

Best Practice in Assessment for Learning

John Izard
Adjunct Professor, RMIT University
john.izard@rmit.edu.au

Abstract

Learning involves changes in knowledge, skills and the sophistication of the strategies employed by the learners. To measure these changes we need at least two relevant measures. One assessment must document the level of achievement prior to a particular stage of learning and a later assessment must document a higher level of achievement. But improving achievement requires more than good assessment. To use a farming analogy: the farmer's maize will grow better if appropriate nutrients and water are provided in timely fashion. Measuring the height of the maize frequently is not going to improve the yield at all. To provide effective learning opportunities for all pupils we need assessment strategies that will be teacher-friendly (helpful in identifying what has to be taught) and teaching strategies to ensure students learn what they currently do not know. This paper looks at the requirements for using assessment for teaching and learning and addresses threats to the validity of assessments to evaluate progress. These threats include failure to use appropriate samples of tasks in assessments, use of uncalibrated tasks, and inappropriate scoring procedures and methods of interpreting data.

THE ASSESSMENT CONTEXT

Traditionally, examinations have been used to assess whether learning has occurred but weaknesses in methodology have meant that examinations often show a current state of knowledge, not whether the teaching was successful. Examinations have been recognised as a powerful influence on what happens in schools by tending to dictate what is taught and not taught. The examinations implicitly define what knowledge is valued and what is regarded as not important. In many educational systems much of the provision of evidence about learning is external to the classroom and the data obtained are not provided to schools and teachers in a form that can be used to help the students learn more. But frequently such examinations assess what is easy to measure rather than what is important to measure. Such assessments are of learning because they are not used to facilitate learning (assessment for learning). But improving achievement requires more than good assessment. To use a farming analogy: the farmer's maize will grow better if appropriate nutrients and water are provided in timely fashion. Measuring the height of the maize frequently is not going to improve the yield at all.

Learning involves changes in knowledge, skills and the sophistication of the strategies employed by the learners. Teacher instructional activities, whether using electronic means or not, are expected to achieve this learning – this is one of the important roles of teachers. Another important role of teachers and those who develop teaching strategies and materials is to provide *evidence of these changes* in the students. To measure these changes we need at least two valid (relevant) measures. One assessment must document the level of achievement prior to a particular stage of learning and a later assessment must document a higher level of achievement. Before we can show that progress has been achieved in a teaching program, we have to indicate the current achievement status of each pupil and the subsequent assessments have to include tasks representative of the skills we intended teaching. Black and Wiliam (1998a, 1998b) argue that facilitating student learning is part of the teacher's role. Without such informed learning based on formative assessment, improvement of standards becomes accidental rather than purposeful. Further, they have reported that studies of formative assessment show effect sizes (Cohen, 1969, 1988) between 0.4 and 0.7 on standardized tests, larger than most known educational interventions. Research in Australian school classrooms (Izard, 1998b; Izard *et. al.*, 1999) and subsequent research at tertiary undergraduate level in Britain (Izard, *et. al.*, 2003) has had similar results.

Use of student assessment for teaching purposes involves identifying where students have reached in their learning, what skills and knowledge are being established currently, what skills and knowledge are not yet within reach, and providing differential teaching according to their needs, based on analysis of the evidence. Many teachers and administrators accept and support analysis of test data to improve teaching and learning, but practical implementation has been found wanting. (Izard, 1998b; Black and Wiliam, 1998a, 1998b). Without valid student assessment practices the actual achievements are never compared in a legitimate way with the intentions. (Izard, 2002a)

SOME PROBLEMS WITH OUR ASSESSMENT STRATEGIES

If I asked you to measure the changes in the height of some plants you would probably reach for a measuring tape, metre-rule or some other measuring device. You would record the height initially, record the height at later times and look at the differences to show how much the plant has grown. This strategy is well known, and is based on considerable experience using measuring scales for length (whether the units are in inches, feet, yards and miles or millimetres, centimetres, metres, and kilometres). There are agreed standards defining the size of the units and community expectations about the accuracy of the measures.

Teachers assume that classroom tests measure what they are intended to measure and take little account of errors of measurement. Short true/false tests proliferate even though high scores can be achieved without seeing the items. Many teacher-made tests are conceived and written without a specification to ensure adequate sampling of topics. Sub-tests often have too few items to provide meaningful information. Test items constructed by teachers without technical support do not always distinguish between

students with relevant knowledge and students lacking such knowledge. Tests differ in number and format of items. Items vary in difficulty but are treated as equivalent in difficulty. Two tests of the same topic are assumed to cover the same work without empirical validation. If we use the analogy of a ruler to represent a test, teacher-made tests may be considered as measuring “tapes” or “rulers” of different lengths with different markings on the “ruler”. When teacher-made tests (and many published tests) are used to assess students in a classroom, there are no defined units for the “measuring tapes” or “rulers”. One “tape” may have large units while another may have small units and the relationship between these units on different “tapes” is unknown. The “tapes” will vary in length and the units will probably not be in equal intervals along the “tape”. Since there is no placing of the rulers together, a reading on one ruler has no meaning in relation to a reading on another ruler. Measuring progress is fraught with difficulty where different tests are used on two or more occasions.

