

Testing for Teaching: A longitudinal formative assessment project

John Izard and Peter Jeffery

RMIT University, Melbourne, Australia
john.izard@rmit.edu.au

Professional Resources Services Pty Ltd
peter@professionalresources.com.au

Presented at the joint NZARE-AARE Conference in Auckland, November-December 2003

Paper Code JEF03075

Abstract

This paper is a progress report on an on-going longitudinal, public-private project [PPP] between a whole school staff at Monbulk Primary School, Victoria, Australia and two consultants connected with Professional Resources Services Pty Ltd. The project involved establishing a climate for change to a model of formative assessment practices in the school, assisting teachers to select appropriate instruments, publishing those instruments, collecting and publishing item response model [IRM] data for the instruments, applying them annually and facilitating follow-up teaching using the results. The project began in 1995 and has been subject to two independent reviews by government authorities since then. Quite remarkable learning gains are being demonstrated consistently every year.

Context for the Project

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)

Government schools in Victoria, Australia, are required by the Victorian Education authorities to prepare a charter that describes their intentions to implement the curriculum requirements and are then evaluated on the progress they have made. Schools are required to use evidence from such areas as parent satisfaction questionnaires, student attendance records, centrally-constructed tests administered as an external examination at Grades 3 and 5 (LAP/AIM, for example, see Victorian Curriculum and Assessment Authority, 2000), teacher ratings of student performance, teacher days absent, accountability for government and parent funds, and so on. The review process involves internal school self-assessment and independent external verification. Every three years an external reviewer hired by the central authority visits the school to examine this evidence. The reviewer has to verify that this evidence represents a fair record of the learning of *every* child, and to discuss what action the school proposes to take to

upgrade its charter, address new priorities, and ensure that *all* students learn.

Using assessment to improve learning is widely accepted. The notion of testing to improve learning appears in assessment policies in many levels of education including higher education. But the methodology has significant gaps: the distinction between summative assessment (assessment of learning) and formative assessment (assessment for learning) is not well understood by those who prepare policy statements about assessment and learning. Formative assessment has serious implications for the behaviour of teachers and their students but these issues have not been addressed well in theory or practice (Izard, 1998; Black and Wiliam, 1998a, 1998b). In spite of policy statements about assessment and learning, methods of reporting information are mostly inappropriate for learning purposes. Technical limitations of current school-based assessment were discussed at the AARE Conference in 2002 (Izard, 2002b). Without valid student assessment practices the actual achievements are never compared in a legitimate way with the intentions (Izard, 2002a).

Much of the collection of evidence about student learning by systems is summative in practice even though many claim that it is gathered for formative purposes. The intention is to ensure that teachers work hard to ensure the best results for their students in national or state assessments. But the presentation of the evidence to teachers ignores the information that could be used for formative purposes. Teachers are often not told which items *their* students found difficult and which items were completed successfully. Sometimes an examiner report will state success rates on items for a region or nation but an individual school may differ from the pattern since the same mean can be obtained from different patterns of item success. Teachers in Victorian schools (Australia) using AIM are given a matrix of information that includes each student's successful and unsuccessful responses by item. While this is an improvement, in practice the teachers are only given the component of the curriculum each item addressed rather than the actual item. (If you are not permitted to see the test item then interpreting the student errors on that item is most difficult.) The combined effects of administration costs, the limited number of subjects and topics within subjects being tested and the limited feedback to teachers results in less effective evidence of progress and less effective use of centrally-collected information for formative purposes.

Monbulk Primary School decided that the central authority external examinations at Grades 3 and 5 were not useful for *formative* assessment because the examinations were administered later in the school year, had a significant delay before results were provided (and sometimes a further delay while errors in the results were corrected). The results did not provide the comprehensive detail sought by the teachers. Further, the information was not in a form that they could use easily and teachers at other Grade levels were ignored. The school decided to implement a different system of assessment that would be more useful for teachers and learners, with tests that suited the school curriculum supported by prompt and early analyses and reports (Silis and Izard, 2001).

