Quality assurance: Asking the right questions.

John Izard
School of Education, RMIT University, Melbourne, Australia

The Assessment Context

Examinations and evaluations of performance used in universities and schools are usually high-stakes assessments. Those responsible for such assessments usually try to ensure that the tasks set as part of such assessments are of high quality, that the tasks represent a balance of relevant content, and that, as far as is possible, each candidate has a fair chance to respond to the tasks and to receive due credit for the responses. Improving the quality of assessments is taken very seriously. IAEA, ACEAB and development agencies such as World Bank, Asian Development Bank and UNESCO have invested considerable resources in conferences, training workshops, consultancies and publications relating to quality assessment. But unless one asks the right questions, those making assessment decisions may, in fact, be threatening the validity and quality of the assessments. In this paper I propose to show, using analyses of actual examination data, how some decisions about assessment strategies threaten the quality of those assessments. Practical advice on how to avoid these threats will be provided in the paper

International collaboration

In 1990, UNESCO commissioned a report from IAEA on recent international trends and major developments on the assessment of learning achievements both in and out of school. The report I prepared included a bibliography of relevant research, and identified the trends and developments in assessment theory and practice that were likely to be significant in the future: *Assessment of learning in the classroom* was published by UNESCO (Izard, 1992).

IAEA and World Bank collaborated in delivering workshops to develop assessment skills. In late 1991, the World Bank commissioned IAEA to organise and conduct three regional workshops on issues related to assessment to monitor educational achievement on a national scale. The regions selected were Latin America, Asia and Africa.

- Latin America: Conducted in Chile, August 31 – September 4, 1992. This workshop was coordinated and chaired by Protase Woodford, Educational Testing Service, USA.
- Asia: Conducted in the Philippines, November 8 – 13, 1992. This workshop was coordinated and chaired by John Izard, Australian Council for Educational Research.
- Africa: Conducted in Kenya, January 11 – 15, 1993. This workshop was coordinated and chaired by Christopher Modu, Educational Testing Service, USA.

Some of the teaching materials used in these workshops were published subsequently by the World Bank (see Izard, 1996, for an example).

IAEA, UNESCO and World Bank, have published other materials to help improve the quality of assessment. For example, IAEA has sponsored *A Teachers Guide to Assessment* (Frith & Macintosh, 1984) to assist teachers in the development of test construction skills. It was published on behalf of IAEA by Stanley Thornes. UNESCO's International Institute for Educational Planning (IIEP) has prepared a number of Training Modules in *Quantitative Research Methods in Educational Planning* (available on the IIEP website http://unesco.org/iiep and at http://www.sacmeq.org/training.htm). Training modules 5 (Withers, 2005), 6 (Izard, 2005a) and 7 (Izard, 2005b) relate directly to assessment.

Annual conferences of IAEA from 1976, (see the IAEA Conference list, available on http://www.iaea.info/index.php?option=com_conferences&Itemid=45 as at 25 Jan. 2006), have been devoted to various aspects of assessment. Topics have included admission to higher education (1976), assessing teacher effectiveness (1978), assessing school achievement in various ways (1983, 1984, 1990, 1993, 1995 and 1996) and setting standards (1977 and 1987), using assessment to increase opportunity (1980, 1981 and 1986) or to serve the needs of learners (2004 and 2005), evaluating educational programs and systems (1979, 1985 and 1991), education for employment (1982, 1988 and 1994), selection for higher education (1989 and 1992), and equity issues (1997, 1998, 1999 and 2003). But assessment is of wider interest. For examples in mathematics and communication see Izard, (1993), Izard & Haines (1994) and Haines & Izard (1994).

Quality assurance strategies

Strategies used to improve quality have been focussed on the representativeness and usefulness of the tasks used to provide evidence of achievement (Izard, 2005c). (A useful test may be regarded as one where students of different achievement levels receive different scores ordered according to their achievement and students of the same achievement level receive the same scores.) These strategies include preparing a specification for the assessment, providing an appropriate variety in the range of task complexity, reviewing tasks (items) before conducting trials with similar samples of students (and checking that scoring rubrics are not ambiguous), using trial data to check whether items can distinguish between able students and less able students (as judged from the test as a whole), and eliminating items that are inconsistent with other items. Where the scoring rubrics require trained scorers, attention has been given to achieving consistent scoring by each judge. (We do not expect perfect agreement, and if there is perfect agreement, some will suspect collusion.) The administration of the tasks, often under secure conditions, attempts to ensure that each candidate has a fair chance to respond to these tasks.