If a test is easier then scores will tend to be high. If a test is more difficult then scores will tend to be low. If two tests are given at the same time to the same students, then it will be possible to see which items are easy and which are difficult. If two tests are given at differing times (without being given together to this group or another comparable group) any changes in the score cannot be interpreted. *One does not know whether the difference in scores was because the tests differed in difficulty, whether learning occurred over the time interval, or whether some combination of these events occurred.* Many teachers do not know the relative difficulties of the tests they give, so interpreting the results from their tests is impossible. A further difficulty with published tests is in the reporting of scores. Traditional published test data (expressed in percentiles or standard scores based on relative position of students) are not appropriate to measure achievement progress. Results are interpreted relative to a reference (norm) group (whether relevant or not). They compare students with students rather than compare each student’s achievements with the curriculum intentions. We do not learn from the data collected what students know or do not know, because this information is ignored in the interpretation of the evidence. (Izard, 2002a, Izard 2002c)

How is a teacher to indicate progress? Assessing progress in a curriculum-related way implies that subsequent assessments will be made, and that these assessments will be compared with the earlier records. These earlier records have to be in a format that permits legitimate comparisons. What tasks can a student now do that could not be done before? It is necessary to change the way in which test data are presented. Since familiarity with the test material may be a plausible explanation for any improvement in score rather than effective teaching, alternate forms of tests are required. (Izard, 2002a)

TECHNICAL LIMITATIONS OF ASSESSMENT

A Sampling of items across the desired range of skills

When test items are prepared to suit single grade minimum competence requirements, many students miss out on the opportunity to demonstrate their skills and knowledge. This inequitable situation is illustrated in Figure 1 (Figure 1 from Izard, 2002b). (The comments have been added to an actual analysis.)

better if the standard errors are taken into account. When gains are measured their gains would be zero regardless of their actual learning because of the deficiencies of the assessment strategy. The problem with floor effects is more subtle. Consider the bottom 96 students that scored 1 or 2 on the test. If this was their pre-test result then their post-test result may be better but the decision about gains is based on success on a single item or two items – hardly a convincing measure of attainment status. Izard (1998a, 2002b) reviews other constraints in giving candidates due credit for their work in his paper on strategies for quality control in assessment.

B Sampling of students across the desired range of items

For a test analysis to provide useful information, the data for the analysis must include the appropriate indicators in order to address the appropriate issues or aspects. For example, if wishing to compare scores obtained by female candidates with those obtained by male candidates one has to know which are male and which are female. Such an analysis is impossible unless the data for each candidate include this information. Similarly, an analysis of the contribution of each test item requires the data for each candidate to include each response to the items. *This information cannot be retrieved from total scores for candidates.*

If there are 4 items and 7 candidates the largest possible number of item-candidate pairs will be 28 (4 X 7). For this simplified example, the matrix as shown in Figure 2 has a bullet (•) for each interaction between an item and a candidate.

	Candidate						
Item	1	2	3	4	5	6	7
1	•	•	•	•	•	•	•
2	•	•	•	•	•	•	•
3	•	•	•	•	•	•	•
4	•	•	•	•	•	•	•

Figure 2 Item-candidate pairs for 4 items and 7 candidates

An analysis is possible even if some items are not attempted by all candidates. Often we can assume that items not attempted provide evidence that the work was not known. Or we may have to refer to such missed items as missing data where such an assumption is inappropriate. For example, if some examination papers omitted some pages, it would not be fair to penalise the candidate for errors made in the production of the examination papers. In the matrix shown in Figure 3 reasonably accurate estimates for those who received the faulty items (shown with a ?) are possible because the gaps in the evidence are limited.

Item	Candidate						
	1	2	3	4	5	6	7
1	•	•	•	?	•	•	•
2	•	•	•	•	•	?	•
3	•	?	•	•	•	•	•
4	•	•	?	•	•	•	•

Figure 3 Incomplete item-candidate pairs for 4 items and 7 candidates

There is another way that such a pattern could arise. If we consider only candidates 2, 3, 4 and 6 the pattern is that same as for an examination that offers students a choice of questions. An analysis is possible provided that there is an *adequate overlap* between items and candidates. By this we mean that there is sufficient evidence from those attempting the same items to gauge whether the items are comparable: we could say we require *connectedness*.

An analysis is possible if many items and candidates are not paired but if there is not some form of connectedness in the data the analyses may lead to ambiguous or misleading results. For example, the Figure 4 matrix will allow reasonably accurate estimates for subsets but, because the gaps in the evidence are substantial, the performance on items 1 and 2 with candidates 4 to 7 cannot be related to the performance of items 3 and 4 with candidates 1 to 3.

