Using Assessment Strategies to Inform Student Learning

John Izard

Faculty of Education, Language and Community Studies,

RMIT University, Melbourne, Australia

IZA02378

Presented at the AARE Conference in Brisbane December 2002

Why have assessment?

Everyday life requires citizens to make choices from restricted ranges of options in some contexts. In other contexts they may have perform in some way (like give a speech, perform a song, or write a letter), produce something (like bake a cake), or collate evidence to reach a decision or to argue their case. Students need to learn skills like these to function in society as adults. Their probability of success in such tasks can be judged from successes on a sample of possible tasks. Test items are, in effect, small sample tasks that can provide evidence of success. From this evidence inferences can be made about the extent of achievement. The inferences are stronger if sample tasks (test items) are of high quality.

Student assessment of educational progress is important as a way of

- certifying that standards have been reached,
- informing the student (and parents) of progress made, and
- providing evidence to inform teaching decisions.

Student assessment can provide management information about the implementation of the nation's school curriculum. Curriculum statements describe *intentions*: without valid student assessment practices the *actual* achievements are never compared in a legitimate way with the intentions. Valid student assessments provide quality assurance for certification of school achievement or professional recognition, for informing management, and for evaluation of innovations and development intervention.

Student assessment is a powerful influence on citizen perceptions of what is important. Successful educational achievement gives access to greater income, better work conditions, and further education and allows greater participation in national, regional and local decision-making. Consequently, national school examinations are usually high-stakes assessments. Those who have succeeded are concerned to protect the quality of the education system and keep the assessment requirements just as demanding for those who follow.

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.

When teacher-made 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 is unknown. The tapes will vary in length and the units will probably not be in equal intervals along the tape. Measuring progress is fraught with difficulty because the 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) 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. (If you wish to demonstrate "progress" give the more difficult test first, then give the easier test. Scores after your "teaching" will rise whether your "teaching" was effective or not. The reverse applies if the easier test is given first.) Many teachers do not know the relative difficulties of the tests they give so interpreting the results from their tests is impossible.

Some teachers use tests of differing lengths. The rationale is that less of the work has been covered at the beginning so the earlier tests should be shorter. If you ask 20 questions at the start, 50 questions at the end, *and the questions are of comparable difficulty*, then the average score will rise since the later test has more opportunities for students to be successful.

How is a teacher to indicate progress? What tasks can a student now do that could not be done before? If we know what the student can now do, what learning is a good "bet" for the immediate future? (For that matter, how can administrators set reasonable targets for future achievement if they do not know what these students can do now? Are targets unrealistic if they are distant in complexity and sophistication? Or is it possible to teach calculus before students can read and write?) Before we can indicate the progress that 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.

Now compare this approach with assessment of students in a classroom using many published tests. Once again, the measuring tapes often have no defined units. One tape may have large units while another may have small units and the relationship between these units is unknown. The tapes will vary in length and the units will probably not be in equal intervals along the tape. Some tapes will only compare student achievements with the performance of a reference group, but there is no evidence to allow the teacher to judge whether the reference group is an appropriate comparison group for the class. (For example, reading ages involve a reference group performance, but often there is no evidence about that reference group. One set of tests used widely in Australia used students from Scotland in the 1940s as the reference group.) Even if the reference group is from the same nation, there is an assumption that the group studied the same curriculum, had the same quality of teaching, and access to the same level of resources. Some of the tapes will be relevant only for a given age group or year of schooling. Students ahead of or behind their classmates will be judged in terms of the age cohort (and remember, by definition, half of any group is above the average for that group).

A further difficulty is in the reporting of scores. Results are interpreted relative to the reference group (whether relevant or not). We do not learn from the data collected in the classroom what students know or do not know, because this information is ignored in comparing students and reporting in terms of their relative standing.

