Review of Queensland Literacy and Numeracy Testing Programs 1995 - 1999

Prepared for the Queensland School Curriculum Council by Professor Sam Ball, Professorial Fellow, Department of Mathematics and Science, The University of Melbourne.

June 2000

QUEENSLAND SCHOOL CURRICULUM COUNCIL

Acknowledgments

I wish to acknowledge the co-operation and helpful advice provided by Mr. Peter Burroughs, Mr. Jim Tunstall and Ms. Chris Tom in their leadership capacities in Queensland School Curriculum Council. I particularly want to thank Mr. Christopher Dean for his detailed help in providing a range of materials necessary for this review. He and the Council staff were of inestimable assistance.
Appreciation should be expressed to the many people who gave their time and their opinions to enhance the review process and the ensuing report. They represented, informally, many of the interest groups that are users and audiences for the Council testing programs in literacy and numeracy ranging from the Minister, The Honorable Dean Wells through government, catholic and independent sector authorities, the parent organisations and specialist organisations that conducted the testing processes, on to practising teachers and principals. Their names are presented in Appendix A.
Special appreciation is expressed for generously-timed conversations with Professor Neil Russell and Dr. Glenn Finger of the Gold Coast Campus of Griffith University. Dr. Finger wrote the Issues Paper that formed a basis for this review and his analytic strengths are freely acknowledged.
It is indicative of the open co-operative attitudes of Queensland educators that everyone approached by this reviewer provided strong support for the review.

ISBN 0 7345 2146 4

Level 27 MLC Centre
239 George Street
Brisbane, Queensland, Australia

PO Box 317
Brisbane Albert Street, Q 4002

Reception 07 3237 0794
Fax 07 3237 1285
Email inquiries@qscc.qld.edu.au
Website http://www.qscc.qld.edu.au

EVAL 00009

Contents

(Continued…)

Executive Summary

The purposes of this evaluation and review report were to draw appropriate conclusions about the conduct of state-based literacy and numeracy testing in Queensland 1995-1999 and to suggest future directions.

The tests that were considered were:

- 1995 Queensland Year 6 Test (Department of Education, Queensland)
- 1996 Queensland Year 6 Test (Queensland School Curriculum Office)
- 1997 Queensland Year 5 Test (Queensland School Curriculum Council)
- 1998 Queensland Years 3 and 5 Testing Program (Queensland School Curriculum Council)
- 1999 Queensland Years 3, 5 and 7 Testing Program (Queensland School Curriculum Council)

In 1999 the Testing Program involved testing a stratified random sample of Year 3 students mainly for systemic reporting and a census testing of Year 5 and Year 7 students for reporting to systems, schools and parents.

In 1999, the student report to parents itemised results separately in

Literacy: Reading
Viewing
Spelling
Writing

Numeracy: Number
Measurement and Data
Space.

Schools received a more detailed confidential report in hard copy indicating their students' performance by item, class reports, school reports, and a comprehensive marking guide.
There were nine focus questions for this review and evaluation. They were:

Q1. What are the major purposes for the QSCC testing programs? Who are the major audiences?

Q2. What is the evidence that the testing programs are providing useful information for the intended audiences?

Q3. To what extent do the current tests assess those parts of the intended curriculum that they are designed to assess? Within the constraints of practicability, are there curriculum areas (both within literacy and numeracy and outside these areas) that ought also to be considered by the QSCC for assessment purposes?

Q4. How does the current QSCC Aspects of Literacy and Numeracy Testing Program relate to the National Literacy and Numeracy Benchmarks?

Q5. Which processes (eg. test development, quality assurance, distribution, marking) of the QSCC testing program should be subject to specific detailed review? What are the perceived problems?

Q6. Are the current reporting regimes satisfactory in terms of providing appropriate useable information to specified audiences? How might reporting be improved to facilitate improved resource allocation, teaching and learning?

Q7. What is the current situation with respect to testing of and reporting on students with special needs? Are there changes that QSCC should consider?

Q8. How does the QSCC testing program compare to other States' testing programs? Is there anything to be considered as a result of these comparisons?

Q9. What are the most promising future developments in assessment that QSCC should consider in the short to medium term (say for possible implementation over the next five years)?

The specific approach taken in order to conduct this review included two related methods of data gathering in a rubric best described by Eisner (1991) as the connoisseurship model of program review.

The reviewer read and analysed a large number of documents mainly from QSCC but also from library and other archived sources. As well, the reviewer interviewed a range of participants and audience members involved in the Council testing programs. These ranged from the Education Ministry, the leadership of Council, leaders in Education Queensland and the Catholic and Independent School Sectors, test developers, union leaders, curriculum experts, principals, teachers and parents. Of course it was not assumed or expected that the review would comprehensively assess in census or survey fashion such major groupings as parents or teachers but rather work through leaders of representative groups. Nonetheless visits were also made to a small sample of schools and discussions held with principals, teachers and parents.

The connoisseurship model involved qualitative interpretations based upon current and archival tests, personal and telephone interviews and expert examinations and analyses of test materials. The model is not unlike the processes underlying art critiquing on the one hand and a review of an engineering project or writing of an environmental impact study on the other. It combines the professional knowledge and expertise of the writer with the evidence uncovered in the investigation.

A number of conclusions were reached:

1.(a) The QSCC tests are providing useful information for a variety of purposes and audiences. Further improvements to optimise usefulness can be made (see below) but the current base is strong (Focus Questions 1 and 2).

1.(b) QSCC should consider a communications strategy that indicates the multiple purposes and uses of its tests. This strategic push should target major audiences and clear up, perhaps through a Question and Answer postscript, some misconceptions. For example, one group of educators in middle management thought QSCC marks to a normal curve. (See Focus Questions 1 and 2.)

2. Consideration should be given to moving the testing period to a month or so earlier and perhaps allowing the writing task to precede the other tests so that the overall reporting can occur by the start of term 4. (Focus Questions 2 and 5.)

3.(a) The current QSCC literacy and numeracy tests are assessing those important parts of the key learning areas of English and Mathematics that they set out to assess. (Focus Question 3.)

3.(b) Conversations might be initiated with relevant groups on the matter of including somewhat greater use of teachers and teacher judgement in the testing progress. This could allow a broader scope for the tests but would require teacher co-operation. The extra workload would not be egregious (Focus Question 3.) Extra resourcing in terms of teacher release time might be needed.

3.(c) There is no need or pressure for QSCC to move outside the areas of literacy and numeracy at this point as part of its testing program. (Focus Question 3).

4.(a) The QSCC tests provide the required information on National Benchmarks for national reporting. (Focus Question 4.)

4.(b) The actual national reports provide important information but the process underlying the national reports could be improved. The idea of a core national test to assess against core national benchmarks is sensible, especially if embedded in each State's own testing program. (Focus Question 4.)

5.(a) The development period for each year's testing is not as long as it should be. Consequently there is evidence of very minor technical problems in production and reporting that do not affect overall the test results but take away from the overall quality of the process. (Focus Question 5.)

5.(b) Tests are distributed and returned efficiently and as effectively as might be expected in a state as vast as Queensland. Designating a test officer in each school and attempting to achieve continuity there would provide assurance of continued effectiveness. (Focus Question 5.)

5.(c) Financial savings could be made if multi-year contracts could be entered into with a test developer. (Focus Question 5.)

5.(d) The marking of Writing has improved a great deal over the past two years. However, more might yet be done. The criteria should be statistically reviewed as well as considered by curriculum and marking experts in light of the 1998 and 1999 experiences. The reliability and relative validity of different markers needs to be calculated and, if necessary, corrections need to be made as part of the marking process. (Focus Question 5.)

5.(e) There is too much confining of the test questions to items deemed to be appropriate in curriculum terms for a given year level. The reality of individual differences is that many Year 5 students are still at Year 3 level and many have already reached Year 7 level. Truncating the range of test items prevents clear reporting of the size of the individual differences in a class and can also lead to lower reliabilities of scores at the extremes of the year distribution. (Focus Question 5.)

5.(f) The consistent use of evaluation reports by QSCC for monitoring purposes is commended. (Focus Question 5.)

6.(a) Sectors, schools and parents are the three audiences that seem to be getting most satisfaction and use from QSCC test reporting. The major problem for teachers is the lateness of the reports. (Focus Question 6.)

6.(b) Improved reporting can now be considered by QSCC through its test data banks. The improvements can include:

i individual student tracking;
ii school tracking through its cohort of students moving from Years 3 to 5 to 7;
iii school tracking from calendar year performance to calendar year performance;
iv the use of like-school grouping to allow schools to make fair comparisons versus the use of "value added" analysis to ensure fairness in comparison since input factors should be taken into account. (Focus Question 6.)