Item	Candidate						
	1	2	3	4	5	6	7
1				•	•	•	•
2				•	•	•	•
3	•	•	•				
4	•	•	•				

Figure 4 Unconnected item-candidate pairs for 4 items and 7 candidates

Items and candidates cannot be put on the same achievement continuum because of the limitations in the data collection design. If the items vary in difficulty and the test analysis takes no account of this, the practical consequence is that some candidates receive an unfair advantage over other candidates. It should be noted that the same

difficulty of interpretation occurs when optional questions are offered on an examination. If an examination provided four questions and told candidates to answer two, those candidates responding to questions 1 and 2 cannot be compared with those responding to questions 3 and 4. In practice, these means that those who choose the easier items receive a higher score than they deserve and those who choose the more difficult items fail to receive due credit for the quality of their work. This issue is taken up again later when I discuss implementation issues.

C Sampling of students relative to the population

The candidate sample from particular classrooms will not be a random sample of the whole population of students. The candidates are atypical in that they share experiences and teaching, and home background with their colleagues in those classrooms. For such a candidate group, intra-class correlation will be higher than expected with a simple random sample from a whole population of candidates. Cluster sample design data analysed under simple random sample assumptions may be expected to lead to greatly underestimated errors for means, correlation coefficients, standardized regressions coefficients and multiple correlation coefficients. Accordingly, traditional inferential statistics will need cautious interpretation due to design effects (Ross, 1978, 1993).

D Reporting differences between students

One way of reporting the effects of learning is to report how much overlap there is between the pre-test and the post-test results (see Cohen, 1977, pp. 20-27), or to report the difference between means in standard deviation units (obtained by taking the square root of the pooled variance – see Cohen, 1969, 1977, 1988). When the magnitudes of improvements are expressed as effect sizes in standard deviation units we can use Cohen's descriptors as a common language to describe these magnitudes.

Table 1 shows these descriptors together with ranges assigned by the writer for this paper. Effect sizes are described as "very small", "small", "medium" or "large". [also uses the idea of overlap of the distribution of scores of groups for illustrative purposes. For example, for two normal populations with equal variability and equally numerous, an effect size of 0 indicates 100% overlap or 0% non-overlap. An effect size of 0.2 indicates 14.7% non-overlap (the component of the combined distribution not shared by the two populations). The corresponding non-overlap values for effect sizes of 0.5 and 0.8 are 33% and 47.4%.]

Note that changes are not always positive. In some cases students fail to make progress, or obtain lower scores on the required skills after the "teaching" of those skills. Houston and Neill (2003) reported on their testing of tertiary undergraduate students in awareness of the steps in the modelling cycle using partial-credit multiple-choice items obtained from three universities but were unable to distinguish between changes in score due to differing test difficulty and changes due to learning or forgetting. Izard, Haines, Crouch, Houston and Neill, (2003) calibrated these three partial-credit tests of mathematical modelling administered at level in Britain and re-analysed the data.

Table 1 Descriptors for magnitudes of effect sizes (after Cohen, 1969, p.23) and assigned ranges

Effect Size Magnitude	Cohen's Descriptor and Cohen's Example	Assigned Range
< 0.2	Very small*	0.00 to 0.14
0.2	Small difference between the heights of 15 year old and 16 year old girls in the US	0.15 to 0.44
0.5	Medium ('large enough to be visible to the naked eye') difference between the heights of 14 year old and 18 year old girls	0.45 to 0.74
0.8	Large ('grossly perceptible and therefore large') difference between the heights of 13 year old and 18 year old girls or the difference in IQ between holders of the Ph.D. degree and 'typical college freshmen'	0.75 or more

* Note that "very small" is a descriptor devised by the author for magnitudes less than "small"

When test difficulty was controlled, there were differences in mean score ranging from -0.56 to 0.84 with corresponding effect sizes ranging from -1.28 (medium decrease in score) to +1.22 (large increase in score) implying that some students failed to retain this knowledge as they progressed through their course. (Houston and Neill, 2003, p. 175)

IMPLEMENTATION ISSUES

A Problems with assessments

Many assessments whether teacher-constructed or from a commercial test publisher are not representative of the curriculum intentions. This is unfair to the students and their parents. Teachers often test what they consider important from what they teach. If they omit significant sections of the curriculum then students are short-changed in the coverage of the curriculum. When such students are attempting external examinations, the teaching is directed often to what the teacher believes will be on the examination rather than to the curriculum that the examination assesses. If the teacher is successful in prediction, students will obtain a better mark than they deserve in terms of their coverage of the curriculum (and the teacher will attract more students). If the teacher is unsuccessful in prediction, the ground will probably shift – the student will be blamed for not working hard enough. Commercial tests prepared in Australia do not address the eight different government school system curricula. Where tests are prepared (and there are not many in evidence) they are prepared either for the most populous State or for some amalgam of the courses of the largest States.