Student assessment has many purposes

Assessment of students can provide evidence for management purposes, for certification of achievement, for teaching purposes, for selection purposes, and for monitoring progress of development projects. Unfortunately, student assessment is also used for inappropriate purposes. (Izard, 2001)

A. ***Student Assessment for Management Purposes***

Student assessment provides a means of ensuring that administrators have measures of quality for management purposes. National system performance in basic education may be assessed by national examinations at the end of elementary schooling and at the end of secondary schooling.. Memory plays an important role in success on such tests. But many of these tests are not equated for difficulty. When the pattern of marks rises or falls, there were several possible explanations: each test is measuring something different, the tests are changing in difficulty or the population achievement level is changing.

B. ***Student Assessment for Certification of Achievements***

Where a certification process becomes well known, the certifying authority serves to reassure the public that meaningful learning is occurring in schools and that assessment of that learning is systematic, valid and fair in that due credit is given for work done for assessment. In some constituencies, students are blamed if they do not learn. In others, teachers are blamed for not teaching well. Others criticize the provision of teaching materials or equipment. Assessment instruments and procedures are criticized too, if students cannot achieve (as judged by pass rates on the assessments). In fact, a sound education system requires all of these components:

- Well-trained teachers who can help students learn,
- Well-designed curriculum that is coherent and builds on past success,
- Teaching and support materials that suit the curriculum and are accessible to all students,
- Students who are motivated to learn and who persevere in their pursuit of knowledge, and
- Assessment strategies that give due credit to the quality of achievement and generalisable skills.

In accreditation systems, factors like these are checked to see whether a course or a teaching institution is worthy of being trusted to operate for limited period (often 3 years) without more frequent review. The assessment strategies are only part of the story but they are often more visible to the general public. The general public's knowledge of what is taught and how it is taught are often based on what happened in their own schooling or on perceptions gained from their children's schooling. They are often not aware of the textbooks a school uses or does not use (unless they are asked to pay for them in private schools).

When the general public or other educational institutions place little trust in the assessment practices of educational bodies such as schools and universities, alternative external schemes are set up to exercise additional quality control of the certification and/or selection.

A. ***Student Assessment for Teaching Purposes***

Use of student assessment for teaching purposes involves identifying where students have reached in their learning with a test or tests, 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 test results. Many teachers and administrators accept and support analysis of test data to improve teaching and learning, but practical implementation has been found wanting. (Izard, 1998; Black and Wiliam, 1998a, 1998b).

B. ***Student Assessment for Selection Purposes***

In one sense, all of the tests administered in basic education classrooms contribute to the school's decision whether the student will be promoted to the next class. In that sense the tests are selection tests. In some systems (such as Australian State education authorities), students are promoted with their peers, and the teachers are required to ensure that every child makes progress.

C. ***Student Assessment for Monitoring Progress of Development Projects***

The use of student achievement as a performance indicator in monitoring progress and contributing to an understanding about benefits of investment and intervention is a recent international development. For example, the Organization for Economic Cooperation and Development (OECD) has commissioned a large study on performance indicators for assessing student achievement [OECD Programme for International Student Assessment (PISA)] with the intention of using regular surveys to monitor national education systems. The Australian Council for Educational Research (ACER) leads the consortium conducting the study with many developed nations involved. Information on this study (still in progress) is available at www.pisa.oecd.org.

D. ***Student Assessment for Inappropriate Purposes***

Much of the duplicated testing conducted in schools is to "make students achieve higher scores". There is a belief that repeated testing leads to higher scores. Further, it is believed that increasing the amount of testing raises the scores even higher. This belief is erroneous. A farming analogy emphasizes this: the farmer's corn will grow better if appropriate nutrients and water are provided in timely fashion. Measuring the height of the corn frequently is *not* going to improve the yield at all.

The need to have informed candidates when tests are given is accepted. There are two types of knowledge needed by the candidate. One is the knowledge acquired through experience and study with the help of the classroom teacher(s). The published curriculum indicates which knowledge is required and the tests should sample competencies and skills from this published curriculum. The second is knowledge of what the testing tasks will be like and how candidate responses are to be recorded. This knowledge must be available to all candidates, not just those who can afford to purchase 'review tests'. (Teachers need similar support when receiving that examination's *results* to ensure that the results are interpreted correctly and consistently.)