6.(c) To allow effective reporting, a system of non-obtrusive student identification will need to be developed. (Focus Question 6.)

6.(d) The scaling of scores on the 300-800 scale seems to be well accepted. There is a need to enhance the behavioural anchoring of major scale points in terms of achievement levels. (Focus Question 6.)

6.(e) Greater use of electronic data transfer and accompanying interrogation programs would enhance reporting. (Focus Question 6, and in greater detail Focus Question 8.)

6.(f) Reports for teacher use should include analyses generated by the computer so that the teacher is not required to carry out initial analytic work. (Focus Questions 6 and 9.)

6.(g) Care should be taken to provide all audiences with warnings against over-interpreting differences in performance that may be within common probability bounds. (Focus Question 6.)

6.(h) Some of these improvements - better reporting, implementing new electronic and computer generated programs and involving teachers in the testing and marking (if desired) - will require extra resources. (Focus Question 6.)

7. Students with special needs seem to be sensitively and appropriately handled by QSCC in its testing programs. (Focus Question 7.)

8.(a) QSCC tests compare well with similar tests from other Australian States. (Focus Question 8)

8.(b) Systematic annual reporting to the Council about the other States' tests could provide a useful data base. (Focus Question 8.)

8.(c) The opportunity for QSCC testing staff to swap places with counterparts from other States for a day or two each year would broaden their experience. (Focus Question 8.)

8.(d) Extra resources and their provision in time to initiate earlier contractual arrangements with test developers would ensure QSCC tests had as much preparation time as tests developed in other states. (Focus Question 8.)

9.(a) The systematic use of information and communications technologies could enhance the role and effectiveness of QSCC testing programs. Examples of such use include:

- On-line professional development programs for teachers in the area of classroom testing and interpreting QSCC test results;
- Electronic transfer of results to schools, along with programs for their analysis;
- Delivery to and transfer back from isolated schools to speed up testing administration;
- Providing teachers with pre-analysed diagnostic information on their use of curriculum and their students' status.(Focus Question 9.)

9.(b) The opportunity to use or further develop computer-based, internet delivered, interactive, tailored testing for students in Years 3, 5 and 7 based on a large item bank is already available. Just how soon primary schools will be sufficiently resourced to take advantage is unclear, but the situation improves dramatically each year. QSCC could begin pilot testing such programs as VSAM so that it is ready to move into the next generation of testing as soon as practicable. (Focus Question 9.)

1 Introduction

1.1 Purposes of the Review

The Queensland School Curriculum Council (QSCC) (often referred to in this review as the Council) decided in 1999 it was timely to review the conduct of state-based literacy and numeracy testing. The timing of the review was seen to be appropriate because:

- The annual testing program had been first administered five years earlier (1995) and a review after five years growth and experience was seen as prudent;
- There had been, during the period 1995-1999, the development of the National Benchmarks Program in Literacy and Numeracy under the aegis of the Ministerial Council of Education, Employment, Training and Youth Affairs (MCEETYA) and the relationship with QSCC testing and national benchmarks needed review;
- There was a feeling in the government, and specifically at the Council, that planning for future curriculum development and literacy and numeracy testing programs would be facilitated by a review of the current program.

This review had the following objectives:

1. To research the QSCC literacy and numeracy testing program and to identify major issues .
2. To collect a small set of data from key stakeholders on the issues arising from the research.
3. To draw suitable conclusions on the conduct of state-based literacy and numeracy testing in Queensland 1995-1999; and to suggest future directions. (QSCC, 04/00: Specification 1, March 2000.)

The first objective was achieved through the work of Dr Glenn Finger and the publication of his Issues Paper. (Finger, December 1999.) This left Objectives 2 and 3 (above) as the targets for this second phase of the review process.

Finger's Issues Paper highlighted eight issues:

- Defining literacy and numeracy;
- National literacy and numeracy benchmarks;
- Differences between testing and assessment;
- Sample and census testing;
- Benefits and concerns related to state-based testing in literacy and numeracy;
- Beyond accountability: the impact of the testing program on improving school literacy and numeracy;
- Students with special needs;
- Future developments of test materials – the use of interactive computer-based technologies.

QSCC, in its Specification Section, also suggested that this second stage of the review process, embodied in this report, should consider these related aspects:

- Level of commitment to state-based testing in literacy and numeracy;
- Level of support, eg. school administrators, teachers, parents, school authorities;
- Test coverage – curriculum, year levels, interstate common items, new basics;
- Mode of testing – paper and pencil/teacher administered tasks;
- Time of testing – month of administration, test length;
- Test development outsourcing/in-house/purchase existing test;
- Project management;
- Test quality assurance processes – panels/committees, proofing/sign off;
- The test materials;
- Exemptions and special considerations;
- Marking of tests, particularly writing;
- Freedom of Information;
- Relevance of class and school reports for teachers and principals;
- Relevance of students' reports to parents;
- Use of data by schools;
- Costs (QSCC, 04/00: Specification 2, March 2000).

As a product of an analysis of the stipulated purposes, Finger's Issues Paper, the related aspects nominated by QSCC (see immediately above) and conversations with senior staff of QSCC, ten questions were developed. These became the focus questions for this review paper and they will be presented, explicated and enumerated as a structure or framework for this report (see below 1.3).

1.2 QSCC Literacy and Numeracy Testing (Years 3, 5 and 7)

As one of the results of the Wiltshire Report (Queensland Department of Education, 1994) statewide census tests of literacy and numeracy were introduced in Queensland in 1995. In the first instance, Year 6 testing was carried out and then, over the next five years, extensions and developments occurred. As Finger (ibid p.1) points out, these included:

- 1995 Queensland Year 6 Test (Department of Education, Queensland)
- 1996 Queensland Year 6 Test (Queensland School Curriculum Office)
- 1997 Queensland Year 5 Test (Queensland School Curriculum Council)
- 1998 Queensland Years 3 and 5 Testing Program (Queensland School Curriculum Council)
- 1999 Queensland Years 3, 5 and 7 Testing Program (Queensland School Curriculum Council)

In 1999 the Testing Program involved testing a stratified random sample of Year 3 students mainly for systemic reporting and a census testing of Year 5 and Year 7 students for reporting to systems, schools and parents.

In 1999, the student report to parents itemised results separately in

Literacy: Reading
Viewing
Spelling
Writing

Numeracy: Number
Measurement and Data
Space.

Schools received a more detailed confidential report in hard copy indicating their students' performance by item, class reports, school reports, and a guide to reports including a comprehensive marking guide.

1.3 Review Process

There are nine focus questions for this review.

Of the nine substantive focus questions, the first two deal with purposes and audiences for the testing programs including audience response to them, two deal with the relationship of the tests to Queensland's intended curriculum and to the National Literacy and Numeracy Benchmarks, one deals with the mechanics of the processes of QSCC testing, one with reporting purposes and procedures, one with students with special needs, one takes a comparative look (vs other Australian states) and one looks at future developments in testing worthy of QSCC consideration.

A consideration of Invitation to Offer QSCC 04/00 (the genesis of this review), of Dr. Glenn Finger's Issues Paper which preceded this review, the Council's Evaluation Reports and of the QSCC 1999 Report to the Minister for Education *(Statewide Performance of Students in Aspects of Literacy and Numeracy in Queensland)* indicates that the major assessment issues worthy of QSCC consideration are being addressed by the focus questions presented below.

Q1. What are the major purposes for the QSCC testing programs? Who are the major audiences?

Q2. What is the evidence that the testing programs are providing useful information for the intended audiences?

Q3. To what extent do the current tests assess those parts of the intended curriculum that they are designed to assess? Within the constraints of practicability, are there curriculum areas (both within literacy and numeracy and outside these areas) that ought to be considered by the QSCC for assessment purposes?

Q4. How does the current QSCC Aspects of Literacy and Numeracy Testing Program relate to the National Literacy and Numeracy Benchmarks?

Q5. Which processes (eg. test development, quality assurance processes, marking, exemptions) of the current QSCC testing programs should be subject to specific detailed review? What are the perceived problems underlying the suggested detailed review?

Q6. Are the current reporting regimes satisfactory in terms of providing appropriate useable information to specified audiences? How might reporting be improved to facilitate improved resource allocation, teaching and learning?

Q7. What is the current situation with respect to testing of and reporting on students with special needs? Are there changes that QSCC should consider?

Q8. How does the QSCC testing program compare to other States' testing programs? Is there anything to be considered as a result of these comparisons?

Q9. What are the most promising future developments in assessment that QSCC should consider in the short to medium term (say for possible implementation over the next five years)?