In some cases tests vary in length. Since, other things being equal, a longer test gathers more evidence of achievement, a longer test will tend to be a more accurate assessment of the student's knowledge. When teachers compare scores over time, no account is taken of variation in the length of tests, or variations in the difficulty of the

items, or the associated errors of measurement. Since it is not clear what these tests are assessing, it is not clear what the scores mean. Further, as shown in Figure 1 above, many public examinations, commercially published tests and teacher-made tests favour some students over others. If you are at the level at which most of the test items are pitched you will have many opportunities to find an item within your capability. But if you are more able or less able there will be fewer items that are pitched at your level. This discrimination is rarely acknowledged, although it has a direct bearing on the usefulness of the scores for selection for further study or employment. For example, universities complain that they cannot make use of examination scores to distinguish between the best students for the purpose of awarding scholarships. If we assume that such examination scores should be available for these purposes, the students have to engage with challenging questions at the highest level. Unless students face questions that they cannot answer we do not know what they do not know.

One of the most serious problems with teacher-made assessments and many external examinations is that the data are reported in ways that militate against the information gathered being used to improve teaching and learning. The traditional ways of reporting scores are *not* in terms of how well the student has satisfied each component of the curriculum. Instead, achievement is reported in terms of place in class (often confused further with meaningless letter grades) or in terms of place in a cohort. These reports are little better than rank orders and mask any evidence that teaching/learning is improving or getting worse. Within schools, cohort reports tend to be used to justify selective schools but do not account for the value added by the teaching within such selective schools. A further difficulty is the assumption that teacher-made test items can distinguish between knowledgeable students and less knowledgeable students. Respectable commercial test publishers do check that items are useful in this way, but most teachers lack the skills and the time to carry out the analyses.

Another problem facing both schools and tertiary courses is the lack of information on the progress made by students over several year levels. This is a consequence of using different tests at different stages of learning without ever asking how the scores on each test relate to the overall continuum of achievement in that subject. As indicated above, research at tertiary undergraduate level in Britain (Izard, *et. al.*, 2003) addressed this issue of distinguishing between the difference in scores because the tests differed in difficulty and difference in scores because learning occurred (or skills were forgotten) over the time interval.

B Using appropriate technology with assessments

Elsewhere I have contended that there are four main issues to be considered with respect to achievement of curriculum intentions in education. The first is the degree of trust that other members of the education community and the general public are willing to place in the various achievements certified by the Education system. The second concerns the relevance of the assessment strategies for their purpose. The third concerns the quality of the available assessment strategies. The quality of assessment for monitoring progress is compromised if there are insufficient items to show curriculum effectiveness, inappropriate statistics for reporting, if valid measures of

change are lacking, and if there is a shortage of assessment expertise. More testing may not be a solution since additional time devoted to *testing* detracts from time for *teaching* and may duplicate effort. The fourth issue concerns the availability of assessment expertise. At least two types of expertise are required: test development and publication in a teacher-friendly mode, and teacher expertise to interpret the assessment evidence provided *and take the appropriate steps to improve student learning*. (Izard, 2002a)

The development of appropriate assessments to inform students and teachers about what students know and do not know has commenced. The methods of analysis are well established (see for example, Wright & Stone, 1979; Wright & Masters, 1982; Wilson, 1992). Lacking are the techniques for applying these strategies to school-based assessments in teacher-friendly ways. It seems obvious that one should seek to teach students what they do not know, rather than continue to teach them what they already know and have demonstrated. It also seems obvious that teachers should assess less and use the assessment information more to ensure that all of their students make progress.

C Assessment-for-learning strategies

One early attempt to use assessment for learning was known as the Individual Mathematics Programme (Izard, *et. al.* 1966, 1969, 1972, 1973, 1974). The kit of materials provided both pencil-and-paper tasks and structured materials, and associated tests. In some of the teaching materials students were able to self-administer a test of the ideas in that booklet. If they knew the material (were above the criterion score on the test) they could move on to the next booklet in the sequence. There were periodic teacher-administered tests to act as a quality control procedure in case students avoided tasks when they should have attempted them. But these materials required experienced teachers who could manage large classes that had students working at different rates, at different levels (spanning several grade levels) and using a range of practical materials.

