Extension of Item Response Modelling Data for Mathematics Tests

John Izard and Peter Jeffery

RMIT University and Professional Resources Services

IZA01148

Presented at the AARE Conference in Fremantle, December 2001

as part of the symposium on Testing for Teaching Purposes

Context for this research

Assessment strategies for use by teachers need to provide useful information so that the learning by their students can be enhanced. Currently there are two main types of published tests available for teacher use. Traditional tests generally have been developed with classical item analysis. [For example, see *ACER Test of Employment Entry Mathematics* (Izard, Woff, and Doig, (1992), *CATIM 6/7* (Izard *et al.* 1976) and *CATIM 4/5* (Izard *et al*, 1979)]. Modern tests generally have been developed with variants of item response modelling (sometimes known as item response theory). [For example, see the *Mathematics Competency Test* (Vernon, Miller and Izard, 1995, 1996), and *Mathematics 7-14* (Professional Resources Services, 1997, 2001). Both traditional and modern testing approaches use the same data matrix of students by items as the basis for analysis but the results differ and vary in their interpretation. Many of the analysis assumptions are common to both approaches. For example, a student with a higher score is judged more able than one with a lower score.

A. Traditional published tests

Handbooks and manuals for traditional tests generally include information about the performance of a reference group (or groups) as a basis for comparison.

Reference group comparisons

The performance of the reference group is often referred to as the norm. The performance of an individual or group is compared with the reference group performance using percentile ranks or standard scores. The former give meaning to a score by showing what percentage of the reference group had that score or less. The latter give meaning to a score by indicating how many (reference-group) standard deviation units this score is away from the mean score of the reference group.

Problems where reference groups are not appropriate

If the reference group is not an appropriate group for comparisons, then the percentile ranks and standard scores cannot give meaning to the scores. Some examples follow. They address issues of curriculum, how much time is spent on each learning area, teaching skill, and the opportunity of students to learn.

Curriculum followed: A reference group should follow a comparable curriculum if comparisons are to be meaningful. For example, secondary schools in USA follow a different schedule to Australian schools for the teaching of geometry. Use of such USA reference group data to judge performance of Australian students would be invalid.

Time allocation for school learning areas: If students in the reference group have 7 hours a week for mathematics then students with a different time allocation for mathematics will be advantaged or disadvantage depending on whether there was more or less time. The comparisons would present difficulties in interpretation.

Teachers varying in skills: If the reference group data were collected in large schools in urban areas, the students in these schools would be likely to have had more experienced teachers, better classroom conditions, and more senior school administrators. Students from small rural schools would be likely to have less experienced, younger teachers (probably also responsible for their school's administration) and multi-grade classrooms. The comparisons would present difficulties in interpretation and are likely to be unfair to the smaller schools.

Students commencing at different ages: In some States, students commence at age 5 in a preparatory year while others commence at age 6 in Year 1. Students tested in Year 3 will vary in the opportunity to learn: those who commenced at age 5 will have had one more year of schooling than those who commenced at age 6.

Reference group performance is often at one age or year level only: Traditional test reference group results are often presented at a single level. A Year 5 test will only have a Year 5 reference-group data set. The corresponding Year 6 test will have only a Year 6 reference-group data set. Measuring progress from Year 5 to Year 6 is usually impossible.

Individual performance is interpreted with respect to the group, not to the items: A test score is compared with reference group scores and expressed as a proportion of students. The items that were used to obtain these scores are ignored in the interpretation. Teachers and students may be exhorted to do better without any knowledge of the skills that need to be taught.

Other problems of traditional published tests include:

Scaled scores may have to be calculated: Some traditional published tests require teachers to do calculations to obtain scaled scores. This effort detracts from the opportunities to teach students.

Summary statistics are not useful to teachers: Traditional test results often are expressed as means and standard deviations. The statistics give no information about the next tasks that could be the bases for further learning.

There is an implication that doing better than "average" is good: Those interpreting traditional test results appear to assume that doing better than "average" is good. They ignore the fact that, logically, half of the group is better than "average".

The reference group performance may also be less than acceptable in curriculum attainment terms.