The constant threat of litigation in developed nations makes many examination boards cautious about publishing results before they check 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) for an example in teacher licensure testing from USA where this quality control mechanism did not apply.]

But several quality assurance issues have not been well addressed. Several examples are given below. The first two address sampling of *items* across the desired range of task complexity. The second two address sampling of *students* across subsets from the desired range of items. In all cases it is explained how wrong inferences may be made. Possible solutions are also presented.

A   Example 1 – Restricted range of items

The first example (illustrated below, with comments added in red to an actual analysis) looks at the sampling of items across the desired range of task complexity.

```
  3.0                            |
                                 |
                                 |
                                 |
                                 |
                                 |
                                 |
                             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     XXXXXXXXXXXXXXXXXXXX   |      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)

The test in the example failed to ensure that the spread of item complexity is balanced, enabling candidates at different achievement levels to have similar numbers of tasks at their level of achievement. Without such balance, the assessment tasks are biased in favour of some candidates. This inequitable situation is illustrated in Figure 1 (Figure 1 from Izard, 2002a). 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 favoured 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. This type of sampling distortion occurs when test items are prepared to suit single grade minimum competence requirements, and is exacerbated when the test produced does not match the actual range of achievement of the students being assessed. Many students miss out on the opportunity to demonstrate their skills and knowledge.

In a development aid context where evidence of improvement after intervention is required, 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 if measurement 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 initial success on a single item or two items – hardly a convincing measure of attainment status. Izard (1998a) reviews other constraints in giving candidates due credit for their work in an earlier IAEA conference paper on strategies for quality control in assessment. **Solutions** to this problem depend upon collection of trial data on tasks over the full range of likely achievement so that students at each achievement level have a comparable number of items to attempt.

B Example 2 – Effects of using item banks

Earlier in this paper a useful test was described as one where students of different achievement levels receive different scores and students of the same achievement level receive the same scores. Izard (2005c) investigated the usefulness of some tests in assessing more than 700 actual students known to be at different achievement levels (from a larger sample of the same tasks). Figure 2 provides a model for the study. In Figure 2 each student is shown by an **X** placed on a vertical linear continuum. A numeral has been added to 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.

An overall test of 21 items was administered to more than 700 primary and secondary school students and the results were analysed using item response modelling software. (The traditional internal consistency index was 0.83.) The 54 students who obtained a perfect score were excluded from the study because their achievement level could not be determined. (We need to know what they cannot do, as well as what they can do, to estimate their achievement level: with perfect scores we have no evidence of what they cannot do.) The results were sorted to obtain groups of students with the same scaled achievement level according to the overall test of 21 items Since classes in mathematics are not typically large, students were sampled randomly from each of these larger same-achievement groups to match as closely as possible the hypothetical students shown in Figure 2. Generally five achievement levels were used. Items were chosen to replicate the series of tests with 5 items. Raw and scale scores were calculated for each student. Apparent achievement levels reflected by the test results were compared with the actual achievement levels.

```
      |  Test    Test    Test    Test    Test    Test    Test    Test    Test    Test
      |   A       B       C       D       E       F       G       H       I       J
      |
      |           B5                              F5      G5              I5
      |
      |           B4                              F4      G4              I4
      |
      |           B3                              F3                              J5
      |
      |           B2
      |
      |           B1
      |
  X1  |
      |
      |                   C5
      |
      |                   C4                                              I3
      |
      |                   C3                                              I2      J4
      |
      |                   C2
      |
      |                   C1
      |
  x2  |
      |
      |                           D5
      |
      |                           D4                      G3
      |
      |                           D3                      G2              I1      J3
      |
      |                           D2                              H5
      |
      |                           D1
      |
  x3  |
      |
      |                                   E1
      |
      |                                   E2                      H4
      |
      |                                   E3                      H3              J2
      |
      |                                   E4                      H2
      |
      |                                   E5
      |
  x4  |
      |
      |   A5
      |
      |   A4
      |
      |   A3                                      F2                              J1
      |
      |   A2                                      F1      G1      H1
      |
      |   A1
      |
```

Figure 2 Alternative possibilities for tests (from Izard, 2005c)

A sample set of results is presented in Figure 3. Each graph has a vertical line or scale representing the continuum of achievement as indicated by the 21 items on the overall test. The difficulty level of each item (numbered with a **bold** numeral on the right hand side of each scale) is shown by its position on the vertical scale. The position of each student sampled is shown by the score obtained, on the left of each vertical scale. For example, the top 3 students on Test A had the same achievement level according to the 21-item test and all scored 5 on Test A. They are shown at the top of the graph and are well clear of the next achievement levels as indicated by the separation on the vertical scale.