1.4 Review Approach

The specific approach taken in order to conduct this review included two related methods of data gathering in a rubric best described by Eisner (1991) as the connoisseurship model of program review.

The reviewer read and analysed a large number of documents mainly from QSCC but also from library and other archived sources. As well, the reviewer interviewed a range of participants and audience members involved in the QSCC testing programs. These ranged from the Education Ministry, the leadership of Council, leaders in Education Queensland and the Catholic and Independent School Sectors, test developers, union leaders, curriculum experts, principals, teachers and parents. Of course it was not assumed or expected that the review would comprehensively assess in census or survey fashion such major groupings as parents or teachers but rather work through leaders of representative groups. Nonetheless visits were also made to a small sample of schools in each sector and discussions held with principals, teachers and parents.

The connoisseurship model involved qualitative interpretations based upon current and archival tests, personal and telephone interviews and expert examinations and analyses of test materials. The model is not unlike the processes underlying art critiquing on the one hand and a review of an engineering project or writing of an environmental impact study, on the other. It combines the professional knowledge and expertise of the writer with the evidence uncovered in the investigation.

It is proper (though unfortunately rare) to acknowledge the background and values of the author of this review because this affects the review.

The author obtained his PhD at the University of Iowa majoring in educational psychology and measurement. He became a professor in psychology and education at Columbia University working with such colleagues as Thorndike, Hagen and MacGinitie (experts in educational measurement). He became Senior Research psychologist at Educational Testing Service, Princeton, New Jersey (1968-1978) before returning to the University of Sydney. He was a foundation member of the NSW Board of Studies and its deputy president before moving to Victoria in 1993 as CEO of the Victorian Board of Studies.

He is the author of many books and journal articles. His books include *Encyclopedia of Educational Evaluation* and *The Profession and Practice of Program Evaluation*.

The values espoused that are particularly relevant to this report are:

- That assessment is a vital component of the educational process, informing decision-making with respect to goals and objectives, pedagogy and curriculum;
- That testing is a useful component of assessment;
- That occasionally more formal kinds of assessment, such as statewide tests, can add useful information too;

- That some of the information from statewide testing cannot be provided by classroom testing developed at the school level;
- That statewide testing can be ineffective if the test development and related procedures are not professionally carried out;
- That if statewide testing is professionally carried out, it can provide useful information to a variety of audiences.

Inevitably the following report is influenced by the author's experiences and values.

2 Focus Question 1 – Purposes and Audiences

Focus Question 1
What are the major purposes for the QSCC testing programs? Who are the major audiences?

2.1 Purposes and Audiences

Testing programs must be reviewed against a backdrop of the reasons why the testing is taking place (purposes) and for whom the results are intended (audiences). The two are interrelated because, unless one completely rejects the philosophy of pragmatism, the major purposes of testing are realised through the audiences that use the test results. Testing is a means, or a feedback tool, in the educational process. It is not an end in itself.

This is at least implicitly understood by educators. In the interviews, many respondents replied to Focus Question 1 by first discussing the audiences for the QSCC testing programs before tracking back to purposes via the route of what various audiences obtain out of the test results.
The responses can be summarised in tabular form showing the interaction of purposes and audiences (see Table 1).

Table 1	***Kinds of evaluative feedback***	
Audiences	*1.Summative/Accountability*	*2.Formative/Developmental*
A. Sectors	Ensuring systems are effective. Providing baseline data to assess new programs. To use in order to participate in national benchmarking	To help provide information for allocation of resources to schools with special problems and teachers needing specific professional development. To help schools assess their effectiveness with "like school" information.
B. Schools/Principals	For use in school annual reports to sector & community. To assess overall effectiveness	To help determine if there are curriculum strands which need greater emphasis? Are there sufficient remedial and/or gifted and talented programs in place.
C. Teachers	To validate own judgement on whether implementing the curriculum effectively. Is my class achieving literacy & numeracy fundamentals.	To provide diagnostic information. Are there students who need more challenge? special remediation? Are there students who have particular gaps in literacy & numeracy. Is my class showing weakness in a particular area or strand?
D. Parents	To indicate overall performance of their child in literacy & numeracy in terms of intended curriculum and the child's state cohort.	To provide (with teacher guidance) areas needing parent help (remedial or higher level challenge).

Table 1 takes into account the point made by a number of respondents that the purposes can be subdivided into the provision of accountability (summative) information and the provision of formative and diagnostic information that will more directly enhance the teaching/learning process.

Audience A (Sectors), is simply a code word for the Minister, ministry, government and non-government educational agencies and the community at large. Of course in the Westminster system the Minister is the person who embodies most of these groupings. The Minister represents the community at large and is responsible for the overall functioning of education.

There was a tendency for those who were ideologically critical of statewide testing to look for the one purpose of testing and then to argue that statewide testing was an inefficient means of achieving that one purpose. For example, one respondent argued that " *the* purpose of testing is to help teaching". But, the respondent continued, "if you spent the money allocated to statewide testing on the professional development of teachers you could more effectively improve teaching". Whether this claim is true or not, it ignores the other purposes and audiences for statewide testing presented in Table 1.

The vast majority of respondents argued for a variety of purposes and audiences, most of which are detailed in Table 1. However, there are some other purposes not presented in Table 1 that were seen as worthy of mention by some respondents:

- To provide the Council itself information that will prove useful as outcomes-oriented curriculum frameworks are developed in the key learning areas of English and Mathematics;
- To provide for the Ministry information needed in dealing with the Commonwealth Government on funding matters;
- To focus public attention (the community at large) on important educational issues of teaching, learning, accountability and resource allocation.

Some respondents regarded students themselves as an audience for the testing program. Certainly students are the subjects of the testing program and they provide the raw data through their performance. Certainly too, the improvement of their learning is an important target of the testing program through changes that might occur to teaching, curriculum and resource allocation.

Polemics aside, a strong case can be made that there are at least four major audiences for Council literacy and numeracy testing and the purposes, which include both summative (accountability) objectives and formative (developmental) objectives, are potentially capable of being achieved.

2.2 Conclusions

1. The QSCC tests are providing useful information for a variety of purposes and audiences. Further improvements to optimise usefulness can be made (see below) but the current base is strong (Focus Questions 1 and 2).
2. QSCC should consider a communications strategy that indicates the multiple purposes and uses of its tests. This strategic push should target major audiences and clear up, perhaps through a Question and Answer postscript, some misconceptions. For example, one group of educators in middle management thought QSCC marks to a normal curve. (Focus Questions 1 and 2.)

3 Focus Question 2 – Evidence of Useful Information

Focus Question 2
What is the evidence that the testing programs are providing useful information for the intended audiences?

3.1 Evidence of Useful Information

In every Australian State and Territory the advent of primary school testing programs has created a level of polemics where, initially at least, political considerations have vied with educational considerations in the ensuing debate.

Part of the problem has been that test results in some parts of the world in earlier times (especially in the first half of the 20th century) were misused. Inputs were ignored and outputs were misinterpreted. In the period from 1950 to 1985, the mental health movement and interpretations of the progressive education movement meant that there was a distaste for formal evaluative feedback for students and a fear, still expressed by respondents during this review, that testing would be used for "school bashing" and "teacher bashing". Appendix B presents the kinds of criticisms that are usually made of statewide testing programs when they are introduced and a response that might be made by proponents. It can be seen that poorly devised and poorly implemented testing can create legitimate concerns but that these concerns do not necessarily need to become reality.

Specifically in terms of the first five years of the Council testing program, it can be asserted with some confidence that despite misgivings among some teachers the worst of the fears set out in Appendix B did not eventuate.

On the positive side, respondents reported that:

- The Government and sector authorities received information that was used for resource allocation and to satisfy Federal Government funding requirements in the areas of literacy and numeracy.
- Schools used the results as part of their reporting processes, especially in their annual reporting;
- Parents appreciated the reports they received.

Some felt that the parents overvalued the Council reports relative to the schools reporting on students. Nonetheless, there was general agreement parents appreciated getting the two kinds of reports (QSCC and the school). See Cuttance and Stokes (2000) for a more general confirmation of this point.

In terms of the quality of the information, there were some concerns that will be followed up in later sections of this review. Specifically, there was consensus that two matters required further attention:

- The information arrives too late in some instances to be useful, especially at the classroom level. For example, Year 7 test results were known to arrive in some schools in the last week of fourth term.
- Teachers (relative to sectors, schools and parents) seemed to find themselves using the information less optimally than Council would wish. This review will return to this concern in the discussions of Focus Questions 5 and 6.

Under Focus Question 5, suggestions will be presented on the problems of lateness of delivery of information to schools. Under Focus Question 6, suggestions will be provided on how to engage greater teacher use of the Council test results and how other audiences might be provided enhanced reporting.