Item-level data are generally ignored and are difficult to interpret: When results are presented in terms of success relative to proportions of a reference group, performance on each item may be neglected. When item-level data are accessible, teachers have to compare reference group success rates with their own class success rates item by item.

Generally interpretation is limited to static total-score comparisons: When testing handbooks are prepared, the data are usually compiled for total scores on the test, and sometimes include sub-scores. Rarely, parallel tests are available to allow a class to attempt more than one test to assess progress.

Sub-scores are usually neglected: If sub-scores are provided, teachers have more information on which to base future teaching.

Errors of measurement are usually ignored: When results are reported, tables often show the scaled score associated with each raw score. All measurement is subject to some error. Invalid conclusions are likely if these measurement errors are not taken into account. If the test is too short, consideration of such errors of measurement may mean that the test is useless (in the sense that any student being retested is likely to obtain any score from zero to the maximum on the next occasion even when no learning has occurred.

Analysis methods are restricted generally to right/wrong (0,1) and multiple-choice types of items (0,1): Test items that require construction often have responses that are less than perfect but are not completely wrong. Students completing such items should receive due credit but traditional tests rarely provide for this. Recognition-type items generally may provide more detailed coverage: a set of similar items may act something like a single partial credit item if the test constructor has been aware of this. *There is a disturbing tendency to limit the assessment tasks to what is easy to assess rather than to what is important to assess.* Further, there is considerable emphasis on what students can do alone in a pencil-and-paper mode to the exclusion of qualities associated with working co-operatively, sharing out tasks as a more efficient way of tackling problems, and being able to accomplish integrated and complex practical tasks.

Comparisons are invalid if any items no longer fit the current curriculum: When there are curriculum changes, some items remain appropriate while others become irrelevant. But the reference group data are collected on all of the items. Without the raw data for the reference group, there is no possibility of reconstituting the reference group performance to suit the new curriculum. Without reconstitution, the comparisons are not valid for the new curriculum.

B. Examples of traditional tests

Three examples are provided. The first example (from Izard, 1998) illustrates how performance relative to a reference group provides meaning to the scores. The second and third examples illustrate how traditional published tests have been modified in an attempt to provide teacher-friendly information.

Test of Employment Entry Mathematics

The ACER *Test of Employment Entry Mathematics* (Izard, Woff and Doig, 1992) presents a table (Table 2, p.11) which shows raw scores and their associated percentile ranks, stanines, T-scores and descriptive ratings. The advice provided to test users is based on a reference group of 3267 apprenticeship applicants for both public and private sector employers, from data obtained from testing in 1988, 1989 and 1990. From the table, a raw score of 25 with an associated percentile rank of 81can be interpreted as follows. Any student who obtains 25 correct answers on that test is as good as or better than 81 per cent of the 3267 apprenticeship applicants. Similarly a candidate's raw score of 15 with an associated percentile rank of 16 can be interpreted as saying that this candidate is as good as or better than 16 per cent of the 3267 apprenticeship applicants. If the test user has considerable experience of apprenticeship applicants, meaning can be gained from the comparison of the user's candidate and performance of the reference group.

Such information does not give help to the test user in curriculum terms: the student ranking does not say what the student probably knows, what progress has been made in achieving curriculum intentions, nor whether there are gaps in knowledge or skill which may impede further progress. The tasks that give rise to the score are ignored in the interpretation, in favour of a comparison with a group that is not necessarily known to the user and which may not even be studying a similar curriculum. Presenting the comparison group results as stanine scores, descriptive ratings or T-scores instead of as percentile ranks does not improve the position. Knowing that a score of 25 is "Above average" and that a score of 15 is "Below Average" offers little useful information unless the test user has a considerable experience of apprenticeship applicants, the test applied to them, and the apprenticeship performance of applicants at different score levels.

CATIM 6/7 and CATIM 4/5

These tests are good examples of a traditional test in that they were developed using traditional test analysis techniques. The tests differ from the traditional approach by focussing on sub-scores and success rates on items. Students responded on answer sheet strips that were then attached directly to a chart (to avoid the transcription clerical chore). The chart supplied for recording scores allowed for analyses of item success as well as pupil success, although the analyses are not as sophisticated as for modern published tests. The success of CATIM 6/7 at Years 6 and 7 encouraged a similar approach at Years 4 and 5 lower down the school.