First Steps for the School

The school 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 and the teaching practice fitted the ‘big picture’ and also wanted information that was obtained from the assessment to have a strong link to student needs. Standardized scores, percentiles and state benchmarks were *not* the objective in this program: improving the school's ability to deliver programs that

matched student needs was their objective. The school sought to implement the concepts of *teacher-owned assessment* and *teaching responses to that assessment*. This was *not* a central authority requirement. Consultants were sought to help the school implement its intentions and there were detailed discussions about how to introduce innovative techniques.

Professional Resources Services Pty Ltd (PRS) was the company chosen to provide the consulting advice. Peter Jeffery and John Izard delivered the initial professional development program at the end of 1995 to prepare for implementation in 1996. The professional development purpose was to make teachers aware of assessment for learning, to ensure that they knew what to look for in selecting assessment strategies that would help them in their teaching and reporting, and to make them aware of as many test instruments as possible, including those developed by local teachers. The consultants collected a wide array of examples of published tests in Reading, Spelling and Mathematics and discussed the key issues in making a choice.

Whole-school Professional Development days were set aside for meeting with the provider (PRS). The project had a realistic time line. The staff learnt through hands-on experience, including sitting some test papers to see what the students would face. Croft's *Test Evaluation Sheet*, (Croft, 1980) was used to consider, choose and eliminate tests. Issues of bias, validity, reliability and objectivity were discussed. Staff then took away test specimen sets for a closer look and selected those they felt would meet the intent of their curriculum. Initially multiple papers were considered particularly in Reading and Spelling before one was selected by the elimination of others by staff.

When the consultants met with the school staff later, a number of tests had been identified. A number of these were from British publishers and lacked Australian data. The tests also needed some adaptation to make them suitable for use in Australia. (For example, pounds and pence had to be altered to dollars and cents.) It was fortunate that PRS was also a test distributor and publisher, as licences had to be sought from the copyright owners to adapt and publish the revised versions in Australia. PRS had to invest substantial funds in producing these tests and manuals for Australian use: without this investment the project could not have proceeded.

The concept of a Public Private Project PPP (Webb and Pulle, 2002) embraces the style of the innovation at Monbulk Primary School and offers some clues to explain why this innovation has continued for long enough to be now designated "longitudinal". The Principal of the school was faced with the problem of inadequate mandated assessment practices which were unlikely to provide the information that he felt was necessary for teachers to facilitate the learning of their charges. He sought out an academic colleague known for entrepreneurial innovations in education and with a track record of assessment advisory contributions to education. After preliminary feasibility discussions he appraised the likely cost implications and initiated a call for funds from his School Council who backed the project to use alternative suppliers and consultants. The publisher also invested time money and expertise in the PPP in the expectation that the publishing investment would be possibly made more worthwhile if other schools in Australia took up some or all of the publications and analysis system. To a limited extent this has transpired. The school was not able to publish the necessary materials due to lack of capital and expertise. A private venture could undertake this with an understanding (a trust relationship rather than contract) that the school would at least purchase all the materials it needed for the project for at least 3 years from the publisher. The publisher also located a consultant measurement expert and other talent such as a graphic designer to make it possible for the measurement statistics and the technical documentation to be created and published. PRS

supplied data analysis and data processing at a subsidised rate so that the initial data work could proceed. The school also contracted directly with the consultants for certain professional development activities and funded replacement teachers for the participants by use of school funds supplemented with School Council grants.

Changing the approach to assessment

The school sought to involve the teacher and to acknowledge the teacher's essential role, keep the accountability measures on an honest footing with teaching staff, provide professional development in the use of formal testing procedures and data interpretation, look at the 'value added' issue and monitor trend data on a school-wide basis. Their approach was called *Testing for Teaching Purposes* (a name jointly agreed by the participants) to emphasise their focus and underline joint ownership of the project. It was recognised that there would be instances of positive and negative teacher behaviour that would impact on the program's intention, but it was considered essential to provide opportunities for improvement and accountability measures to be discussed in relation to student learning needs and curriculum change.