A second and important source of evidence about the quality and usefulness of the Council test results is provided by the program evaluation work conducted each year by the Council.

Most recently, a survey of principals, teachers and parents/caregivers was conducted (QSCC, June 2000). The survey concluded that the test materials overall were seen as effective with only a small percentage of schools indicating concerns about specific items. The reports were seen as appropriate, and schools saw good uses in terms of student diagnosis and school curriculum program development (with the proviso that late distribution of results curtailed this use).

It is noteworthy that this general survey's results and the growing acceptance and use of the test results are consistent with the findings that came from the review presented in this report. (See especially focus questions 5 and 6).

3.2 Conclusion

Consideration should be given to moving the testing period to a month or so earlier and perhaps allowing the writing task to precede the other tests so that the overall reporting can occur by the start of term 4. (Focus Questions 2 and 5.)

4 Focus Question 3 – Test Items and the Curriculum

Focus Question 3
To what extent do the current tests assess those parts of the intended curriculum that they are designed to assess? Within the constraints of practicability, are there curriculum areas (both within literacy and numeracy and outside these areas) that ought also to be considered by QSCC for assessment purposes?

4.1 Test Items and the Curriculum

The first part of this focus question addresses whether the test items are curriculum valid – whether they target those aspects of the curriculum they are intended to target. The answer is demonstrably that they do. The test specifications and the subsequent test development ensure that there is a close relationship.

There are three caveats that should be attached to this conclusion. The caveats do not deny the conclusion but they do set it in a context that has implications for future considerations.

- The current aspects of literacy and numeracy that are tested are seen as relatively narrow in terms of the total Queensland curriculum. However, it would be costly of funds and teacher workload to extend the current test domains to include, for example, media literacy, oracy and science literacy. The current test domains are basic and vital for future student success. It is hard to imagine students progressing through secondary education if they cannot decode print text and if they cannot comprehend what they have decoded. It is hard to imagine, similarly, students taking well to higher levels of mathematics if they are struggling with basic functions such as addition and division and if they are unable to solve basic quantitative problems such as those featured in the Council tests of aspects of numeracy. In psychometric terms, there are not many separate factors operating in the realms of literacy and numeracy. Indeed, as pointed out by Keeves (1998), most of the variance can be attributed to a global literacy factor and a global numeracy factor and even those two are correlated. The need to assess various strands (say number, measurement and space) is more as curriculum pointers for teachers. They are not separate factors from a measurement viewpoint. In summary, the perceived narrowness of the testing domain does not deny the importance of the domain that is tested. Also, while broadening the testing domain in literacy and numeracy is possible, it could also create a burden on the budget and on teacher workload. This extra burden is not cost effective from a measurement viewpoint but would please several curriculum groups.

- A second caveat relates to the current Council curriculum in English and Mathematics which, all seem to agree, are outdated. In fact, the curriculums are already in the process of being revised. As part of this revision, they are being developed in terms of outcomes rather than processes. This makes subsequent assessment development a clearer procedure since the relationship between assessment task and curriculum outcome is more direct.

- The third caveat is that there is a “Catch 22” situation related to the perceived narrowness of the current test formats. If they were broadened to include more complex assessment tasks, it would both lengthen the testing time and create an increased workload for teachers (which, by and large, teachers do not want to incur). However, to leave the testing in its current format risks the criticism by teachers that the tests call for too narrow a range of student responses.

It would seem appropriate for Council to enter an informed and constructive dialogue in the first instance with teachers and parents to see if some variation of current formats is warranted. In the meantime, the current formats seem quite adequate. (See also Focus Question 8.)

The second part of the focus question dealt with extending the range of subject matter beyond the current realms of literacy and numeracy. Other areas that could be considered include citizenship (civics), science, information technology and fitness (Health and Physical Education).

There was no major support for Council becoming involved, at this point, in such an extension. While the subject areas are regarded as important, there is no current support for their statewide testing in Queensland. If testing were to occur, it would not be focussed particularly on giving parents information, but rather providing sectors and the Government with data for monitoring purposes. A national taskforce is beginning to consider such monitoring activities which might later be put to QSCC, but is not under current consideration and certainly not being considered as a series of census tests.

4.2 Test Items and the Future Curriculum

While it is not explicit in Focus Question 3, it can be seen that results of student performance on the Council test items are an important data source when considering future curriculum development. There is sometimes a gap between expert perceptions of what students ought to be able to achieve and what they actually can achieve (or even have already achieved). This is especially a problem with students at either end of the achievement range.

The current test items and the results students achieve are especially important when QSCC develops outcomes-based curriculum.

4.3 Conclusions

1. The current QSCC literacy and numeracy tests are assessing those important parts of the key learning areas of English and Mathematics that they set out to assess.
2. Conversations might be initiated with relevant groups on the matter of including somewhat greater use of teachers and teacher judgement in the testing progress. This could allow a broader scope for the tests but would require teacher co-operation. The extra workload would not be egregious Extra resourcing in terms of teacher release time might be needed.
3. There is no need or pressure for QSCC to move outside the areas of literacy and numeracy at this point as part of its testing program.

5 Focus Question 4 – QSCC Tests and National Benchmarks

Focus Question 4
How does the current QSCC Aspects of Literacy and Numeracy Testing Program relate to the National Literacy and Numeracy Benchmarks?

5.1 QSCC Tests and the National Benchmarks

The National Benchmarks for Literacy and Numeracy were developed by the Curriculum Corporation during the waning years of the twentieth century. It was a laborious and lengthy process but the final documents have been accepted by all of the States and Territories. A preliminary paper based on the Year 3 Reading National Benchmark results for 1999 has just been released (MCEETYA, 2000) and this is the first major reporting against the benchmarks.

There is acceptance that the QSCC aspects of literacy and numeracy tests have a very close relationship with the benchmarks. Indeed, efforts are made to ensure this is so and to link, where possible, with the other States and Territories to reduce major differences in testing styles. The remark was made that Queensland bends its tests to the benchmarks. Certainly there was a sincere desire to co-operate within the constraints of being true to Queensland's own curriculum.

Nonetheless, at the national level, there are clear problems with the current situation where States develop their own tests in their own ways, administer them to various sub-populations (different from State to State) and then attempt to create scaling equivalencies based upon subjective judgements. Appendix C provides a critique of the process and the analyses that emanate from the process.

As Andrew Porter (1991) argues: "If this practice of separate assessments continues, can the results be somehow equated so that results on one can also be stated in terms of results on the other? There are those who place great faith in the ability of statisticians to equate tests, but that faith is largely unjustified. Equating can be done only when tests measure the same thing." (p35)

A senior Queenslander in education made the point that since the benchmarks are common across Australia, why should there not be a common core test? Then States that wished to do so could elaborate with other items that assessed areas deemed important but not covered by the nationally developed benchmark core tests.

This is indeed a suggestion worthy of further exploration. The cost of each State and Territory inventing its own tests of the same benchmarks seems less than efficient. The resulting attempts at equating could be eliminated and, in terms of the arguments presented in Appendix C, that would be an improvement.

The problem here is essentially a political one. There seems no clear educational argument for many different tests of the same benchmarks. The educational issue is simply whether it would be worthwhile for Council to supplement a national test. Until such a national test is decided upon, and as long as territorial arguments prevail, it is necessary for Queensland, through Council, to continue its aspects of literacy and numeracy testing otherwise Federal funding penalties would occur.

At this stage then, it can be argued that the current Council tests are seen to relate closely to the National Benchmarks. There is much less certainty about the order of accuracy of the subsequent equating.

5.2 Conclusions

1. The QSCC tests provide the required information on National Benchmarks for national reporting.
2. The actual national reports provide important information but the process underlying the national reports could be improved. The idea of a core national test to assess against core national benchmarks is sensible, especially if embedded in each State's own testing program.

6 Focus Question 5 – Development, Administration and Marking

Focus Question 5
Which processes (e.g. test development, quality assurance, distribution, marking) of the QSCC testing program should be subject to specific detailed review? What are the perceived problems?

6.1 Test Development, Administration and Marking Processes

In this question respondents were invited to comment on the micro aspects of Council testing looking over the processes of test development, quality assurance, marking and the distribution and return of test materials. Overall the responses were positive. For the most part the tests were well prepared according to almost all of the respondents.