Which of these purposes are relevant for school-based assessment?

In the search for valid assessment strategies those responsible for national school assessments and certification of school achievement have sought to go beyond relying on external examinations (whether open-ended, multiple-choice, or some combination of these. Further, teachers and students need to build on what is already known, and to act as soon as possible to teach and learn what is not known but is within reach. 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.

School-based assessment of students can provide evidence for management purposes within the school including monitoring student progress, can contribute to certification of achievement, and is essential for teaching purposes. Assessment strategies that show progress in these curriculum terms are available now. This approach has been shown to work from the first years of schooling through to university entrance level, in undergraduate studies, and at graduate level (for example, see Griffin & Forwood, 1991; Haines & Izard, 1993; Izard, 1994;). The methods of analysis are well established (see for example, Wright & Stone, 1979; Wright & Masters, 1982; Wilson, 1992). The approach can handle performance tasks as well as traditional pencil-and-paper tests (for example, see Haines, Izard, Berry, et al., 1993; Masters & Forster, 1996a, 1996b; Forster & Masters, 1996a, 1996b, 1997; Izard, 1997), and has been used in national and international studies (for example, see Izard, 1992, 1996; Lokan, Ford & Greenwood, 1996). What has been lacking is the techniques for applying these strategies to school-based assessments in teacher-friendly ways.

Technical limitations of current school-based assessment

Teachers assume that the 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. If we use the analogy of a ruler to represent a test, teacher-made tests may be considered as rulers of different lengths with different markings on the ruler. 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.

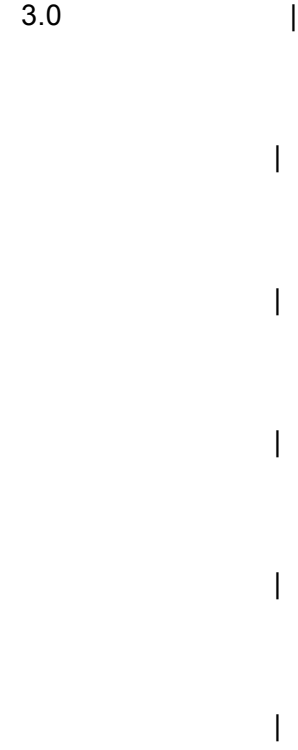
Test items constructed by teachers without technical support do not always distinguish between students with relevant knowledge and students lacking such knowledge. Tests differ in number and format of items. Items vary in difficulty but are treated as equivalent in difficulty. Two tests of the same topic are assumed to cover the same work without empirical validation. Difficulty is not controlled from one occasion to the next. So measuring progress becomes problematic: one does not know whether a score difference is due to the test changing in difficulty, the students making progress, or the tests being unrelated to each other.

Some published tests suffer the same technical limitations. Further, many published tests cannot show progress across grade levels because each grade-level test only allows comparisons within that grade-level and the relation between tests is not provided in the test manual even if known. To extend the ruler analogy used earlier, there are several distinct rulers but we do not know how the marks at one grade-level relate to marks on another grade-level.

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. (The comments have been added to an actual analysis.)

Most of the students represented in Figure 1 do not have sufficient items at or around their level of skill. Because items on this test are not spread evenly over the range of student achievement, some students are favored in the number of items that match their level of achievement. In Figure 1 the majority of items are higher in difficulty than the achievement level of the majority of students. Items 1, 24, 38, 10, 16, 31, 22, 2 and 20 are the only items pitched at the level of most of the students. By noting what each item measures, examiners and teachers are able to identify the topics within reach of the students at that level. When test results are used for monitoring progress purposes, single grade minimum competence results obscure any progress that has been made due to ceiling effects and floor effects. For example, consider the 48 top scorers shown in Figure 1. If this was their pre-test result then their post-test result cannot be any better. 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 a single item or two items - hardly a convincing measure of attainment status. Izard (2002a) reviews other constraints in giving candidates due credit for their work in his paper on strategies for quality control in assessment.