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 which permits legitimate comparisons. Traditional test data (expressed in percentiles or standard scores based on relative position of students) are *not* appropriate to measure achievement progress. It was necessary to change the way in which test data were presented in the adapted tests used in *Testing for Teaching*. Since the interest was in achievements and achieving standards, it was decided to present scores in a standards-referenced format. Scores on a first testing occasion could be compared with scores on a second testing occasion. But familiarity with the test material may be a plausible explanation for any improvement in score, rather than effective teaching. So where possible, tests which had alternate forms were to be chosen. (A recent paper by Izard, Haines, Crouch, Houston, and Neill, N., 2003, reports a similar approach at university level in Britain.)

In 1996, procedures were developed and teachers and school management gained experience. Staff participated in professional development to support this new learning in educational measurement, and the associated consequences of using formative assessment (active involvement of students in their learning, building upon students' experiences and interests, linking theory to practice, and providing opportunities for risk-taking and learning from errors). They were aware that school management and School Council fully supported the program. Many staff underwent a shift in their understanding and attitudes regarding empirical measurement tools.

Assessment instruments

It was essential that teachers were provided with current information about the students with whom they were going to work closely for that year. Accountability issues were important but secondary to teachers knowing the learning needs of *this* group of students at *this* point of time (after the school long vacation and ahead of when teachers had responsibility for their students' learning). The three main test series chosen (Diagnostic Spelling Test by Vincent and Claydon, 1996; Effective Reading Tests by Vincent and de la Mare, 1995; and Mathematics 7 – 11 including some calculator and non-calculator usage (now Mathematics 6 – 14) (Professional Resources Services, 1997, 2001) were described elsewhere (Izard, 2002a).

To make a real comparison of achievements from one time to the next, one has to ensure that the scores on either form are equivalent. This can only be determined by empirical data collection and test analysis. The design of the data collection was arranged so that this information could be gathered. For example, the Diagnostic Spelling Test had two forms (Form A and Form B). Initially, some students attempted both Forms. The results were analysed and the two forms were scaled against each other. By this we mean that a score on Form A can be compared with a score on Form B, because the same “ruler” is used for both forms.

If progress relative to the intended skills has been made, then higher scores will be obtained on the scales for that key learning area. There is a possibility that progress will not be made: students may not learn some topics as well as others, and their learning may be affected by health or other influences such as family circumstances. But the key issue is that teachers should know what has been learned and what needs to be learned *so that appropriate teaching action can be taken*. The testing process can confirm or moderate the teacher’s beliefs about the learning stage reached by each student.

Initial implementation

The tests chosen for each grade level were administered by the teachers along with other instruments chosen by the consultants to provide Australian-data comparisons. PRS scored all of the tests, carried out double data entry procedures (to ensure correct entry) and conducted analyses using the Item Response Modelling (IRM) test analysis (See Wright and Stone, 1979) with computer software known as QUEST (Adams and Khoo, 1993). The 1996 results showed that teachers had underestimated the knowledge and skills of their students.

In 1997, testing was scheduled for term one with the aim of providing information to teachers by the close of that term. It was decided to use two tests at each year level where possible. For example, in mathematics, test M7 and test M8 were administered to Grade 3 students as shown in Figure 1 (from Izard, 2002a).

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

Figure 1: Data collection for mathematics (from Izard, 2002a)

The analyses by the consultants commenced with the Year 4 students and calibrated M8 and M9 items on a common scale. After checking item fit, anchor files were created for M8 and M9 items. A second analysis used the M8 items as anchors for the Year 3 data to place the M7 items on the common scale, and create anchor files for M7. A similar procedure with M9 items as anchors for the Year 5 data was used to place the M10 items on the common scale, and create anchor files for M10. These in turn were used with the Year 6 data to place the M11 items on the enlarged common scale, and create anchor files for M11. The effect of these analyses was to place all of the tests on the one “ruler” and to provide standards-referenced benchmarks against which future progress could be gauged.

External Review

Earlier in this paper, it was noted that the review process in Victorian schools involves internal school self-assessment and independent external verification. Every three years an external reviewer hired by the central authority visits the school to examine this evidence. We now look at the events of two of these reviews, the first for the period from 1996 to 1998 and the second for the period 1998 to 2000.