6.1.1 The Tests

The reviewer in Appendix D presents a critique of one test (Aspects of Numeracy, Year 3, #1, 1999) to exemplify that there are some minor criticisms, inevitably, that might be made and that the few, more serious criticisms of the tests are open to rebuttal. Nonetheless it is argued in Appendix D that despite having an experienced organisation carry out the test development and despite having a number of panels of review set up by Council (though the members were not trained in test development) it is quite possible to have a few problems of at least a minor kind in a published test. One way of overcoming this concern is to allow for a longer development and review period (see 6.1.3).

6.1.2 Distribution

On the question of distribution of test materials and subsequent collection there was no evidence of major problems. In a state like Queensland, remote areas will always present problems but Council was praised for the efficiency of its management of the contractors and the consequent relative smoothness of delivery and later receipt of the test materials. Ensuring each school had a designated officer to attend to all matters from the test contractor would be helpful.

6.1.3 Timing

Of major importance is the criticism of the timing of the test administration, the timing of the delivery of the results and the related issue of the amount of time provided for test development. As things currently stand, the test developer has less than a desirable time period to get the tests developed, have Queensland panels review the proposed items, have the items trialled and refined and then put together in a well formatted mock up for printing and for distribution (along with the administration guides, practice tests (etc) which need also to be prepared.) So far a lead time of about eight months is the best obtained by contractors.

The related timing problem is that schools receive their results toward the end of October at the earliest and about the start of December at the worst. Both times are too late in the year to allow the best use of the results. Reasons given by Council for testing in late August need to be carefully reconsidered. The Brisbane Show in early August is an impediment but need not be a brick wall to advancing the tests from late August to late July or early August. One month earlier would bring it closer into line with other states. The writing task could be a few weeks earlier still as it is slower to

mark. The contractors should be capable of a turn-round of six to eight weeks. The results coming back to the schools in time for the final term of schooling should be the aim.

A final point is that the contractor does not get the contract agreement to begin until too late to do the work as effectively as it could be done. The reason for this is obviously related to state budgetary considerations which go beyond the remit of this review. However, the point might be put by Council that considerable saving can accrue if a test developer can trial two or three years test items in one application. Thus, a long-term commitment, if possible, could provide better and less expensive tests. In any case, earlier annual contracting would be valuable.

6.1.4 Marking Writing

Much improvement has been made in the past year of the Writing component of the Aspects of Literacy test. The main problem was there were too many criteria and too many levels per criterion. Curriculum experts tend to want to fine detail the marking criteria arguing that each criterion is important. What needs to be understood is that no matter the curriculum theory, psychometrically the markers do not make these theoretical distinctions and the correlation among most criteria are as high as the reliability of the individual criterion marks can sustain.

In 1999 the number of criteria were reduced from seven criteria – each with seven levels – to three criteria with seven levels. Still the emphasis was on a particular pedagogy (the argument about the pre-eminence of genre) and somewhat less on the writing product. Thus, a piece of well-written text, seen to be of the wrong genre, could expect to get minimal marks. Similarly, having Spelling as one of the three criteria for Writing has validity problems. There are separate tests and a separate mark for Spelling. A mark for Writing should not normally be confounded with performance in other curriculum strands.

It is not the province of this review to provide definitive answers to problems but perhaps to point out the way to possible solutions and to draw attention to the need for special attention by Council. For the sake of criteria review alone, the area of marking of Writing needs attention. However, two other areas should be considered too.

First is the related question of marker reliability. With one marker only for most writing tasks and with no way of later identifying the marker with a particular mark, it is hardly possible to obtain reliability indices necessary to judge the quality of marking. Furthermore, it is not possible, should one wish to do so, to correct statistically should it be shown (inevitably it occurs) that there are differences in “marker harshness”.

The second, more political question, is whether to have classroom teachers mark Writing tasks too. It is clearly useful professional development for teachers to participate in the marking of the Writing tasks. It is clearly also useful in improving the reliability of the Writing scores (two expert observers provide a more reliable measure than one). However, such a move while better valuing teachers’ contribution to statewide assessment would create an increase in workload of about two hours for the Year 3, 5 and 7 teachers, mainly in terms of providing training in the marking system. This then becomes a budgetary problem for Council.

6.1.5 The Truncation of Task Difficulty

Most teachers tend to resent having tests with items that assess curriculum from an earlier year level and they resent even more items that they see as too challenging – that is, that go beyond the

curriculum designated for their class year level. Thus, say for year 5 teachers, items should test year 5 work.

The problem with this view is the reality of marked individual differences which are often insufficiently appreciated. Here are two pieces of data that illustrate this point.

- Only about 25% of Year 6 students are reading at the average level of Year 6 students. More than a third are reading at Year 7 or above levels and more than a third are reading at Year 5 or below levels. The situation fans out further for Year 7 students.
- In terms of Australian benchmarks, if you take a test where 85% of Year 5 students "pass" the Year 5 benchmarks and give that test to Year 3 students, more than 60% of them will also pass the Year 5 benchmark standard.

Some schools with predominantly high-achieving students find the QSCC tests too easy. In such schools almost all Year 5 students, for example, are achieving at levels beyond Year 5.

By attenuating, or truncating the curriculum spread of the tests in order to target the middle group working at Year level, Council maximises the reliability of scores in the middle ranges, which are already high, but minimises the reliability of scores at the two ends of the distribution. This means it is more difficult to decide which students are at risk (working at more than two years below Year level) and which are in need of special challenge being at more than two years above Year level. Such discrimination becomes important if the tests are to be used to emphasise the need in schools for remedial programs and for special challenge programs.

In future test construction, Council might consider resisting those review panel members who argue against items that assess curriculum elements not typically taught at a given Year level. The present tests tend to pander to their viewpoint. In fairness, however, it should be emphasized that there is a reasonable spread of difficulty level in the Council tests. That is, for a given level, there are some very difficult items; but the items refer to that Year's curriculum. As Council moves to outcomes-based curriculums, a spread of items, in terms of Year level, becomes even more desirable.

6.1.6 Review Panels

Review panels are an essential ingredient in the mixture that goes to make tests that are relevant to a state's students. There seems to be no denying that the QSCC review panels work conscientiously in trying to ensure tests that are fair to both sexes and to minority groups and that are curriculum valid (with the caveat mentioned in 6.1.5).

It is almost inevitable that some tension will occur between test constructors and review panels. Such tension does not seem to have become an undue problem in the development of the Aspects of Literacy and Numeracy Testing Program. Nonetheless, some changes to the operation of review panels has taken place in recent years to ensure a constructive spirit continues.

This is an area that demands consistent monitoring as the changes to the operation of review panels are implemented.

As a general principle, the review panels are advisory to QSCC which is the organisation ultimately accountable for its testing program. The panels' work should be front-end loaded. They should consider the initial array of test items and assessment tasks at the stage when many have been developed and a relatively few are later to be chosen. It is at this point that their advisory input is most important. Later, after trialling, the decisions with respect to choice of items and tasks should be based on technical, psychometric considerations.

6.1.7 Summary

These are the six areas (6.1.1 – 6.1.6) this reviewer felt most needed comment in the complex process of test development, administration and marking. It is unfortunately true that over five years of statewide testing there has been each year one or two problems somewhere in the process. Some of these problems are due to the "teething" problems new and complex procedures inevitably produce. All other States have had somewhat similar experiences.

The QSCC staff that manages the testing programs is not large and the temporary nature of the testing program staff does not help. Of course to change the situation involves budgetary matters. Mistakes can be minimised by giving a more realistic timeframe to the process, ensuring the right balance of skills is temporarily brought in to supplement management if and when staff changes occur, and ensuring some consistency is established with the various contractors providing services to the Council testing programs.

As a final point, this review commends the use of evaluation reports on a variety of relevant issues. It is indicative of an understanding of the continuous quality improvement process that Council has commissioned reports on such matters as:

- Inclusivity issues
- School survey to obtain feedback on student problems, the information guide, parent brochures and other related issues concerning test administration and reporting.
- Evaluating the form and nature of the Year 3 test (1998-99)
- Evaluating the 1998 Year 3 and 5 Testing Program via principal and teacher surveys
- Evaluating the 1998 Test Resource Kit
- Reporting on the work of the Rural and Remote Forum

These and other work reported in *Interlink* (a newsletter of QSCC) provide evidence of a professional approach to a complex remit.

6.2 Conclusions

1. The Queensland development period for each year's testing is not as long as it should be. Consequently there is evidence of very minor technical problems in production and reporting that do not affect overall the test results but take away from the overall quality of the process.
2. Tests are distributed and returned efficiently and as effectively as might be expected in a state as vast as Queensland. Designating a test officer in each school and attempting to achieve continuity there would provide assurance of continued effectiveness.
3. Financial savings could be made if multi-year contracts could be entered into with a test developer.
4. The marking of Writing has improved a great deal over the past two years. However, more might yet be done. The criteria should be statistically reviewed as well as considered by curriculum and marking experts in light of the 1998 and 1999 experiences. The reliability and relative validity of different markers needs to be calculated and, if necessary, corrections need to be made as part of the marking process.