C. Modern published tests

Modern tests use an Item Response Modelling approach developed from initial research by a Danish statistician, Georg Rasch (1960). Wright and his colleagues have extended these ideas. [For example, see Wright, B.D. & Stone, M.H. (1979) and Wright, B.D. & Masters, (1982)].

Comparisons with reference group(s)

Comparisons with a reference group are more informed because the relative difficulty of each item may be shown explicitly on a variable map. The performance of an individual or group is compared with the reference group performance using scaled scores that which imply the probability of success on other items. Meaning is given to a score by showing *what items a student could do* and *what items could not be done*. A teacher only has to consider what teaching is required to ensure that each student is successful on the next items.

Reference groups may vary

Modern published tests allow comparisons with multiple reference groups including with the same students on a previous occasion. The teacher can see whether topics that presented difficulty initially have now been mastered. Provided the test publisher has anticipated the need, teachers can observe the progress of a cohort over several years of schooling using different tests.

Problems where reference groups are not appropriate

As in the case of traditional published tests, if the reference group is not an appropriate group for comparisons, then the comparisons will lack meaning. However the presentation in terms of the items rather than in terms of the total-score performance of the group gives a greater opportunity for teachers to check whether the reference group is appropriate or not. Further, since comparisons for the same cohort are facilitated, those using a set of tests over time can make comparisons with prior performance *of the same students*. (It is difficult to argue that previous performance of the *same* students is not an appropriate reference point.)

The advantages of modern published tests include:

Individual performance is interpreted with a scaled score on the test(s): Teachers have a single "ruler" to describe the curriculum continuum, regardless of the test from the set being used (provided that students are successful on some of the items of their test but not all).

There is an implication that improvement in score is good: With a single "ruler" a baseline is established for performance of a class. The aim is to improve the score *for each student.*

Improvement implies changes: Since there is a single "ruler" teachers can follow the progress of a cohort easily. If a cohort fails to make progress, the teacher is aware of the need to improve the learning of those students.

Measuring change requires two or more test administrations: The aim is to improve the score *for each student* by teaching the requirements of the curriculum. (Teaching the test would not make sense because subsequent tests are different.)

Scaled scores are provided automatically: The total score for the test (or sub-test) is all that is required to assess current standing of a student on the curriculum continuum addressed by the collection of tests.

Errors of measurement are usually considered: Scores are provided with a table of associated errors to assist in interpretation or (better) the plotting of scores automatically provides for consideration of error.

Summary statistics and diagrams are useful to teachers: The variable map shows whether the test is appropriate for the group of students, shows which items are difficult relative to particular students' achievement and which are "within reach". In extreme cases, the variable map shows students not well matched to that test either because they know much more or are much lower in achievement.

Can use the concept of wide and narrow tests: Wide range tests may be used to give a broad perspective on student achievement. Narrow tests investigate a smaller group of topics in more detail. With modern published tests the type of test is obvious from the variable map.

Can cope with total-score and sub-score comparisons: Profiles can be shown for sub-scores as well as total scores.

Sub-scores are not usually neglected: Provided that the sub-tests have sufficient items, and the categories are consistent from test to test, progress can be measured with sub-scores as well as with total scores.

Analysis methods are suited to right/wrong (0,1), partial credit (0,1,2 etc), multiple-choice types of items (0,1), rating scales and performance tasks: Modern published tests can combine information from many assessment approaches. *The use of the "new" approaches can give credit for high quality work whether as a product of an individual or of group work.*

Can adjust for items that no longer fit the current curriculum: While some knowledge is required, it is possible for a teacher to delete irrelevant items and reconstitute the reference group data. [Hand work-sheets for this have been available since the early 1980s. For example, see Izard and White, (1981).]

There is investigation of item performance and student performance: Modern published tests allow teachers to judge student performance in terms of item success and to judge item performance in terms of the number of students successful on that item. These judgments take due account of the variation in item difficulty: no longer do we assume all items are equally difficult (when we know they are not).