```
     Test A              Test B              Test F              Test J
   (N=15 L=5)          (N=15 L=5)          (N=15 L=5)          (N=18 L=5)
----------          ---------           ---------           ---------
 555 |                   |                   |               555 |
     |                   |                   |                   |
     |                   |                   |                   |
     |                   |                   |                   |
     |                   | 11                | 11                |
 555 |                   |                   |                   |
     |                   |                   |                   |
     |                   | 12                | 12                |
     |                   |                   |                   |
 545 |                   |                   |                   | 11
     |                   | 13                | 13                |
 555 |                   |                   |               454 |
     |                   | 14                |                   |
     |                   |                   |                   |
 545 |                   | 10                |                   |
     |                   |                   |                   | 13
     |                   |                   |               433 |
     |                   |                   |                   |
     |                 . |               232 |                   |
     |                   |               222 |                   | 16
     |                   |                   |                   |
     |                   |               222 |                   |
     |                   |                   |                   |
     |               000 |               232 |               222 |
     |                   |               222 |                   |
     |                   |                   |                   | 17
     | 18            000 |                   |                   |
     |                   |                   |                   |
     |               000 |                   |               111 |
     |  8                |                   |  3                |
     |                   |                   |                   |
     |               100 |                   |                   |  9
     |  1                |                   |  4                |
     |                   |                   |                   |
     |  3                |                   |                   |
     |                   |                   |                   |
     |                   |                   |               000 |
     |  4            000 |                   |                   |
---------           ---------           ---------           ---------
```

Figure 3 Student scores on alternative tests A, B, F and J (from Izard, 2005c)

Test A was designed as a very easy test (with all items near the bottom of the scale). The majority of students achieved the same score of 5. Although they are clearly *different* on the overall test, the results of Test A imply that most are *identical* in achievement as shown in Table 2. Students from *different* achievement levels should receive *different* scores: most students from the five achievement levels received the *same* score. ***Test A fails to detect real differences between students.***

Test B was designed as a very difficult test (with all items near the top of the scale). The majority of students achieved the same score of 0. Although they are clearly different on the overall test, the results of Test B imply that most are identical in achievement as shown in Table 3. Students from *different* achievement levels should receive *different* scores: most students from the five achievement levels received the *same* score. ***Test B also fails to detect real differences between students.***

Table 2 **Comparisons of Apparent and Actual Achievement Levels: 5-Item Test A Compared with Actual Test of 21 Items**
(from Izard, 2005c)

Students (*n*=3 for each group)	Actual Levels of Scaled Scores (logits)	Apparent Levels on Test A	Decision based on Test A
Group 1	3.47	13 of the 15 students are at the same level	There is no difference between the 13 students
Group 2	2.67		
Group 3	2.16		
Group 4	1.76		
Group 5	0.56		

Table 3 **Comparisons of Apparent and Actual Achievement Levels: 5-Item Test B Compared with Actual Test of 21 Items**
(from Izard, 2005c)

Students (*n*=3 for each group)	Actual Levels of Scaled Scores (logits)	Apparent Levels on Test B	Decision based on Test B
Group 1	-0.16	14 of the 15 students are at the same level	There is no difference between the 14 students
Group 2	-1.14		
Group 3	-1.43		
Group 4	-1.77		
Group 5	-2.67		

Test F was designed to have very difficult items and very easy items (with some items near the top of the scale and some items near the bottom of the scale). The majority of students achieved the same score of 2. Although they are clearly different on the overall test, the results of Test F imply that most are identical in achievement as shown in Table 4. Most students from the five achievement levels received the *same* score. ***Test F fails to detect real differences between students.***

Table 4 **Comparisons of Apparent and Actual Achievement Levels: 5-Item Test F Compared with Actual Test of 21 Items**
(from Izard, 2005c)

Students (*n*=3 for each group)	Actual Levels of Scaled Scores (logits)	Apparent Levels on Test F	Decision based on Test F
Group 1	0.12	13 of the 15 students are at the same level	There is no difference between the 13 students
Group 2	-0.12		
Group 3	-0.36		
Group 4	-0.61		
Group 5	-0.87		