5. There is too much confining of the test questions to items deemed to be appropriate in curriculum terms for a given year level. The reality of individual differences is that many Year 5 students are still at Year 3 level and many have already reached Year 7 level. Truncating the range of test items prevents clear reporting of the size of the individual differences in a class and can also lead to lower reliabilities of scores at the extremes of the year distribution.
6. The consistent use of evaluation reports by QSCC for monitoring purposes is commended.

7 Focus Question 6 – Reporting Regimes

Focus Question 6
Are the current reporting regimes satisfactory in terms of providing appropriate useable information to specified audiences? How might reporting be improved to facilitate improved resource allocation, teaching and learning?

7.1 Reporting Regimes

A perusal of the QSCC *Guide to Reports* (QSCC, 1999) indicates clearly enough that considerable thought and energy has gone into the devising of appropriate and useful reports – student reports, class reports and school reports. Sector authorities indicated satisfaction with the sector reports they received, except perhaps for the previously mentioned problem of timeliness. (See 2.5.3 above.) Teachers and principals agreed with a senior education official that the impression is one of satisfaction and of evolutionary improvement over the five-year history of the program if only schools were to obtain their reports earlier.

Parents apparently have considerable satisfaction too. Indeed some teachers seemed to think the Council reports to parents were valued too highly by parents at the expense of the school's own reporting to parents. This "embarrassment of riches" favouring Council parent reports suggest a continued need to explain to parents that the Council reports supplement and complement school reporting. To have the two sets of reports go to parents as a package makes sense.

Some argument was heard that the parent reports were too detailed. Parents are a highly diverse group bound together by a common bond of genetically inspired accountability for the care and nurturing of their young. Their diversity ensures that Council will also stand accused of being too detailed or too simplistic in their reporting to parents (depending on the kind of parents being reported to). This reviewer thinks it better to err on the side of detail, especially since there are excellent graphic presentations of overall progress by strand for a given child.

The second aspect of Focus Question 6 deals with ways to improve reporting. Five positive constructive suggestions will be presented and the answer to Focus Question 6 will end with a specific warning. The suggestions are:

- Now that the testing pattern for Years 3, 5 and 7 is in place, it makes sense to track students to ensure steady progress. Because total scores are more reliable than part scores, the tracking could best be done by Literacy and by Numeracy collapsed across their respective strands. Care will need to be taken to ensure statistical comparability when conducting this tracking and a statistical consultant might need to be engaged. In order to facilitate the procedure, given the mobility rates of students is about ten per cent per year (much more in some locations), it would be wise to provide students with a unique identification code. A combination of date of birth with one or two identifiers will usually allow a data bank to track some 90 per cent of a cohort. Some modeling of existing data needs to be part of the process of establishing an identifying code.

  Communicating the benign educational purpose of this exercise is important and parents who do not wish to have their child's progress monitored could be allowed to opt out by some active indication on their part.

  A similar tracking of school performance is also a possible option to be taken by QSCC in conjunction with sector authorities or individual independent schools. The tracking of school performance should be in terms of value adding taking into account students continuing in that

school from Years 3 to 5 to 7. The use of this information should be to help schools in need, to help identify exemplary schools and to describe the practices of exemplary schools (those that add above average value to student achievement).

- As well as tracking schools against themselves as their students move from Years 3 to 5 to 7, it is also possible to establish "like schools" and to give schools an indication of how they are progressing in comparison to similar (like) schools. In a sense, this will mirror the tracking of schools against themselves but it provides a different kind of comparison. In association with EQ which has already begun work on this useful approach to reporting, further development might be considered. Principals seem not to know the criteria for determining "like schools" and some seem suspicious of the accuracy of the groupings.

- A second improvement would occur if the abstract statistical scaling 300-800 could be further anchored in curriculum and performance realities. One way of doing this will become available when Council upgrades and updates its curriculums to incorporate outcomes. Since this report reviews the testing programs (and not the curriculum) it is beyond its scope to delve deeply into this point. The potential is available as the curriculum revisions occur.

- Reports to schools are generally provided by Council as hard copy in the traditional way. It could be a saving to funds and to forests if the detailed reports were electronically transmitted. Thus, for example, schools could be provided summary information in hard copy and the detailed follow up electronically with simple programs that allow principals and teachers to interrogate their data as they wish.

- Teachers are more likely to use digested, analysed information that has been flagged for them than if they have to run through a total procedure. Therefore, the hard copy that accompanies the electronically delivered data should identify those students and areas of the curriculum that stand out statistically (outliers).

  For example, here are three number questions where your class performed relatively poorly and here are four measurement questions where your class performed relatively well. Are there any curriculum planning implications to be drawn? Or, here are two students who performed well above average but got three simple items wrong. Have they missed something?

  The computer can be programmed to throw up hints to teachers. Professional teachers can then investigate those hints without necessarily ploughing through the whole data field.

As well as these five suggestions, a warning on over-interpretation needs to be kept constantly in the thinking of those who report. Two particular instances come to mind as consistent problems in educational reporting. First is the problem of over-interpreting individual difference scores. If an individual takes a test on two occasions, it is a seduction for teachers to report that the student is improving (or getting worse) on the basis of the difference in performance from time A to time B. It needs to be remembered that the reliability of a difference score is typically much lower than either of the two component scores especially when the two component scores are highly correlated. Council would do well to inform teachers how big the differences should be before they become excited or concerned at the result.

The second illustration occurs when Council is reporting on particular subgroups. The problem is based upon the fact that students' scores seem to fan out as they get older. For example, if Disadvantaged Group X is developing educationally at four-fifths the rate of Advantaged Group Y, then at Year 3 there will be on average about one year separating their achievement levels, but if they continue to develop educationally at the same rate then by Year 8 there will be two years separating

them in achievement levels. This is simply an indication that the level of disadvantage is acting as a constant.

The area of reporting is vital to the proper use of statewide testing. Council has been commended for its evolving improvement to its reporting processes. There is more that can be done and Council is encouraged to continue exercising continuous improvement strategies.

Perhaps the most urgent and important area of concern is the in-school and in-classroom use of the test results. Assuming that schools (and teachers) receive the results earlier (say before the start of term 4) and assuming that at least some analyses are provided with the reports and that the reports include an electronic version with interrogation programs, then teachers will be more likely to use the results for curriculum planning, curriculum delivery and decision-making about differential curriculum programs for clearly delineated different groups of students in the classroom. Thus the results will be used to further individualise and improve instructional programs.

The various points made in 7.1 will of course need to be discussed within Council but a factor that will have to be considered is that of budget. New reporting programs including tracking and extra computer analysis of results cost money. This is a resource issue that cannot be solved solely by Council.

7.2 Conclusions

1. Sectors, schools and parents are the three audiences that seem to be getting most satisfaction and use from QSCC test reporting. The major problem for teachers is the lateness of the reports.
2. Improved reporting can now be considered by QSCC through its test data banks. The improvements can include:
3. individual student tracking;
4. school tracking through its cohort of students moving from Years 3 to 5 to 7;
5. school tracking from calendar year performance to calendar year performance;
6. the use of like-school grouping to allow schools to make fair comparisons versus the use of “value added” analysis to ensure fairness in comparison since input factors should be taken into account.
7. To allow effective reporting, a system of non-obtrusive student identification will need to be developed.
8. The scaling of scores on the 300-800 scale seems to be well accepted. There is a need to enhance the behavioural anchoring of major scale points in terms of achievement levels.
9. Greater use of electronic data transfer and accompanying interrogation programs would enhance reporting. (Focus Question 6, and in greater detail Focus Question 8.)
10. Reports for teacher use should include analyses generated by the computer so that the teacher is not required to carry out initial analytic work. (Focus Questions 6 and 9.)
11. Care should be taken to provide all audiences with warnings against over-interpreting differences in performance that may be within common probability bounds.
12. Some of these improvements – better reporting, implementing new electronic and computer generated programs, and involving teachers in the testing and marking (if desired) – will require extra resources.

8 Focus Question 7 – Students with Special Needs

Focus Question 7
What is the current situation with respect to testing of and reporting on students with special needs? Are there any changes that QSCC should consider?

8.1 Students with Special Needs.

There is always a tension between ensuring students with special needs are treated inclusively (in other words, are not made to feel left out) and not placing unfair demands on those who might have difficulty coping. The area of educational testing is the place where this tension can come to the forefront.