Examples of modern mathematics tests

Mathematics Competency Test

The *Mathematics Competency Test* (Vernon, Miller & Izard, 1995, 1996) uses an Item Response Modelling approach. The questions on the *Mathematics Competency Test* are sampled from four important areas of mathematics: Using and applying mathematics; Number and algebra; Shape and space; and Handling data. Figure 1 shows the *Mathematics Competency Test* item difficulty by key area and sub-scale score also by key area. For a given scale score, the teacher can identify the types of question on which a particular student is likely to be successful, those which are need consolidation, and those which are not yet likely to be within reach. Figure 1 below shows how achievement profiles can be drawn for key areas and for the total score. In the case of the total score, the diagram can show the band that includes the error associated with that estimate of achievement. A teacher can place student charts from two different occasions next to each other to gauge progress.

Mathematics 7 - 14

The *Mathematics 7 - 14 Tests* (Professional Resources Services, 1997, 2001) use an Item Response Modelling approach. The questions on the *Mathematics 7 - 11 Tests* are sampled from five important areas of mathematics: Understanding Number, Non-Numerical Processes, Computation & Knowledge, Mathematical Interpretation, and Mathematical Application. The later tests (*Mathematics 10 - 11*) have sub-tests with and without a calculator. The questions on the *Mathematics 12 -*

14 Tests are sampled from five important areas of mathematics: Application of Skills, Application of Patterns and Relationships, Computation and Knowledge, Application of Concepts, and Interpretation and Evaluation. These tests also have calculator and non-calculator sections. For a given scale score, the teacher can identify the types of question on which a particular student is likely to be successful, those which are need consolidation, and those which are not yet likely to be within reach. Australian Item Response Modelling data and Profile Sheets are included in Australian Supplements to Teachers Guide. The research to develop these Australian data is discussed in the rest of this paper.

D. Issues to be addressed

In order to achieve better practical assessments, the design of the instruments has to take account of the need to be teacher-friendly. Visual methods of showing progress, detecting students needing help, and identifying the next steps in student learning make teacher planning-decisions easier, avoids complex calculations by the teacher and gives more time for teaching (instead of focussing on clerical chores). By use of the results to improve teaching/learning, more direct support is available for each learner. Teaching to the curriculum (instead of to the test) is encouraged by the use of different tests that vary in their coverage of the curriculum.

Previous developments

The research reported in this paper built on previous research. The previous data collection had been designed so that each student's progress could be tracked over time. Quality data cannot be obtained if data collection and entry do not meet high standards. Previous data collection and data entry procedures met high standards and new proposals would have to continue to meet these standards.

The previous data collection for mathematics involved a series of tests given in an overlapping design, as shown in Figure 2. The M7 code stands for the age 7 mathematics test, adapted under licence with the permission of the publisher in United Kingdom. The M8, M9, M10 and M11 codes have similar meanings.

