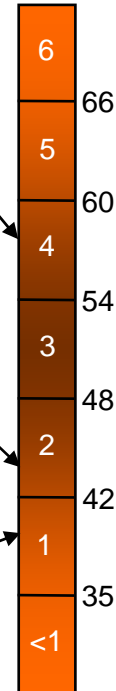
[Source: OECD (2004), *Learning for tomorrow's world: first results from PISA 2003*, p.48]

Mei-Ling from Singapore was preparing to go to South Africa for 3 months as an exchange student. She needed to change some Singapore dollars (SGD) into South African rand (ZAR).

During these 3 months the exchange rate had changed from 4.2 to 4.0 ZAR per SGD. Was it in Mei-Ling's favour that the exchange rate now was 4.0 ZAR instead of 4.2 ZAR, when she changed her South African rand back to Singapore dollars? Give an explanation to support your answer. [586]

On returning to Singapore after 3 months, Mei-Ling had 3 900 ZAR left. She changed this back to Singapore dollars, noting that the exchange rate had changed to: 1 SGD = 4.0 ZAR. How much money in Singapore dollars did Mei-Ling get? [439]

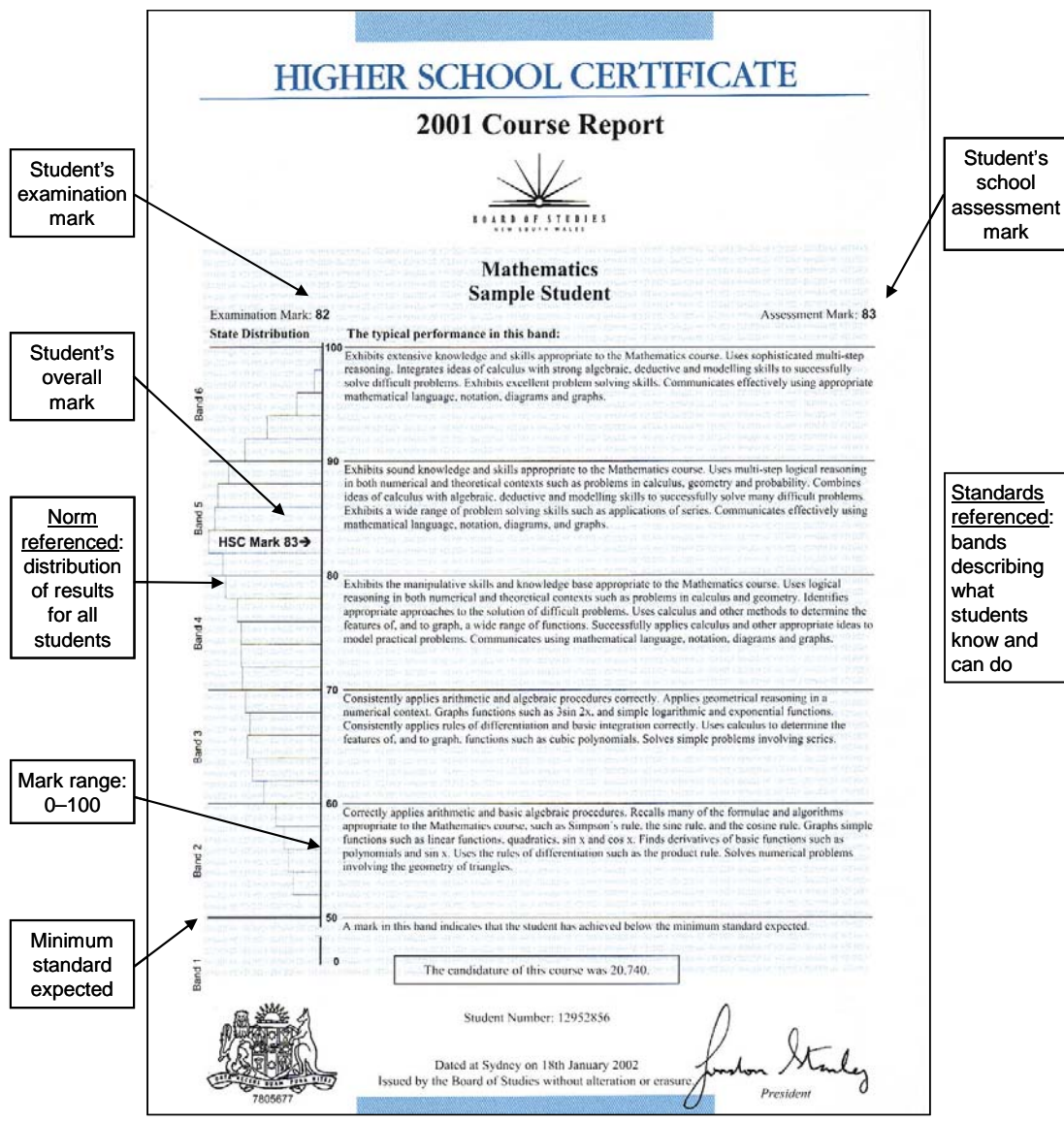
Mei-Ling found out that the exchange rate between Singapore dollars and South African rand was: 1 SGD = 4.2 ZAR. Mei-Ling changed 3000 Singapore dollars into South African rand at this exchange rate. How much money in South African rand did Mei-Ling get? [406]