Results for the first triennial review

The school and the outside consultants collaborated in preparing a report for the external reviewer. Using the data collected for the previous three years, analyses of student data were conducted. Two types of comparison can be made. The first comparison involves seeing how Year groups have performed from year to year. Those involved in this comparison are those who attempted the test(s). The second comparison involves seeing how the same group has performed from year to year as they moved through Year levels. Those involved in these comparisons are those who had results each year. (Students who left the cohort are not included because we do not have complete data on their progress.) In this paper only the second comparisons (year to year) are reported.

Progress in Spelling is shown in Table 2. The test form used alternated from year to year. The magnitudes of improvements are expressed as effect sizes in standard deviation units (Cohen, 1969, 1977, 1988) using descriptors provided by Cohen. Table 1 shows these descriptors together with the ranges assigned for this report. In Table 2 and subsequent tables, effect sizes are described as “very small”, “small”, “medium” or “large”. [Cohen (1977, pp. 20-27) 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% nonoverlap. An effect size of 0.2 indicates 14.7% nonoverlap (the component of the combined distribution not shared by the two populations). The corresponding nonoverlap values for effect sizes of 0.5 and 0.8 are 33% and 47.4%.]

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 authors for magnitudes less than “small”

Table 2 Scaled Scores Showing Progress in Spelling 1996 – 1998

	Year 3	Year 4	Year 5	Year 6
Cohort 1				+2.41
Cohort 2			+1.43	+2.05
<i>Improvement</i>				+0.41 sd units <i>small</i>
Cohort 3		+1.11	+1.47	+2.55
<i>Improvement</i>			+0.19 sd units <i>small</i>	+0.65 sd units <i>medium</i>
Cohort 4	-0.25	+1.64	+2.45	
<i>Improvement</i>		+1.18 sd units <i>large</i>	+0.57 sd units <i>medium</i>	
Cohort 5	-0.15	+1.08		
<i>Improvement</i>		+0.93 sd units <i>large</i>		
Cohort 6	-0.16			

Note: 1996 is shown in **bold**.

Progress in Mathematics is shown in Table 3. The comparison involved seeing how the same group has performed from year to year as they moved through Year levels. (Students who left the cohort are not included because we do not have complete data on their progress.) But in the case of Mathematics, *the tests were not parallel*. A series of separate age-level-based Mathematics tests was used. Each Mathematics test measured part of the scale only but overlapped other Mathematics tests (see Figure 1 above). As for Table 2, the improvements are expressed as effect sizes in standard deviation units (Cohen, 1969, 1988).

Table 3 Scaled Scores Showing Progress in Mathematics 1996 – 1998

	Year 3	Year 4	Year 5	Year 6
Cohort 1				+2.05 (M10)
Cohort 2			+1.79 (M9)	+1.49 (M10)#
<i>Improvement</i>				-0.29 sd units <i>negative, small</i>
Cohort 3		+1.76 (M8)	+2.27 (M9) +1.03 (M10)#	+1.98 (M11)
<i>Improvement</i>			0.51 sd units <i>medium</i> -0.74 sd units <i>negative, medium</i>	-0.26 sd units <i>negative, small</i> 0.87 sd units <i>medium</i>
Cohort 4	-0.37 (M7)	+1.57 (M9) +1.58 (M8)	+1.49 (M10) #	
<i>Improvement</i>		1.87 sd units 1.88 sd units <i>large</i>	-0.07 sd units -0.08 sd units <i>negative, very small</i>	
Cohort 5	+1.58 (M8)	+1.90 (M9)		
<i>Improvement</i>		0.31 sd units <i>small</i>		
Cohort 6	+0.00 (M7)			

Notes: 1996 is shown in **bold**. M7 denotes Test M7, M8 denotes Test M8, etc.

Problem with calculators: tested in M10 but not in M9. It had been assumed that students were familiar with calculators when some were not. Action was taken to resolve the problem.