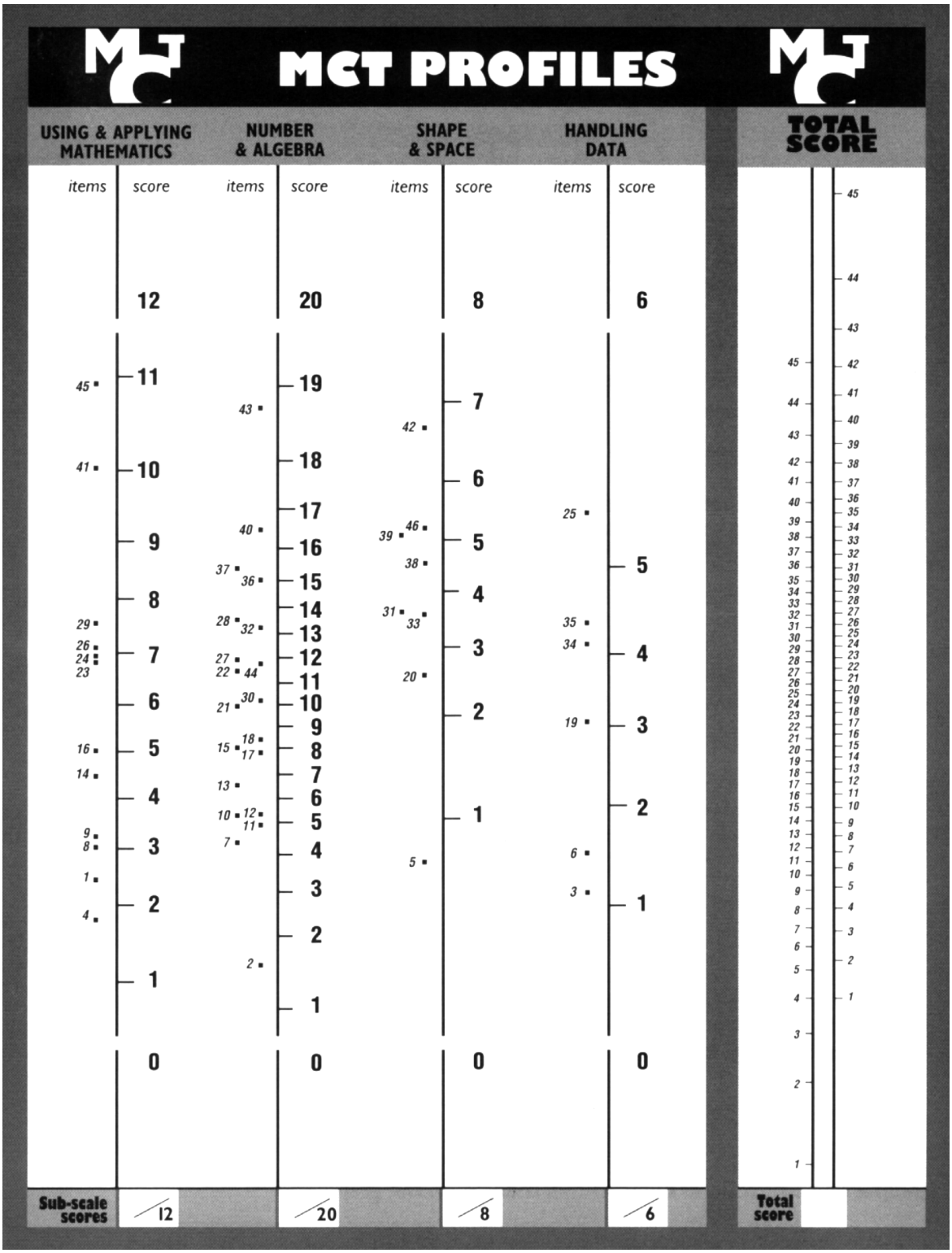

Figure 1 Mathematics Competency Test Profile Chart

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

Figure 2: Previous data collection

The analyses commenced with the Year 4 students and calibrated M8 and M9 items on a common scale. A second analysis used the M8 items as anchors for the Year 3 data to place the M7 items on the common scale. 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 these in turn were used with the Year 6 data to place the M11 items on the enlarged common scale.

E. Design for data collection

After the previous study, the M12, M13 and M14 tests became available and were adapted under licence in a similar way. The extension of the plan to include these new tests had to use a different approach because the categories of item used in M11 did not continue in M12, M13 and M14. It was decided to collect the data for the new tests (including the new categories) alongside responses on the existing M11 test, as shown in Figure 3.

Year Test	M11	M12	M13	M14
Year 7	4	4		
Year 8	4		4	
Year 9	4			4

Figure 3: Second data collection

F. Analyses

The use of standard student codes so cohorts may be tracked was continued. Use of overlapping tests differed from the earlier study but preserved the feature of allowing the new tests to be added to the common scale. Sub-tests for M12, M13 and M14 were based on the new categories for these tests. As before, the double-entry procedure for data was employed again and any discrepancies resolved before the analyses commenced. As in the previous instance, QUEST (Adams and Khoo, 1993) was used for the analyses. The analyses used the anchor values for M11 to place each of the M12, M13 and M14 items on the common scale. Test scales for each new test were prepared in each run with the M11 anchor items. At the same time a

new anchor file was prepared for each new test. Separate scales for calculator and non-calculator items were also prepared using these anchor files.