The most serious technical limitations are reporting procedures that compare students with students rather than compare students with curriculum intentions. Tests that provide evidence of learning in a global benchmark-comparison sense report comparisons between people. A better alternative is to report on a classroom skills-people matrix basis so that the teacher can identify what needs to be taught from the results. Without this we do not have *formative* assessment.



```
                    |
                  X |
2.0               X | 384 students represented by X have no items
                    | that match their estimated achievement level.
                  X | They do not receive due credit for their achievements.
                  X |
                  X |
                  X |
                 XX |
                XXX |    44    48  <- Most difficult items.
1.0              XX |    43
                XXX |
               XXXX |     3    21    50
```

```
                    XXX  |      7

These students had      XXXX  |     17     26     33     35     42

more items to show     XXXXXX  |      8      9     13     23     25     27     36

what they could do.     XXXX  |     29     32     45     46     49

 0.0             XXXXXXXXX  |     11     14     28     30     34  <- average difficulty

                    XXXXXXX  |      5     19

                XXXXXXXXXXX  |     15     18     39     40

                    XXXXXXX  |      4      6     12     41

              XXXXXXXXXXXXXX  |     47

                 XXXXXXXXXX  |      1     24     38

                XXXXXXXXXXX  |     10

               XXXXXXXXXXXX  |

-1.0     XXXXXXXXXXXXXXXXXXX  |     16     31 <- Modal student score.
```

XXXXXXXXXXXX | At least 2304 students (who scored at least 1 correct item)

XXXXXXXXXX | 22 represented by X only have up to 3 items which match

XXXXXXXX | 2 their estimated achievement level. They do not have

XXXXXXX | enough suitable items to show what they know. These

XXXXX | students are additional to those who scored 0. (Not shown.)

XXX |

X | 20 <- Easiest item.

-2.0 X |

X |

|

|

Each X represents 48 students

Some thresholds could not be fitted to the display

The page was not wide enough to show all the items that were like 8, 9, 13, 23, 25, 27, and 36.

Figure 1 Examination Item Estimates (Thresholds) for a 50 item Grade 4 English test (from Izard, 2002a)

Improving assessment strategies for learning purposes

To realise the benefits of formative school-based assessment some changes need to be introduced in schools and the public must be informed about why this is happening. Deterministic approaches need to be replaced by probabilistic approaches. With real students we can never be certain that they will be perfectly consistent from one time to the next. 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.

Since the range within any grade-level is much wider than the usual grade-level test (and the problem is exacerbated by minimal competency tests), grade-level tests must be scaled onto a longer scale. This will allow all students in a given classroom to demonstrate progress in a fair manner. Alternative and equivalent forms of tests will be necessary if assessment is to avoid teaching to the test. But improving the assessment (or evidence-collecting) process is only half of the critical issue. *Teachers need support to respond to the information and to provide differential teaching to meet curriculum intentions for each student.*

Research to improve school-based assessment

Black and Wiliam (1998a, 1998b) 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. The research now described has results consistent with this view. In the government school chosen to illustrate this different approach to assessment, concerns about assessment arose in 1995. The outer suburban school with an enrolment of approximately 280 students sought to "validate school achievement by the use of external indicators obtained from published tests selected to match the school's stated goals." (Izard, Jeffery, Silis and Yates, 1999.) The school's management wanted a view of how the students' achievements and the teaching practice fitted the 'big picture' and also wanted information that was obtained from the assessment to have a strong link to individual student learning needs. Standardized scores, percentiles and comparisons with state benchmarks were *not* the objective in this program, but improving the school's ability to deliver programs that matched each student's learning needs was the objective.