Because the school initiative in assessment was focussed on formative assessment rather than summative assessment and the extra information was not part of the government scheme, the independent reviewer had some difficulty in accepting the conceptual basis for the school's actions. Since this approach allowed for professional development of teachers to enhance focussed-teaching practices to improve student learning, teachers chose to continue it regardless of the reviewer's comments. The focus differed from the official LAP/AIM testing (Victorian Curriculum and Assessment Authority, 2000) in that *Teaching for Testing* tests were administered early in the school year so that teachers had information on the achievements of each individual and on the next topics that are likely to be achievable without distressing students (and parents). *Teaching to the test* was recognised by teachers (as joint planners of the project) as a useless strategy since annual measures of progress were gathered at the beginning of the following school year (after the long vacation).

Results for the second triennial review

Once again the school and the outside consultants collaborated in preparing the report. Using the data collected for the previous three years, analyses of student data were conducted. The same type of comparison is reported as for the first triennium.

Progress in Spelling is shown in Table 4. The test form used alternated from year to year. This test had been designed for ages 7 to 11. When focussed teaching was used as described in the first triennial review report, it was found that students were achieving maximum scores more quickly. It was decided that the test would only be used for students who had not achieved the maximum score. Consequently, the data analysed here exclude *all students who achieved perfect scores* in the three-year period.

The results (means) for those who were tested in all three years from 1998 to 2000 *and who did not achieve perfect scores* are tabulated below.

Table 4 Scaled Scores Showing Progress in Spelling 1998 – 2000

	Year 4	Year 5	Year 6
Cohort A	1.36	2.49	3.34
<i>Improvement</i>		<i>0.79 sd units large</i>	<i>0.59 sd units medium</i>

Progress in Mathematics is shown in Table 5. The results are reported regardless of the test form attempted. Note that the 1998 data for the previous triennial review is not comparable with the data for this triennial review. This is a consequence of transfers to and from the school: only cases with complete data in a triennium are reported.

Table 5 Scaled Scores Showing Progress in Mathematics 1998 – 2000

	Year 2	Year 3	Year 4	Year 5	Year 6
Cohort A			+0.60	+0.88	+1.84
<i>Improvement</i>				<i>0.19 sd units small</i>	<i>0.67 sd units medium</i>
Cohort B		+0.17	+0.69	+1.33	
<i>Improvement</i>			<i>0.36 sd units small</i>	<i>0.49 sd units medium</i>	
Cohort C	-0.70	+0.11	+0.74		
<i>Improvement</i>		<i>0.56 sd units medium</i>	<i>0.44 sd units small</i>		

Note: 1998 results are shown in **bold**.

Teacher Perspectives

Recently, the teachers were invited to comment on *Testing for Teaching* with respect to both benefits and disadvantages for students, teachers, and the school. Teachers in the infant grades considered that main disadvantage for younger children was the early use of formal testing. They felt that some pupils found the written tests intimidating in a group situation and suggested that 1-on-1 testing may give different results. Some warned that some young students may feel inadequate or uncomfortable. Another disadvantage was the withdrawal from other activities while the test was done but the teacher making this comment stated that the time missed was minimal. Teachers considered that a major benefit for students was the opportunity to see what had been retained over the long vacation. Other benefits for students were the identification of individual needs and the pitching of the program at individual needs, and the chance to see and document individual progress over time including reporting to parents about their children. If progress is not made, the matter can be investigated.

Experienced teachers found a need to complement the testing with classroom observation and follow-up assessment as teaching progressed. Some experienced teachers considered that they now had a better idea of what the testing might address, found the earlier feedback of considerable value, and were now hoping for better instruments in some teaching areas than those available currently. There was a view that the tests backed up the teacher's own assessments, although some felt that more emphasis was placed on the published tests by the school community when progress was evaluated. Some teachers considered that it was important not to duplicate assessment effort. Teachers with less experience needed a period of adjustment before being comfortable with this school's different approach. The benefits included a focus on areas of strength and of need, assistance in planning their teaching towards achieving curriculum intentions and their own professional development, and seeing each child's progress over several years and sharing that with the parents.