G. Extending the continuum

In the initial research, logit table values and the associated raw scores (adjusted for perfect score items) were used to create graphs to show progress. Logit table values and the associated raw scores (adjusted for perfect score items) were used to create graphs to show progress as before. These graphs had many advantages. The graphs allowed the raw score on a test to be converted easily to a scale score. The scale score was common to all of the mathematics tests. Teachers could compare scores on different tests. Progress could be measured without being subject to the charge that there was teaching to the test rather than to the curriculum. The scale score graph had raw scores listed on each side of the bar: the left-hand-side indicator for a given score was positioned at (raw score - error) and the right-hand-side indicator for the same score was positioned at (raw score + error). By incorporating errors of measurement in the graph, teachers were less likely to make unsubstantiated claims about changes over time.

Implications for teachers and students

More efficient testing: Time spent testing and in processing results has to be taken from the time taken in teaching and learning. A more efficient testing regime gives richer information with a lower expenditure of time. Same-test-retest situations are always matters for concern because of the possibility that the improved scores are a consequence of increasing familiarity with that same test

Catering for individual differences: The provision of a wider range of tests allows advanced students to gain credit for their knowledge and skill. Weaker students have an opportunity to aim for skills and concepts within their reach rather than being intimidated by increasingly impossible tasks. Most students can be assessed: those who achieve very high scores on one test can be given the next higher test. But the greatest benefit is *not* the improved assessment approach.

Teacher-friendly information: Teachers now have evidence of what students need to be *taught* rather than what position the student has in a group. Measures of progress are easy to interpret by teachers, parents and students. With the extension of the common scale across the transition between primary and secondary schools, teachers now have a resource to ensure that there is no plateau between primary and secondary learning while the new teachers find out what the students know. The system has been action at several schools now for some time. The first school commenced in 1996. With teaching focussed by the assessment, teachers have achieved better results (both in their own view, in the view of school management, and through external system-based evaluation). Parents like the system because they can see what has been achieved and where the next stage will take their child.

The future

Assessment has a long history of being used for external motivation or punishment. The technical advances of modern test development have shifted the focus from comparisons with reference groups to item success. This Item Response Modelling approach provides evidence of what is already known by that student, as a basis for further learning by that student. This is better information for teachers to help students learn more, thereby achieving curriculum intentions sooner and helping build student self-esteem. Item Response Modelling also provides a strategy for gauging the progress made over time

(expressed in curriculum terms) as a consequence of the intervention. These approaches have been illustrated in this paper with mathematics but there has been similar experience with other tests such as spelling (Vincent and Claydon, 1996). One wonders why the earlier traditional approach survives when these more teacher-friendly approaches are available. Perhaps teachers in the future can spend more time teaching successfully and less time being criticised with evidence gained through inappropriate comparisons.

References

Adams, R.J. & Khoo, S.T. (1993). *Quest: The interactive test analysis system.* (Computer software & Manual) Melbourne, Vic.: Australian Council for Educational Research.

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. et al. (1976). *ACER Class Achievement Test in Mathematics CATIM YEAR 6/7*. Hawthorn, Vic.: Australian Council for Educational Research

Izard, J.F. et al. (1979). *ACER Class Achievement Test in Mathematics CATIM YEAR 4/5.* Hawthorn, Vic.: Australian Council for Educational Research

Izard, J.F. & White, J.D. (1982). The use of latent trait models in the development and analysis of classroom tests. In D. Spearritt (Ed.) *The Improvement of Measurement in Education and Psychology*.Hawthorn, Vic.: Australian Council for Educational Research

Izard, J.F., Woff, I. & Doig, B.A. (1992). *ACER Test of Employment Entry Mathematics*. Hawthorn, Vic.: Australian Council for Educational Research.

Professional Resources Services. (1997, 2001) *Mathematics 7-14.* Melbourne: Professional Resources Services.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Vernon, P.E., Miller, K.M. & Izard, J.F. (1995). *Mathematics Competency Test*. London: Hodder & Stoughton. [includes items with pounds & pence]

Vernon, P.E., Miller, K.M. & Izard, J.F. (1996). *Mathematics Competency Test*. Melbourne, Vic.: Australian Council for Educational Research. [includes items with dollars & cents]

Vincent, D. & Claydon, J. (1996). *Diagnostic Spelling Test* [Australian Edition] Melbourne: PRS

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.