Discussions with key stakeholders as part of this review indicated clearly that Council procedures and the underlying values upon which the procedures are based receive a high level of commendation. There was praise for the sensitivities exhibited in, for example, the special efforts made for the visually challenged. Council, in collaboration with schools, provide “wonderful help” was the way one parent put it.

The procedures involve ensuring that Council has data on record of the special needs students, their kinds and levels of problems, and the decision made with respect to exclusion or inclusion. This information is not collected in some other States.

As well, equity panels have been a feature of Council procedures and over the years these panels have been encouraged to be active across the testing process.

The equity panels look for problems of unnecessarily difficult or obscure verbiage especially in numeracy testing and of inappropriate topics in literacy testing. For example, a text for testing comprehension should not focus on boys’ sports (unfair to girls) on airline travel overseas (unfair to those on low incomes) or on shopping at large shopping malls (unfair to students from isolated communities). The test developers not only listen to the panels that contain expertise from Aboriginal and Torres Strait Island communities, non-English speaking backgrounds and various groups with expertise in children with special needs. They also use trial test results (carried out interstate) to examine whether items are behaving fairly and consistently.

Another example of the Council’s concern for students and families outside the mainstream is its translations of associated reports and materials into any one of seven community languages (on request).

The importance of diligence in this matter is that test results should validly indicate achievement in literacy and numeracy. Group differences, in test results, should be because there are real achievement differences that exist (and that need attention) rather than that the observed differences are due to unfair testing.

An example of the important need for clarity in interpreting test results can be seen in the test results in literacy for boys and for girls. On the Council literacy tests, on average, boys perform less well than girls. The question can legitimately be raised whether this observed difference is due to unfair, inappropriate test items that discriminate against boys. The procedures adopted by Council would suggest the difference is real and not an artefact of bad test construction.

This is confirmed when one looks across Australia (eg NSW basic skills tests, Victoria's LAP test) and when it is realised that in Year 12 end-of-secondary-education testing in English boys on average are at about the 45th percentile rank while girls on average are at about the 55th percentile rank.

The conclusion to be reached is that all education systems need to consider carefully this differential between boys and girls in literacy. Certainly it is a concern of the Queensland government and the point is that having this Council testing program can provide evidence that the problem is being ameliorated when appropriate remedial programs and cultural value shifts occur.

In summary, this sensitive area of students with special needs should continue to be considered by Council as an important part of the total testing landscape. As new technologies develop (new electronic and computerised programs, for example) even better means will be found of ensuring the tests are sensitive to special needs students. If Council continues as it has in this area, there should be no need for concern.

8.2 Conclusion

Students with special needs seem to be sensitively and appropriately handled by QSCC in its testing programs.

9 Focus Question 8 – Interstate Comparisons

Focus Question 8
How does the QSCC testing program compare to the other States' testing programs? Is there anything to be considered as a result of these comparisons?

9.1 Interstate Comparisons

The educational measurement industry in Australia has no more pillars than the four pillar banking industry. Only four organisations have recent experience in test development in terms of statewide testing. (There are others with the potential, but they are untried at this level.)

The most experienced is the Australian Council for Education Research (ACER) which has been involved in testing for some seventy years. ACER currently develops QSCC tests.

The second most experienced is the Educational Testing Centre (ETC) from the University of New South Wales. It was the previous developer of QSCC tests. It currently provides tests for Victoria and Western Australia.

The third major statewide testing organisation is a part of the Education Department of New South Wales and it develops the tests for New South Wales and South Australia. The fourth group is Patrick Griffin's group from the University of Melbourne. It is not currently engaged in statewide testing but it did work for Victoria a few years ago.

With only three organisations actively developing statewide tests in mainland Australia (and with some interchange among those three and the five client States), it is not surprising that the various State testing program have considerable similarity. There are some minor differences in reporting and in some specific areas of testing. For example, Victoria uses more electronic reporting than Queensland and teachers participate in the marking of Writing. New South Wales tests are mainly put together by classroom teachers. South Australia contracts out much of its testing to New South Wales.

Because of the close relationship between providers and clients, there are few matters that Council can learn from other States' testing programs that have not already been considered. Under the other Focus Questions in this review, where appropriate, ideas (like electronic transmission of results – see Focus Question 6) have already been canvassed. Similarly, in the follow up discussion (to Focus Question 9) a few ideas from interstate will also be presented.

In general, QSCC tests are not unlike the tests from other States. If one were to take test items from all the States' literacy tests or numeracy tests, shuffle them and then ask an innocent but technically able observer (from overseas?) to say which goes with which, that observer would have an impossible sorting task. Even items termed Viewing in Queensland might appear in other States but under a different title.

However, the world of testing is dynamic and it would be a good idea if the office of the Council were to provide the Council with a short annual analysis of the previous year's tests from the other States. A little comparison "shopping" would certainly do no harm and could provide benefits over time for Council. Similarly, it would be useful for Council staff to arrange swaps for a few days with counterparts from other States, thereby broadening their experience.

As was mentioned before in conclusion 6 h (above) it is again true that such matters as earlier contractual arrangements with test developers and providing staff with opportunities to study at first hand other state testing programs require extra resource allocations or earlier resource allocations.

9.2 Conclusions

1. QSCC tests compare well with similar tests from other Australian States.
2. Systematic annual reporting to the Council about the other States' tests could provide a useful data base.
3. The opportunity for QSCC testing staff to swap places with counterparts from other States for a day or two each year would broaden their experience.
4. Extra resources and their provision in time to initiate earlier contractual arrangements with test developers would ensure QSCC tests had as much preparation time as tests developed in other states.

10 Future Developments

Focus Question 9
What are the most promising future developments in assessment that QSCC should consider in the short to medium term (say for possible implementation over the next five years)?

10.1 Future Developments

The major developments in the short to medium term that QSCC should be considering are related to the use of information and communication technologies. There have been a number of suggestions of a more general kind that have been put forward in response to Focus Questions 1 to 8. Those, and related suggestions for possible future action by Council, will be presented under Focus Question 10.

The use of information and communication technologies for QSCC testing purposes could include:

- On-line professional development materials on assessment and testing to help teachers in their own classroom testing and in their use of Council results. Presumably, this could be put together under contract to QSCC

- The electronic transfer of results by internet or by CD ROM along with analysis tools to allow principals and teachers to interrogate their data;

- The delivery to and transfer back from isolated schools by net or by fax in order to speed up the administration of the testing program;

- The provision of computer-analysed diagnostic information for teachers on such matters as their class's curriculum strengths and weaknesses indicated by their class results, students with apparent gaps in their performance profile, and students who should be considered for special help, either of a remedial or challenging nature.

Of perhaps greatest interest in the Australian context is the integrated use of central computers, internet, item banks and local area networked personal computers under the program called VSAM developed by the Victorian Board of Studies. VSAM is well described by Finger in his Issues Paper (ibid, pp15-16) as being computer-based, interactive testing using the internet. Its properties include:

- Students are assessed in terms of their own achievement levels so that able students quickly move to challenging questions and less able students are posed easy questions and are not made to suffer questions outside their achievement level;

- Results are available virtually immediately;

- Test items can be wider in scope and more interesting and realistic to the student (e.g. on-screen calculators that can be used by the student to exhibit their understanding of the correct set of procedures and "drag and drop" operations so that the student can group or isolate pictures to show order, similarities and groupings.) Eventually, of course, VSAM can include audio and video stimulus materials;

- The opportunity for teachers to develop their own classroom tests based on the VSAM item bank;

- The opportunity for students to be tested at times outside the state-decreed testing period to check on progress.

This on-demand property has many flexible uses and can be eventually linked to computerised teaching programs.

VSAM and the above dot points on other information and communication technology applications are well within the realm of possible uses within the next five years by QSCC. However, to move into the information age requires not only that schools become better IT equipped – this is happening with speed anyway – it also requires that QSCC develop strategic plans and allocate resources and person-time to begin trials and pilot programs.

10.2 Conclusions

1. The systematic use of information and communications technologies could enhance the role and effectiveness of QSCC testing programs. Examples of such use include:
    a. On-line professional development programs for teachers in the area of classroom testing and interpreting QSCC test results;
    b. Electronic transfer of results to schools, along with programs for their analysis;
    c. Delivery to and transfer back from isolated schools to speed up testing administration;
    d. Providing teachers with pre-analysed diagnostic information on their use of curriculum and their students' status.
2. The opportunity to use or further develop computer-based, internet delivered, interactive, tailored testing for students in Years 3, 5 and 7 based on a large item bank is already available. Just how soon primary schools will be sufficiently resourced to take advantage is unclear, but the situation improves dramatically each year. QSCC could begin pilot testing such programs as VSAM so that it is ready to move into the next generation of testing as soon as practicable.