Tables 2, 3 and 4 show that Tests A, B and F provide misleading results for the students scored on those tests. For the easy test, Test A, the results inform us that 13 of the 15 students are at the same level when we know that there are substantial differences between student groups. Similarly, for the difficult Test B, the results inform us that 14 of the 15 students are at the same level in spite of substantial differences between student groups. Combining easy and difficult tests, as in Test F, does not resolve the issue: the results inform us that 13 of the 15 students are at the same level when they are not. **If the results tell us that students are identical in achievement when they are not, we do *not* have quality tests.**

Test J was designed to have a range of items spread from very difficult to very easy (with some items near the top of the scale, some around the middle of the scale, and some items near the bottom of the scale). The majority of students achieved the scores consistent with their level of achievement on the larger test of 21 items. Students are clearly different on the overall test, and the results of Test J reflect the majority of those differences in achievement as shown in Table 5. This is evidence of a quality test: differences (and similarities) between student achievements are reflected in the results.

Table 5 Comparisons of Apparent and Actual Achievement Levels: 5-Item Test J Compared with Actual Test of 21 Items
(from Izard, 2005c)

Students (*n*=3 for each group)	Actual Levels of Scaled Scores (logits)	Apparent Levels on Test J	Decision based on Test J
Group 1	3.47	All scored 5	16 of the 18 differences between the students matched their actual differences
Group 2	1.76	4, 5, 4	
Group 3	1.13	4, 3, 3	
Group 4	0.12	All scored 2	
Group 5	-0.61	All scored 1	
Group 6	-0.87	All scored 0	

Using this model to explain and predict: One needs to be aware of sampling fluctuations in the way representatives of each group were chosen for Tables 2 to 5. For example, for the 81 students with the highest (not perfect) score, 78 had a score of 5 on Test A, and 3 had a score of 4. Although the three students sampled had a higher probability of scoring 5, some scores of 4 were possible. Similarly, with the next highest score, 73 of the 77 students scored 5 and 4 scored 4. The next two scores split in different ways. For the higher score, 78 of the next 81 scored 5, while 13 scored 4. For the lower score, 57 of the 73 scored 5, 14 scored 4, and 2 scored 3. Similar concerns apply to the other tests. Small samples of items from a larger population need to be sampled with care to ensure that the inferences from the results are well founded.

Solutions to this problem depend upon collection of trial data. Without prior knowledge of the properties of the test items, those constructing assessment strategies cannot know whether they have tests like Tests A to I (with all their faults) or like Test J - capable of providing useful information at all achievement levels. Sample tasks might be collected for an item bank but trial data must be obtained so that those assembling tests from the items can ensure the resulting tests are like Test J rather than Tests A to I. Receiving due credit for responses to tests depends upon the qualities of the tests. The quality of a test to give due credit can be judged by looking at the proportion of questions available to each band of achievement. Quality assessment has not been achieved unless the test is equitable for all candidates.

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. For example, a composite test comprising the items from Tests A, B, C, D and E would have similar properties to Test J but would be more precise because there would be more items in the test. Tests of differing complexity can be used provided there are valid scaled scores and the balance of item difficulties is preserved. 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.

C Example 3 – Item/candidate matrix issues

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.* The following discussion (adapted from Izard, 2004) shows contexts where dependable inferences may be made and contrasts them with contexts where inferences are invalid. (Even though it is possible to generate the analysis, the results have little meaning.)

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 4 has a bullet (•) for each interaction between an item and a candidate.

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

Figure 4 **Item-candidate pairs for 4 items and 7 candidates**
(from Izard, 2004)

A valid 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 penalize the candidate for errors made in the production of the examination papers. In the matrix shown in Figure 5 reasonably accurate estimates for those who received the faulty items (shown with a ?) are possible because the gaps in the evidence are limited.

There is another way that such a pattern could arise. If we consider only candidates 2, 3, 4 and 6 (shown in ***blue***) the pattern is that same as for an examination that offers students a choice of questions and where each student has to attempt 3 out of 4 questions.

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

Figure 5 Incomplete item-candidate pairs for 4 items and 7 candidates
(from Izard, 2004)

A valid 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 6 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.

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

Figure 6 Unconnected item-candidate pairs for 4 items and 7 candidates
(from Izard, 2004)