This longitudinal research in actual classrooms commenced in 1996 and continues. The procedures for monitoring student achievement through value-added assessment had to address a number of issues (see Izard, 2002b). Existing published tests were reviewed to identify those that matched the curriculum of the school. An initial data collection was essential to scale separate age-based tests in mathematics onto a common mathematics

achievement scale. A similar data collection was needed to equate two parallel forms of a wide-range spelling test. Similar work was carried out in other subject areas. The results from the tests had to be presented as an item-pupil matrix so teachers knew the achievements for each student and in terms of scaled scores so that progress could be evaluated. Teachers had to adapt their teaching to ensure that a student's earlier work was mastered before proceeding to more advanced work.

The results of this research show effect sizes for standardized spelling tests of about 4 over three years (for those not achieving perfect scores). The test form used alternated from year to year. Initially the wide-range spelling tests covered the range of spelling achievement of students in Grades 3 to 6. But as teachers addressed the teaching/learning needs reflected in the results, the achievement levels gradually increased until many students achieved perfect scores. Since the method of analysis excludes students with zero or maximum scores (on the grounds that odds of success on items cannot be estimated) this meant that consideration had to be given to expanding the spelling tests or concentrating the effort on a smaller group. The staff argued that further spelling lessons were unnecessary for students who had developed proficiency in spelling and, presumably, a method of attack for new words. Student with this proficiency would be better served by working in some other area of English language. At the second external triennial school review in 2001 a consequence of the staff decision was that students achieving perfect scores had no measures of progress in spelling. Only those within the achievement range of the standardized spelling tests and staying within that range over the period from 1998 (scale score mean 1.36) to 1999 (scale score mean 2.49) to 2000 (scale score mean 3.34) were in the analysis. (These students were the weakest in spelling skills.) The effect size for the change from 1998 to 2000 was 4.09. It was estimated from the difference between the 1998 and 2000 table entries (1.98) and the average error for each score (0.48).

The results of this research show effect sizes for a scaled set of standardized mathematics tests of between 2.9 and 3.5 over three years. The test form used differed from year to year but it was possible to interpret the results on a common scale because the tests had been linked empirically. Because there was a wider range of tests, more students were in the progress records. (Only those who had left the school or who had recently joined the school were not reported.) Those in Grade 4 in 1998 had a scaled score of +0.60 in 1998, +0.88 in 1999 and 1.84 in 2000. The effect size for the change for 1998 to 2000 was 3.46. It was estimated from the difference between the two 1998 and 2000 entries (1.25) and the average error for each score (0.36). Those in Grade 3 in 1998 had a scaled score of +0.17 in 1998, +0.69 in 1999 and 1.33 in 2000. The effect size for the change for 1998 to 2000 was 2.91. It was estimated from the difference between the two 1998 and 2000 entries (1.16) and the average error for each score (0.40). The testing was extended to Grade 2 in 1998. Those in Grade 2 in 1998 had a scaled score of -0.70 in 1998, +0.11 in 1999 and +0.74 in 2000. The effect size for the change for 1998 to 2000 was 3.46. It was estimated from the difference between the two 1998 and 2000 entries (1.45) and the average error for each score (0.42). These results show that cohorts differ in average achievement, sometimes by substantial amounts, but all classes showed progress.

The results reported above are from the tests the school chose to match their curriculum. An important question is how the results of external testing conducted by the State education authority matched those from essentially internal assessment. The report of the second external reviewer states, "A strong record of academic achievement is a feature of this school and all assessment indicators confirm this finding. In addition to the accountability assessment requirements the staff have developed an extensive program to monitor and guide their teaching approach. This locally developed assessment program called *Testing for Teaching* is very important in shaping the learning of all students in the school and monitoring their progress in particularly in key areas of literacy and numeracy. This

assessment effort complements DEET accountability requirements and appears to make an important contribution to the very strong academic performance in the school" (from Izard, 2002b). It is clear that this approach is a better practical model for describing progress in schools.

Needed research and development

The conduct of this research has shown that changes in teaching approach take time to implement. This is partly because normative assessment (reading ages, percentile ranks, place in class, and so on) is so heavily entrenched, in spite of the lack of useful information for teaching purposes. Four main areas need further attention (but teachers do not need to wait for this research before helping their students improve their learning). These are development of better tests to provide feedback to teachers and students, development of linked tests to cover a wider range of achievement, preparation of teacher guides to interpreting the assessments in learning/teaching terms, and development of exemplary teaching materials so that teachers can act on the information.