Using assessment for learning has implications for the design of the assessment instruments and the way the information gathered with those instruments is reported. Since the assessments have to be representative of the curriculum intentions, the assessment specification has to include each of those intentions. For example, if swimming skills is part of the physical education course, the assessment specification has to include that component. The fact that swimming cannot be assessed through a pencil-and-paper test does not excuse a teacher from making an assessment of swimming achievement if that is in the intended curriculum - only the *mode* of assessment alters. Secondly, assessment for learning implies that there will be more than one assessment. Equity considerations lead to an assumption that multiple assessments will be comparable, otherwise progress will not be revealed. To be comparable, each test should be related to the curriculum intentions, and linked technically to the other tests through the pool of items and assessment strategies. Practical considerations about undesirable teaching to the test (instead of to the curriculum) imply that the subsequent tests should be different from earlier tests.

Achieving comparability between tests that are given at different times presents difficulties not often tackled by examination and assessment boards. Some appreciate that this year's examinations should be comparable with last year's examinations and assessments, but this is a comparison between cohorts. Hopefully the distribution of content and complexity in the questions in one year is comparable with another. Without this comparability some groups of candidates would be treated unfairly. Some examination boards *assume* that the papers are comparable and fail to check whether the assumption is correct. When we administer the papers of several years to the same students (in a balanced design) we can collect evidence to check the assumption. I am aware of examples where papers in one subject became more difficult while papers in another subject became easier over time. But the more telling criticism in that instance was the low correlation between results on each of the papers. Were they assessing a common body of knowledge? (The curriculum had been constant over that period.)

But assessment for learning requires tests that are given at different times to the *same* students. The meaning of comparability is different in this context. It would be inefficient to give two *parallel* tests. At the time of giving the first test many of the items could be considered too difficult because the concepts had not been taught and, if the teaching/learning process had been successful, at the time of giving the second test many of the items could be considered too easy. Students would find the earlier test daunting and may be discouraged from engaging in the learning that the curriculum intends. The development of suitable instruments for assessing the progress of learning requires an explicit domain or continuum, perhaps represented by a large pool of assessment tasks. The tests for tracking or monitoring the learning of students over time have to sample the stages in the continuum

One approach to achieving such comparability in assessment involved a calibration of a commercial range of tests and was described at the ACEAB Conference in Malta (Izard, 2002a). Figure 5 (Figure 1 from that paper) shows the strategy of using two tests at each year level.

Year / Test	M7	M8	M9	M10	M11
Year 3	4	4			
Year 4		4	4		
Year 5			4	4	
Year 6				4	4

Figure 5 Data collection for mathematics (from Izard, 2002a)

While this procedure provided the information needed for that study, there are limitations. For example, the instruments were all pencil-and-paper in format. Secondly, the tests were chosen because they matched the curriculum of that school reasonably well. But if the curriculum changes and certain items are no longer relevant, the items will have to be excluded from the scale. (This possibility can be addressed without difficulty using the information already collected provided that the appropriate technical

expertise is available.) However when the curriculum changes there are sometimes additional issues added and a study would have to be carried out to place the associated assessments on the original scale, otherwise progress measured would not include the new area.

D Developing assessment-for-learning instruments

In this section of the paper I address the practical requirements for achieving appropriate assessment instruments and strategies to evaluate progress in learning.

Earlier in this paper, in the section on *Sampling of items across the desired range of skills* I drew attention to deficiencies in tests prepared to suit single grade minimum competence requirements (see Figure 1). Now I will describe the more general case for useful tests for assessment for learning purposes. The discussion will refer to the usefulness of several hypothetical tests (with open-ended items scored right/wrong) in assessing some hypothetical students as illustrated in Figure 6 (an elaboration of a diagram in Izard *et. al.*, 1983c). For each test, each student is shown by an **X** placed on a vertical linear continuum in the same way as students are represented in Figure 1. However a numeral has been added so that the discussion can distinguish between students. Higher achieving students are shown at the top part of the diagram and lower achieving students are shown in the lower part of the diagram. For example, student X1 shows high achievement, and X4 shows low achievement. Items for each of the five-item tests (A to J) are shown to the right of the vertical line representing the achievement continuum. The vertical placement of each item is in terms of item difficulty. Easy items like A1, A2, A3, A4, A5, F1, F2, G1, H1 and J2 are near the bottom of the diagram. Difficult items like B5, F4, G5, I5 and J5 are near the top of the diagram.

Test A would not be useful in distinguishing between the 4 students because the item difficulties are much lower than most of the student achievement levels. If an initial test, most or all of the students would get perfect scores. After a period of learning, this test would not be able to detect improvements in achievement of the 4 students because they cannot obtain any higher than a perfect score. In a development aid context, such lack of improvement may be interpreted as a failure of the intervention but the fault lies with the choice of Test A.

Test B would not be useful in distinguishing between the 4 students because the item difficulties are much higher than most of the student achievement levels. If an initial test, most or all of the students would get zero scores. After a period of learning, this test may be able to detect improvements in achievement of the 4 students because they can obtain a higher score than zero. But all would have to progress from their earlier level to the level above that of X4 to receive credit for their learning. In a development aid context, a lack of improvement may be interpreted as a failure of the intervention but the fault lies with the choice of Test B.