In effect, the two subsets of students are attempting *different* tests. Items and candidates cannot be put on the same achievement continuum because of the limitations in the data collection design. If the items in a test vary in number and 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. (It is possible to scale scores for candidates attempting questions 1 and 2, use those results in a further analysis for candidates attempting questions 1 and 3 to get estimates of difficulty for question 3, validate these estimates for question 3 in a further analysis for candidates attempting questions 2 and 3, then use those results in further analyses for candidates attempting questions 1 and 4, 2 and 4, and 3 and 4 to get estimates of difficulty for all four question. But most just ignore the problem of unequal item difficulties and total raw scores.) In practice, these means that those who choose the easier items receive a higher score than deserved and those who choose the more difficult items fail to receive due credit for the quality of their work. *Without equity, we cannot have a quality assessment.* In a selection context, candidates choosing the "soft" option of the "easiest" questions (whether by accident, or on purpose but without any evidence based on actual data) will do better than they deserve on the basis of their achievements unless item difficulty is taken into account. Conversely, when there is no adjustment for item difficulty, other candidates will have lower scores than expected from their achievements.

The same difficulty of interpretation occurs when pre-test questions differ from post-test questions in an evaluation of effects of an educational intervention as shown in Figure 7. In effect we are comparing results on two *different* assessments. If the pre-test results vary from the post-test results we do not know whether this is because the items differ in difficulty, or because the students differ as a consequence of the intervention, or both.

			Item				
	1	**2**	**3**	**4**	**5**	**6**	**7**
Candidate							
1 (pre-test)				•	•	•	•
2 (pre-test)				•	•	•	•
etc							
1 (post-test)	•	•	•				
2 (post-test)	•	•	•				
etc							

Figure 7 Unconnected item-candidate pairs for 7 items and 2+ candidates (pre- and post-tests)

If the results do not vary from the initial results, we do not know whether it is due to an inappropriate range of items (ceiling effects or floor effects, as in Example 2 above), un-related tests differing in difficulty, the variation in the number of items, or a lack of progress - implying an unsuccessful intervention).

Assessment using such a research design is inadequate to demonstrate student progress over time. Similarly, assessment of progress during a course at school or university is usually impossible because the tests for successive years are usually unrelated. The assessment strategy for one year and the corresponding strategy for the following year differ: when the "rulers" are un-related and lacking common units we have no valid measure of added value (Izard, 2002a). The same problem occurs when we compare distributions of scores in successive years on a national or regional examination.

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 if there is to be equity between years. 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.

Solutions to these problems depend on knowing the properties of the items, obtained from appropriate trials. For example, in one development aid context, the national examinations for three successive years (denoted A, B and C) were administered (after the event) to the same 700 or so students in a balanced way over several days (one-third A, then B, then C; one-third B, C, A; and one-third C, B, A). (The balanced design was to control possible practice effects.) Without such a design it was impossible to distinguish between two explanations for differences over time: *results in some subjects show a decrease in achievement* or (alternatively) *questions in the same subjects varied in difficulty*. In fact, there was no decrease in achievement: differences were explained by changes in item difficulty. Further examples, where changes over time were documented appropriately using calibrated items to ensure that the respective tests were comparable, are available at university level (Izard, Haines, Crouch, Houston & Neill, 2003), at primary school level (Izard, 2002b) and at both primary and secondary school level (Watson, Kelly, & Izard, 2004) .

D Example 4 – Documenting Learning

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 (Izard, 2004). The traditional ways of reporting scores are *not* in terms of how well the student has satisfied each component of the curriculum. 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). Sometimes 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. Cohort reports tend to be used to justify selective schools but *do not account for the value added by the teaching* within schools.

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, those who develop teaching strategies and materials, and those involved in implementing educational interventions 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. Both the pre-tests and the post-tests must be linked as illustrated in Example 3 above.

Concluding comments

Agencies such as World Bank, UNESCO, the Asian Development Bank, and national foreign aid providers (like AUSAID) use accountability provisions to check that the funds provided are spent as intended, but are unable to show that *learning* has occurred unless this is built into the project from the outset. In my experience, those reviewing development projects often lack *measurement* expertise so do not recognise the threats to validity. Further, when a development aid project has a design incorporating such measures, reviewers do not understand why they are necessary. The basic problem is that the accountability mechanisms of development agencies do not, in general, ask the right questions, such as, "Did the individual students participating in the intervention *change* their achievement level?" Similarly, the various Australian Year 12 examinations offer no evidence that students have *learned*. Much of the public comment about internal and external assessment in State and private schooling in Australia and elsewhere fails to recognise the necessity for measures of learning as distinct from a single snapshot of achievement. This basic issue is a consequence of not recognising that providing evidence of learning requires *at least two* relevant measures of achievement *for the same students*. Because we do not know what *learning* has occurred we cannot judge which schools are best in causing learning to occur, and therefore we are unable to reward those schools if we see that as desirable. (Instead we reward schools for their selection policies.)