**Figure 2: Scale values of sample mathematics questions in PISA 2003**  
 [Source: OECD (2004), *Learning for tomorrow's world: first results from PISA 2003*, p.75]

Level	Space and shape Figures 2.4a-c	Change and relationships Figures 2.7a-b	Quantity Figures 2.10a-b	Uncertainty Figures 2.13a-c
6 668.7	<b>CARPENTER</b> Question 1 (687)	<b>WALKING</b> Question 5 – Score 3 (723)		<b>ROBBERIES</b> Question 15 – Score 2 (694)
5 606.6		<b>WALKING</b> Question 5 – Score 2 (666)		<b>TEST SCORES</b> Question 6 (620)
4 544.4		<b>WALKING</b> Question 5 – Score 1 (605) <b>GROWING UP</b> Question 8 (574)	<b>EXCHANGE RATE</b> Question 11 (586) <b>SKATEBOARD</b> Question 13 (570) <b>SKATEBOARD</b> Question 14 (554)	<b>ROBBERIES</b> Question 15 – Score 1 (577) <b>EXPORTS</b> Question 18 (565)
3 482.4	<b>NUMBER CUBES</b> Question 3 (503)	<b>GROWING UP</b> Question 7 – Score 2 (525)	<b>SKATEBOARD</b> Question 12 – Score 2 (496)	<b>OECD average = 500</b>
2 420.4	<b>STAIRCASE</b> Question 2 (421)	<b>GROWING UP</b> Question 7 – Score 1 (420)	<b>SKATEBOARD</b> Question 12 – Score 1 (464) <b>EXCHANGE RATE</b> Question 10 (439)	<b>EXPORTS</b> Question 17 (427)
1 358.3			<b>EXCHANGE RATE</b> Question 9 (406)	
Below Level 1				

**Figure 3: Map of selected PISA 2003 mathematics tasks**  
 [Source: OECD (2004), *Learning for tomorrow's world: first results from PISA 2003*, p48]




**Figure 4: New South Wales Higher School Certificate student's subject report**



# HIGHER SCHOOL CERTIFICATE

## Record of Achievement




*This is to certify that Sample Student of Sample High School  
has received the results shown below:*

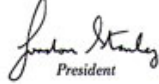
Year	Unit Value	Course	Assessment Mark	Examination Mark	HSC Mark	Performance Band
2001	2	English Standard	80	78	79	4
	2	Mathematics	90	92	91	6
	2	Chemistry	76	73	75	4
	2	Legal Studies	63	63	63	3
	2	PD, Health, Phys Ed	86	88	87	5
2000	2	English Standard (Preliminary)				
	2	Mathematics (Preliminary)				
	2	Chemistry (Preliminary)				
	2	Legal Studies (Preliminary)				
	2	PD, Health, Phys Ed (Preliminary)				

**ELIGIBLE FOR HIGHER SCHOOL CERTIFICATE**

Student Number: 65487965



*Dated at Sydney on December 2001  
Issued by the Board of Studies without alteration or erasure.*



Jonathan Stanley  
President

All HSC courses listed with:  
(School) Assessment Mark  
Examination Mark  
(Overall) HSC Mark  
Performance Band

All Preliminary courses listed

Figure 5: New South Wales Higher School Certificate Record of Achievement