Test C would be useful in distinguishing between X1 and the remaining students. Probably X1 would be correct on all of the items. Test C would not be useful in

distinguishing between the remaining students because all would probably score zero even though they are at different attainment levels.

	Test A	Test B	Test C	Test D	Test E	Test F	Test G	Test H	Test I	Test J
x1		B5				F5	G5		I5	
		B4				F4	G4		I4	
		B3				F3				J5
		B2								
		B1								
x2			C5							
			C4						I3	
			C3						I2	J4
			C2							
			C1							
x3				D5						
				D4			G3			
				D3			G2		I1	J3
				D2				H5		
				D1						
x4					E1					
					E2			H4		
					E3			H3		J2
					E4			H2		
					E5					
	A5									
	A4									
	A3					F2				J1
	A2					F1	G1	H1		
	A1									

Figure 6 Alternative possibilities for tests

Test D would be useful in distinguishing X1 and X2 from X3 and X4. Probably X1 and X2 would be correct on all of the items. Test D would not be useful in distinguishing between X1 and X2 or between X3 and X4 because X1 and X2 would probably score the maximum and X3 and X4 would probably score zero even though all four are at different attainment levels.

Test E would be useful in distinguishing X1, X2 and X3 from X4. Probably X1, X2 and X3 would be correct on all of the items. Test E would not be useful in distinguishing between X1, X2 and X3 because X1, X2 and X3 would probably score the maximum even though they are at different attainment levels.

Test F would not be useful in distinguishing between the students. Probably all would be correct on 2 of the items and wrong on the remaining 3. Results on Test F would imply that the 4 students were the same even though they are at different attainment levels.

Test G would be useful in distinguishing X1 and X2 from X3 and X4. Probably X1 and X2 would be correct on 3 items and X3 and X4 would be correct on 1 item. Test G would not be useful in distinguishing between X1 and X2 or between X3 and X4 even though they are at different attainment levels.

Test H would be useful in distinguishing X1 and X2 from X3 and from X4. Probably X1 and X2 would be correct on all items and X3 would be correct on 4 items and X4 would be correct on 1 item. Test H would not be useful in distinguishing between X1 and X2 even though they are at different attainment levels. Further, the total scores do not reflect the actual differences between the achievement levels. X3 is as far above X4 (with a score difference of 3) as X2 is above X3 (with a score difference of 1).

Test I would be useful in distinguishing X1 from X2 and X2 from X3 and X4. Probably X1 would be correct on 3 items and X2 would be correct on 1 item. Test I would not be useful in distinguishing between X3 and X4 even though they are at different attainment levels. Further, the total scores do not reflect the actual differences between the achievement levels. X3 is as far above X4 (with a score difference of 0) as X2 is above X3 (with a score difference of 1). Similarly X1 is as far above X2 (with a score difference of 2) as X2 is above X3 (with a score difference of 1).

Test J would be useful in distinguishing between each of the students. Probably X1 would be correct on 4 items, X2 would be correct on 3 items, X3 would be correct on 2 items, and X4 would be correct on 1 item. Further, the differences between the attainment levels are reflected in the differences between the scores.

We can contrast Tests H and I with Test J. Tests H and I do not have an even distribution of item difficulty for the range of achievement. These tests discriminate against some of the candidates by giving others a better opportunity to obtain a higher score. Test J avoids this discrimination by using a rectangular distribution of item difficulty. In a development aid context, any improvement over time can be detected and

the difference in scores may be interpreted as the extent of the success for the intervention. Clearly, examination and assessment boards that have assessment strategies like Test A to I should move to using tests like Test J.

Although this example was based on open-ended items scored right or wrong, the same ideas apply to partial credit items. (Partial credit items have more than one mark available. For example, one partial credit item may receive a score of 0, 1, or 2 while another may receive a score of 0, 1, 2, or 3, and so on.) These ideas also apply to multiple-choice items where the test (or sub-test) is long enough for the influences of random guessing to have reduced effect on achievement scores.

For the purposes of explanation this discussion has been confined to small tests. In practice the tests should be much longer (but retain the rectangular item difficulty distribution) in order to make valid inferences about achievement and improved learning.

A package of open-ended computational tests (the Review and Progress Tests in Mathematics [RAPT]) published by ACER (Izard and White, 1982; Izard *et. al.*, 1983a, 1983b, 1983c, 1983d) addressed these issues and others discussed below. The RAPT package illustrates two complementary approaches to test development for formative purposes. The first type of test may be considered a “wide” test, covering a number of objectives in hierarchical fashion with rectangular distributions of item difficulty. In Figure 7 (showing the objectives for the RAPT multiplication tests) the “wide” tests are called Review Tests. Review Test 1 covers the first 5 multiplication objectives. If a very high or perfect score is obtained, Review Test 2 (covering the 4th to the 9th objectives) can be given. If a very high or perfect score is obtained, Review Test 3 (covering the 8th to the 12th objectives) can be given. The scores on the Review Tests are on a common scale so that learning over time can be evaluated.