Quality assessment gives all candidates due credit for their efforts, facilitates evaluating the effectiveness of learning and teaching, and encourages students to be consistent in evaluating their own work. 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 traditional ways of reporting scores do *not* contribute to informing teachers and students of gaps in knowledge because reports are not in terms of how well the student has satisfied each component of the curriculum. We also lack information on the progress made by students over several year levels, as 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.

Earlier in this paper, it was pointed out that the quality of a test to give due credit can be judged by looking at the proportion of questions available to each band of achievement. A balance of questions has been demonstrated to be better generally at distinguishing between students who differ in achievement level and, conversely, generally not distinguishing between students at the same level. Since quality assessment requires tests to be equitable for all candidates, the assessment techniques described in this paper are required to avoid these problems. The design of sound assessments, the development of improved assessment skills, and methods of describing progress are essential requirements for education system (including examination and accreditation boards), teachers and the students. 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.

The quality of assessments can be improved considerably at little cost – we just have to ask the correct questions. Do the assessment tasks chosen for a test represent the relevant domain or continuum? Does this assessment provide valid evidence of improved performance that allows us to infer that learning has occurred? Assessment must communicate relevant useful information to teachers, assisting them to evaluate what they have added to student knowledge through the teaching and learning process. The assessment will be ineffective unless it indicates the action teachers need to take to make the previous estimate of achievement obsolete. A quality assessment strategy complements quality teaching approaches: we need the former to provide evidence about the latter.

References

Frith, D. & Macintosh, H. (1984). *A Teachers' Guide to Assessment*, Cheltenham: Stanley Thornes,

Haines, C.R. & Izard, J.F. (1994). Assessing mathematical communications about projects and investigations. *Educational Studies in Mathematics, 27*, 373-386

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]

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.F. (1993). Challenges to the improvement of assessment practice. In M. Niss, (Ed.) *Investigations into assessment in mathematics education: An ICMI study,* (pp. 185-194). Dordrecht, The Netherlands: Kluwer Academic Publishers

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. (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). Constraints in giving candidates due credit for their work: Strategies for quality control in assessment. In F. Ventura & G. Grima (Eds.) Contemporary Issues in Educational Assessment. (pp. 15-28). MSIDA MSD 06, Malta: MATSEC Examinations Board, 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. In F. Ventura & G. Grima (Eds.*) Contemporary Issues in Educational Assessment.* (pp. 241-252). MSIDA MSD 06, Malta: MATSEC Examinations Board, University of Malta for the Association of Commonwealth Examinations and Accreditation Bodies.

Izard, J.F. (2002c). Using Assessment Strategies to Inform Student Learning. Paper presented at the AARE Conference in Brisbane, December, 2002. (http://www.aare.edu.au [search code IZA02378]). Brisbane, Qld.: Australian Association for Research in Education.

Izard, J.F. (2004). Best practice in assessment for learning. Paper presented at the Third Conference of the Association of Commonwealth Examinations and Accreditation Bodies on *Redefining the roles of educational assessment,* March 8-12, 2004, Nadi, Fiji: South Pacific Board for Educational Assessment. [http://www.spbea.org.fj/aceab_conference.html]

Izard, J. (2005a). *Overview of Test Construction: Module 6.* Paris: International Institute for Educational Planning (UNESCO).

Izard, J. (2005b). *Trial Testing and Item Analysis in Test Construction: Module 7.* Paris: International Institute for Educational Planning (UNESCO).

Izard, J.F. (2005c). Assessing progress in learning mathematical modelling. Paper presented at the Twelfth International Conference on the Teaching of Mathematical Modelling and Applications (ICTMA12), July 10-12, 2005, City University, London

Izard, J. & Haines, C. (1994). Validating the assessment of projects and investigations. *Acta Didacta Universitatis Comenianae (ADUC) - Mathematics*, Issue 3, 83-96.

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

Watson, J., Kelly, B. and Izard, J. (2004). Student change in understanding of statistical variation after instruction and after two years: An application of Rasch analysis. Refereed paper presented at the AARE Conference in Melbourne, Nov.-Dec. 2004. (http://www.aare.edu.au [search code WAT04867]). Melbourne, Vic.: Australian Association for Research in Education.

Withers, G. (2005). *Item Writing for Tests and Examinations: Module 5.* Paris: International Institute for Educational Planning (UNESCO).