Item response modelling (sometimes known as item response theory) was the only way of tackling the analyses for the research described above. But there is more to be done here too. For example, the effects of expanding the number of items at a given grade level need to be compared with the effects of expanding items across grade-levels. Another example is the development of better methods of linking the teaching implications to the scores obtained. A further example is the validation of teaching strategies to make sure that they do improve learning.

Another key area for research and development is the issue of dissemination. There is a substantial lag in time between discovery and application in Education, and this is particularly true in the case of educational research. A prime example is the conduct of educational research. Some of the persistent (inappropriate) analysis methods are still being used at the beginning of the twenty-first century even though the faults in the techniques have been presented in undergraduate textbooks since the 1960s at least. (Izard, 2001) Statements are made that imply one can trade off validity against reliability. If it is not valid it does not measure what is intended: there is no point in proceeding. It cannot be valid for any purpose unless it is reliable (internally consistent).

Much of the use of statistical significance testing in educational research is flawed. Educational researchers have logical problems in dealing with the null hypothesis. Failure to reject the null hypothesis does *not* mean that there *is no difference*. Logically it means we have *no evidence* of a difference. We may have a true difference but the *power* of our statistical test to detect that difference may be inadequate. Or there may not be a difference. *We just do not know*.

There are other misconceptions about testing for statistical significance. Most educational research uses students from a limited number of schools, colleges or universities. Within the group chosen, the subjects are more alike than if simple random sampling was used to choose them. For a start, they have shared the same schooling, they probably live in similar socio-economic conditions. This is reflected in a higher intra-class correlation. The effect of this is to over-estimate the statistical significance in most cases (Ross, 1993). It is necessary to adjust for this *design effect*.

Substantive significance is more useful than statistical significance. When a new teaching method is compared with another method, the magnitude of the difference between the

methods may not be large enough to be worth the expense of changing methods. Effect sizes give an indication of the response to the question "How big a difference is it?".

With these dissemination difficulties among the educational research community, the likelihood of an easy, and fast, transition to better learning in schools through formative assessment is small. It is essential to find better ways to convince teachers (and parents) of these better methods.

Conclusion

Perceptions (of both technical experts and lay persons) have a considerable effect on decisions. Where there is no strong culture of scientific endeavour, opinions are what count. The opinions of national decision-makers count most, and such persons do not have to consider empirical evidence from a scientific perspective. This is particularly evident in developing nations where assessment expertise is lacking. Such decision-makers are replaced periodically in democratic nations, but do not wear the effects of their decisions on students. Student learning that is compromised by ineffective management and failure to support all teachers with adequate resources for learning condemns many students (particularly those living in poverty) to an unnecessarily limited education.

By way of contrast, the constant threat of litigation makes many examination boards cautious about publishing results before they check (through item analysis) that every item (as scored) distinguishes between able candidates and less able candidates *in the right direction* (able candidates scoring higher on the item than less able candidates). (But see Ludlow, 2001 [at http://epaa.asu.edu/epaa/v9n6.html] for an example in teacher licensure testing in the United States of America where this did not apply.)

The main lesson is that public accountability requires some understanding of what is happening and its underlying purpose. To many, including reviewers of development aid programs, one 'test' is as good as any other 'test'. Tests for different purposes are seen as interchangeable, without consideration of the implications for the quality of the inferences. To use an analogy from a transport context, a motor bike can be interchanged with a 40-seater bus. If there is only one person to transport, the motor bike may even have cost benefits. But fitting 40 people on a single motor bike is ludicrous. In the student assessment context it is essential to recognise why the technical properties of one test are not necessarily directly interchangeable with another.