Teacher perceptions of the benefits for the school included the fact that the tests had been chosen by them rather than imposed on them, the improved planning for learning that targeted individual needs more accurately, the monitoring of each child's progress from year to year, and the ability to point out each child's progress to parents. One teacher felt that the tests covered the basics and this was consistent with parental interests, while another suggested that it was time to tackle new areas such as creative writing. It was seen as important for the school to keep up with modern assessment trends since this was valuable PR for the school when shared with the parents, school council, other schools and the community. Some identified disadvantages,

such as the additional workload implied by addressing each child's needs and placing heavier demands on staffing and timetabling, and the cost of doing the testing each year. One teacher was unsure of any disadvantages and felt that the feedback was positive.

Conclusion and Future Directions

This paper has shown the benefits of a public-private project with focussed teaching as a consequence of informed assessment. 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 many cases the able students are working at levels higher than their classmates.

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 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." (Monbulk Primary School, 2001)

The work commenced in 1995 continues. The complicated web of understandings and funds commitments between all of the (PPP) participants led to a sense of mutual confidence between the parties so that all felt confident and committed to the project. Joint "ownership" thus encouraged the longevity of the project and adherence to the originally negotiated aims. As the positive outcomes for the children and teachers emerged from the various "investments", all participants were encouraged to continue the project. Other schools are starting to use these ideas too. In sharing with you this use of assessment for formative purposes, we hope that you and others will explore this approach so that schools and universities will be more effective and students can enjoy greater learning opportunities.

References

- Adams, R.J. and Khoo, S.T. (1993). *QUEST: The interactive test analysis system*. Hawthorn, Victoria: Australian Council for Educational Research.
- 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. (1977). *Statistical power analysis for the behavioural sciences*. (Revised Ed.) New York: Academic Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd Ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

- Croft, C. (1980). The Foundations of School Testing. *SET: research information for teachers*. Number two, Item 9.
- 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. (2002a). 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. (2002b). 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).
- Monbulk Primary School. (2001). Verification Report dated 25 October 2001. (Document held in Monbulk Primary School records.)
- Professional Resources Services. (1997, 2001). *Mathematics 7-14*. Melbourne: Professional Resources Services.
- Silis, G. and Izard, J. (2001). Monitoring student progress: To use a published test or teacher developed tests. Paper presented at the Australian Association for Research in Education Annual Conference, Fremantle, Western Australia, December 2001.
- Webb Richard and Pulle Bernard, (2002). Public Private Partnerships: An Introduction. Research Paper No. 1 2002-03, Parliament of Australia, Department of the Parliamentary Library, <http://www.alph.gov.au/library/pubs/rp/2002-03/03RP01.htm> 24/11/2003.
- Wright, B.D, and Stone, M. H. (1979). *Best Test Design: Rasch Measurement*. Chicago: Mesa Press.
- Victorian Curriculum and Assessment Authority (VCAA), (2000). *Achievement Improvement Monitor: AIM Testing 2001 Reporting Guide, English and Mathematics Testing Component, Years 3 and Years 5*. East Melbourne, Victoria: VCAA.
- Vincent, D. and Claydon, J. (1996). *Diagnostic Spelling Test* [Australian Edition] Melbourne: Professional Resources Services.
- Vincent, D. and de la Mare, M. (1995). *Effective Reading Test* Windsor: NFER-Nelson Publishing Company.

Appendix 1.

Brief details of the assessment tools used at Monbulk Primary School

The *Diagnostic Spelling Test* is available in two equivalent forms. The forms are parallel in both difficulty and content and consist of a dictation and seven subtests. The subtests are: homophones, common words, letter strings, nonsense words, dictionary use and self-concept.

Mathematics 6 - 14 [ages] has Australian data as well as UK norms. The Australian publisher [PRS] has created *Profile Graphs* which add easy to use powerful interpretation features for teachers. The earlier level tests are orally administered to young children. Icons are used to advise teachers of curriculum content classifications "at a glance". The tests take about 30 minutes each but are not timed.

The *Effective Reading Tests* are group reading tests with two forms at four levels. Progress Tests give a single achievement score. Skills Tests are for diagnostic assessment.

Further information is available from www.professionalresources.com.au or Professional Resources Services Pty Ltd PO Box 71 Coldstream 3770 Victoria Australia +61 3 59649296.