Where errors are made on a Review Test, the score indicates a band of achievement. Scores (and associated errors of estimate) are graphed on each test. This band of achievement can be related to the list of objectives so that the teacher knows the likely objectives yet to be mastered and can take action to teach those skills. Review Tests have two parallel forms so that the improvement is judged from different evidence on each occasion.

The second type of test may be considered a “narrow” test. In Figure 7 these tests are called Progress Tests. Each Progress Test addresses a single objective and each has a parallel form. When the teacher knows the likely objectives to be mastered, the appropriate Progress Test can be given. If this confirms that teaching to this objective is needed then the teacher can act. After learning the required skills the student attempts the parallel form. If the student is successful the next Progress Test can be given.

Although these test materials are now out of print, the conceptual basis for their development was a formative assessment model. The teacher was given the means to identify a stage of progress and then to teach the appropriate skills to ensure progress.

The assessment strategy allowed the new skills to be confirmed. The key component required for a successful implementation was the use of item response modelling in the test analyses. The effect of such analyses is to place all of the tests on the one “ruler” and to provide benchmarks against which future progress may be gauged.

Objective	Progress Test	Review Test 1	Review Test 2	Review Test 3
Multiply a 2- or 3-digit multiplicand by a 1-digit multiplier, with no regrouping and all digits between 1 and 5.	1	4		
Multiply a 2- or 3-digit multiplicand by a 1-digit multiplier, with regrouping to the highest place value only and all digits between 1 and 5.	2	4		
Multiply a 2- or 3-digit multiplicand, which is a multiple of 10, by a 1-digit multiplier, with no regrouping or regrouping to the highest place only and all digits 5 or less.	3	4		
Multiply a 2-digit multiplicand by a 1-digit multiplier, with regrouping from the units place only and the digits in the multiplier 5 or less.	4	4	4	
Multiply a 2-digit multiplicand by a 1-digit multiplier, with regrouping from both the units and the tens places and all digits between 1 and 5.	5	4	4	
Multiply a 2-digit multiplicand by a 1-digit multiplier, with regrouping from both the units and the tens places and no zeroes.	6		4	
Multiply a 3- or 4-digit multiplicand by a 1-digit multiplier.	7		4	
Multiply a 2-digit multiplicand by a 2-digit multiplier, with both multiplicand and multiplier multiples of 10.	8		4	4
Multiply a 2-digit multiplicand by a 2-digit multiplier which is a multiple of 10,	9		4	4
Multiply a 2-digit multiplicand by a 2-digit multiplier.	10			4
Multiply a 3- or 4-digit multiplicand by a 2-digit multiplier.	11			4
Multiply a 4-digit multiplicand by a 3-digit multiplier.	12			4

Figure 7 Multiplication Objectives for Mathematics (from Izard *et al.* 1983c)

So, those developing instruments for formative assessment have to address two important issues at once. Firstly they need to ensure that their assessment tasks represent the relevant domain or continuum, and to ensure that the tests assembled from those tasks can indicate the proportion of the domain that has been mastered or the progress the student has made along the continuum. Without achieving success with respect to these issues, they will not be able to address the issue of assessing each student's learning in a sound way. Nor will they be able to gauge the effectiveness of their teaching.

Secondly they need to ensure that effective teaching occurs. This may well reduce predictive validity: students not expected to be successful turn out to be successful when good use is made of formative assessment.

NEXT STEPS

This paper has reviewed the problems faced in devising better assessments to monitor learning, and provides practical suggestions for meeting these problems. While no single assessment method is capable of providing evidence about the full range of achievement, the quality of assessments can be improved considerably at little cost – we just have to ask the correct questions. Does this assessment provide valid evidence of improved performance that allows us to infer that learning has occurred? Capricious assessment practices fail to give candidates due credit for their efforts, make it difficult to evaluate the effectiveness of teaching, make it difficult for students to be consistent in evaluating their own work, and discredit the institution making and reporting such assessments.

The design of sound assessments, the development of improved assessment skills, and methods of describing progress are essential requirements for both teacher and the students. The assessments need to have curriculum relevance, be practical and fair, and provide useful information for further learning. The assessment strategies and the approaches to analysis of assessment data presented in this paper are applicable to traditional examinations, project and investigation reports, presentations and posters, judgments of performance and constructed products, and observations of participation, collaborative group work and ingenuity. The reporting has to tell teacher and student what the student probably knows and what is within reach. Equity considerations imply that each student should have ample opportunity to attempt items that are optimal for that student, and that the teacher will assist the student to learn what is not known so that the initial result becomes obsolete. Will you use your assessments to help students learn?