Assessment information should inform the teacher (and the student) of what tasks can be attempted successfully, what skills and knowledge are being established currently, and what skills and knowledge are not yet within reach. Assessment strategies that show progress in these curriculum terms are available now. The direct information about what is known and what is yet to be learned is far better information than the old "normative" approach that compares student with student without mentioning what each can do and what each has yet to learn.

The benefits of focussed teaching as a consequence of informed assessment are obvious. Non-teachers can understand the progress made by students over several years because the results can be presented in graphical formats. Because the parents can see the progress made, they support the school program. They understand better what the children are studying because success can be described by the teachers in terms of "your child can do tasks like this ..." and "we are now moving on to tasks like this ...". The extreme groups in the classroom (whether more able or less able) are receiving attention and in some cases the able students are working at levels higher than their classmates. Like Black and Wiliam

(1998a, 1998b), I wonder why the earlier approach survives when better approaches are available.

References

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

Black, P. and Wiliam, 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.

Forster, M. & Masters, G. (1996a). *Portfolios.* Melbourne, Vic.: Australian Council for Educational Research.

Forster, M. & Masters, G. (1996b). *Performances.* Melbourne, Vic.: Australian Council for Educational Research.

Forster, M. & Masters, G. (1997). *Projects.* Melbourne, Vic.: Australian Council for Educational Research.

Griffin, P. & Forwood, A. (1991). *Adult literacy and numeracy competency scales. An International Literacy Year Project*. Melbourne, Vic.: Phillip Institute of Technology.

Haines, C.R. & Izard, J.F. (1993). Authentic assessment of complex mathematical tasks. In S.K. Houston (Ed.) *Developments in curriculum and assessment in mathematics* (pp. 39-55). Coleraine, Northern Ireland: University of Ulster

Haines, C.R., Izard, J.F. & Berry, J.S. et al. (1993). Rewarding student achievement in mathematics projects, (Research Memorandum 1/93). London: Department of Mathematics, City University

Izard, J.F. (2002a). 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. (2002b). 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. (2001). Consultancies and research projects in Asian countries. Paper presented at the Research Conference for Postgraduate Students and RMIT Staff June 1st 2001. Melbourne: RMIT University

Izard, J.F. (1998). 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. (1997). Assessment of complex behaviour as expected in mathematical projects and investigations. In S.K. Houston, W. Blum, I.D. Huntley & N.T. Neill(Eds.) *Advances and perspectives in the Teaching and learning mathematical modelling: Innovation, investigation and applications*. (pp.109-124). Chichester, West Sussex, England: Albion Publishing

Izard, J.F. (1996). The design of tests for national assessment purposes. In P. Murphy, V. Greaney, M. Lockheed & C. Rojas (Eds.) *National Assessments: Testing the system*. (pp.89-108). Washington, D.C.: The World Bank.

Izard, J.F. (1994). Strategies for assessing projects and investigations: Experience in Australia and United Kingdom. In Mauritius Examinations Syndicate (Eds.), *School-based and external assessments*. (pp.214-222). Mauritius: Mauritius Examinations Syndicate.

Izard, J.F. (1992). Assessment of learning in the classroom. (Educational studies and documents, 60.). Paris: United Nations Educational, Scientific and Cultural Organisation

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)

Lokan, J., Ford. P. & Greenwood, L. (1996). *Maths & Science on the line: Australian junior secondary students' performance in the Third International Mathematics and Science Study*. Melbourne, Vic.: Australian Council for Educational Research.

Ludlow, L. H. (2001) "Teacher Test Accountability: From Alabama to Massachusetts." *Education Policy Analysis Archives*, Vol 9 No. 6 February 22, 2001 [http://epaa.asu.edu/epaa/v9n6.html]

Masters, G. & Forster, M. (1996a). *Developmental assessment.* Melbourne, Vic.: Australian Council for Educational Research.

Masters, G. & Forster, M., (1996b). *Progress maps.* Melbourne: Australian Council for Educational Research.

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, (1982). *Rating scale analysis.* Chicago, IL.: MESA Press.

Wright, B.D. & Stone, M.H. (1979). *Best test design*. Chicago, IL.: MESA Press.