References

Black, P. and William, D. (1998a) "Assessment and Classroom Learning," *Assessment in Education*, Vol. 5, pp. 7-74.

- Black, P. and William, D. (1998b) "Inside the Black Box: Raising Standards Through Classroom Assessment," *Phi Delta Kappan*, p.139
(Also at www.pdkintl.org/kappan/kbla9819.htm)
- Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. New York: Academic Press
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd Ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.
- Houston, K. and Neill, N. (2003). Investigating students' modelling skills. In Q-X Ye, W. Blum, K. Houston and Q-Y Jiang (Eds.) *Mathematical modelling in education and culture*. (pp.54-66). Chichester: Horwood Publications.
- Izard, J.F. (1998a). Quality assurance in educational testing. In National Education Examinations Authority (Eds.) *The effects of large-scale testing and related problems: Proceedings of the 22nd Annual Conference of the International Association for Educational Assessment*. (pp.17-23). Beijing, China: Foreign Language Teaching and Research Press.
- Izard, J.F. (1998b). Validating teacher-friendly (and student-friendly) assessment approaches. In D. Greaves & P. Jeffery (Eds.) *Strategies for intervention with special needs students*. (pp.101-115). Melbourne, Vic.: Australian Resource Educators' Association Inc..
- Izard, J.F. (2002a). Describing student achievement in teacher-friendly ways: Implications for formative and summative assessment. Valetta, Malta: Ministry of Education, Malta and the University of Malta for the Association of Commonwealth Examinations and Accreditation Bodies
- Izard, J.F. (2002b). Constraints in giving candidates due credit for their work: Strategies for quality control in assessment. Valetta, Malta: Ministry of Education, Malta and the University of Malta for the Association of Commonwealth Examinations and Accreditation Bodies
- Izard, J.F. (2002c). Using Assessment Strategies to Inform Student Learning. In P. Jeffery (Compiler): *Proceedings of the Annual Conference of the Australian Association for Research in Education Brisbane December 2002*. (<http://www.aare.edu.au> [search code IZA02378]). Melbourne: Australian Association for Research in Education.
- Izard, J.F., Haines, C.R., Crouch, R., Houston, S.K., and Neill, N. (2003). Assessing the impact of the teaching of modelling: Some implications. In S.J. Lamon, W.A. Parker, and K. Houston (Eds.) *Mathematical Modelling: A Way of Life: ICTMA 11*, (pp. 165-177.) Chichester: Horwood Publishing
- Izard, J., Jeffery, P., Silis, G.F., and Yates, R. L. (1999). Testing for Teaching Purposes: Application of Item Response Modelling (IRM) teaching-focussed assessment practices and the elimination of learning failure in schools. In Peter Westwood & Wendy Scott. (Eds.) *Learning Disabilities: Advocacy and Action* (p 163-188). Melbourne. Australian Resource Educators' Association Inc. (AREA)
- Izard, J.F. & White, J.D. (1982). The use of latent trait models in the development and analysis of classroom tests. In D. Spearritt (Ed.) *The Improvement of Measurement in Education and Psychology*. Hawthorn, Vic.: Australian Council for Educational Research

- Izard, J.F. *et. al.*. (1983a). *ACER Review and Progress Tests in Mathematics - Addition*. Hawthorn, Vic.: Australian Council for Educational Research
- Izard, J.F. *et. al.*. (1983b). *ACER Review and Progress Tests in Mathematics - Subtraction*. Hawthorn, Vic.: Australian Council for Educational Research
- Izard, J.F. *et. al.*. (1983c). *ACER Review and Progress Tests in Mathematics - Multiplication*. Hawthorn, Vic.: Australian Council for Educational Research
- Izard, J.F. *et. al.*. (1983d). *ACER Review and Progress Tests in Mathematics - Division*. Hawthorn, Vic.: Australian Council for Educational Research
- Izard, J.F. *et. al.*. (1966, 1969, 1972, 1973, 1974). *Individual Mathematics Programme (IMP)*. Adelaide: ACER and Rigby.
- Ross, K. N. (1978). "Sample Design for Educational Survey Research", *Evaluation in Education, Vol. 2*, pp. 105-195.
- Ross, K. N. (1993). *Sample Design for International Studies of Educational Achievement*. International Institute for Educational Planning Annual Training Programme Module on Monitoring and Evaluating Educational Outcomes. Paris, France: UNESCO.
- Wilson, M. (1992). Measurement models for new forms of assessment. In M. Stephens & J. Izard. (Eds.) *Reshaping assessment practices: Assessment in the mathematical sciences under challenge*. (pp. 77-98). Melbourne, Vic.: Australian Council for Educational Research.
- Wright, B.D. & Masters, G.N. (1982). *Rating scale analysis*. Chicago, IL.: MESA Press.
- Wright, B.D. & Stone, M.H. (1979). *Best test design*. Chicago, IL.: MESA Press.