Appendix A: Those who were interviewed

A few of these interviews were conducted by telephone, but most were carried out person-to-person. Where group interviews occurred, there may be some group members present whose names were not captured. Some of the most senior interviewees were not formally interviewed.

Reg Allen	Deputy Director, Board of Senior Secondary School Studies
Murray Baulch	Principal, Upper Mount Gravatt State School
Damien Brennan	Assistant Director, Religious Education and Curriculum, Brisbane Catholic Education Centre
Peter Burroughs	Chair, QSCC
Michael Byrne	Principal Adviser, Performance, Measurement and Review Branch, Education Queensland
Ian Davis	Project Manager, Hermes Precisa Pty Ltd
Christopher Dean	Principal Project Officer, Quality Assurance, QSCC
Sandra Easey	Principal, Inala State School
Tim Eltham	Senior Policy Adviser, Office of Minister for Education
Garry Everett	Deputy Director, Queensland Catholic Education Commission
Glenn Finger	Lecturer, Education and Professional Studies, Gold Coast Campus, Griffith University
Kaylene Forrester	Visiting Principal, Inala State School
Chris Freeman	Assistant Director (Projects), Educational Testing Centre, University of New South Wales
Deidre Jackson	Manager, Assessment Services, ACER
Dawn Lang	Principal, A.B. Paterson College
Ross Linegar	Principal, Payne Road State School
Gabrielle Matters	Director, New Basics Unit, Education Queensland
Bob McHugh	Assistant Director General, Education Services, Education Queensland
Martin Murphy	Project Officer, ACER, Melbourne
Harry Newman	Assistant Education Officer, Curriculum, Brisbane Catholic Education Centre
Bill Perrin	Manager, P-10 Assessment, Victorian Board of Studies
John Pitman	Director, Board of Senior Secondary School Studies
Patricia Reust	Council Nominee, Federation of Parents and Friends Association, Queensland
Neil Russell	Associate Professor, Education and Professional Studies, Gold Coast Campus, Griffith University
Judy Scotney	Deputy Principal, Inala State School
Robin Thomas	Senior Project Officer (Testing), QSCC
Chris Tom	Assistant Director, QSCC
Julian Toussaint	Queensland Teachers' Union Council Nominee
Jim Tunstall	Director, QSCC
The Hon. Dean Wells	Minister for Education
Tommy Yip	Account Manager, Hermes Precisa Pty Ltd

Appendix B: Initial Criticisms, and Rebuttals, of Statewide Testing

Criticisms	*Rebuttals*
The tests are only a "snapshot", "point in time" or slice of time observations of student performance. This "one off" assessment is an unreliable indicator of student performance.	The tests do provide reliable data consistent over time. Snapshots are useful as indicators of current status. (Note e.g. wedding photos). Teachers must always realise test scores of individual students need interpretation but this does not invalidate them.
The tests are only assessing a narrow range of competencies. This sample of competencies does not tell you the whole picture of academic performance.	The range is relatively narrow but the assessments are of vital areas of competence. Further, the correlation between them and other areas of academic performance is high.
It is only a rough screening device. The tests do not provide fine-grained diagnostic information. They only tell teachers what they already know.	The tests are not intended as fine-grained diagnostic tests to explain, for example, what specifically are a student's problem in adding fractions. Often they affirm a teacher's judgement. So they should. But they also provide information that can provoke thought even in the best teachers. Also this is only one of the many purposes of statewide testing.
They promote competition among students, parents and schools.	There is no evidence of this occurring as a result of statewide testing.
League tables can be established that make unfair comparisons of schools and harm schools in depressed areas.	In Australia, particularly Queensland, league tables have not been published because care is taken to ensure it does not happen.
Statewide testing costs too much. The money could be better spent elsewhere in education.	The cost overall is about $1 out of every $2,000 spent on education, which is marginal. To eliminate this cost would have virtually no impact on class sizes, for example.
The testing takes up too much time which could have been used for teaching.	Over seven years, or about 1,400 days, 3 days (maximum) is used for testing. Assessment is an important element in the teaching–learning process.
The test adds to the workload of teachers. They are already under pressure.	The tests as presently constituted take no more time for teacher preparation than if the teacher were preparing lessons to be given over the same time period.
Students can be traumatised by statewide testing	Across five years of QSCC testing there are no documented reports of trauma among students. Taking a test is a useful experience for students when under the guidance of caring teachers.

Appendix C: Testing, Reporting and the National Benchmarks in Literacy: A Critique

The Australian consortium, the Ministerial Council on Education, Employment, Training and Youth Affairs (MCETYA) has decided that there will be national literacy and numeracy benchmarks and that there will be an annual reporting by each State and Territory of the percentage of students at specified year levels that have failed to reach the benchmark standard.

The current situation is that each State/Territory conducts its own tests. A technical panel considers each test and determines subjectively a cut-point deemed to be so positioned that performance below the cut point is declared to be performance below the benchmark.

A number of technical and political concerns arise out of this process.

- States/Territories test students using a variety of item types and assessment devices. Some contain more open-ended items, some emphasize recognition-type items such as multiple choice. Some clothe numeracy items with a heavy loading in verbiage so that when a student is incorrect, it is not clear whether it is because he or she cannot read or cannot calculate. Some States/Territories intermix easy and hard items so that if a student fails to finish a test, that student will not have successfully answered some easy questions. These are a few of the major differences in testing.

- The students taking the test in a given State/Territory constitute different samples of students. Thus, in South Australia it is mainly government school students who are tested. In Victoria, all government and Catholic schools are involved and about 50% of independent schools also participate. In New South Wales, there has been a mix of all government schools, some Catholic schools and some independent schools. The failure to report like with like in terms of the percentage and kinds of students taking the benchmark tests constitutes a high hurdle to fair interpretation.

- The equating process has an unknown level of unreliability attached to it. When the Technical Equating Committee carried out its first equating exercise, the results were ludicrous. As one of the experts put it “we saw the results and just threw up our hands”. The actual process has not changed, although there have been some obvious attempts to improve things at the micro level. For example, the teachers who are asked to judge the difficulty level of items and tasks now get more training. The equating for the 1999 testing could be accurate but the reliability of the judgements relating to benchmarks and items is either not obtained or if obtained, not made public. Thus, it is not possible to do more than trust that a process that produced flawed results the first time around could produce accurate results given some fine-tuning.

- The actual publication of the results (the percentages of students below the benchmarks by State/Territory) is based upon each authority carrying out its calculations relatively independently.

States seem to have some latitude of at least a few percentage points in determining the actual percentage published. It is not an exact process and the rules are not spelled out nationally to a fully comprehensive degree. The use of confidence intervals is quite appropriate but around which point estimate?

Given these four concerns about the benchmark calculations, it is surprising that there was some opposition to the insertion of some common items in the 1999 tests. Nonetheless, common items were included in four States’ tests.

The results of the relevant States' students on these common items have not been published.

A full accounting should be made.

In the year 2000 testing, despite Queensland's willingness to continue with the insertion of common items across States, there was no agreement to have common items.

Appendix D: Critique of the Test Aspects of Numeracy Year 3, #1, 1999

The following are the minor and in some instances disputable criticisms of the QSCC test named in the heading. They are presented to illustrate the major points in the text of the review (see Focus Question 5) where it is argued regarding the tests themselves:

- There are only a relatively few minor criticisms. Some are trivial and warrant no special action.
- There are no major criticisms that cannot be rebutted.
- Despite being developed by an experienced and reputable organisation and being screened by a number of QSCC panels, there seems to be a few minor points worth tightening.
- Some of the criticisms are dependent on values and purposes for the test held by the reviewer. They are open to an intellectually honest rebuttal.
- While there is no way to be completely assured that slight problems will be eliminated, having sufficient time for test development, trialling and review is essential to minimise problems. At the moment, because of budgetary constraints and timetables, QSCC provides less time for test development than any of its interstate counterparts.
- The vast majority of the items for each of the tests developed for QSCC are reputable and of high quality and have excellent item characteristics from a statistical viewpoint.

With these points in mind here is the critique:

1. Practice Question 3 notes in the stem. “The children in class voted for their favourite fruits: Here is a graph they made. Which is the most popular fruit?” The graph is headed “Favourite Fruit”. This kind of ‘elegant’ variation from one term to another and back to the first is inappropriate even if it were a test of literacy. In a test of numeracy, it is misleading.
2. The instruction for Practice Question 5 tells students “If you don’t need a box, leave it blank”. There follow two examples, neither of which involve a box that should be left blank. A practice question that does not allow a student to practise is not optimal.
3. Of the 33 questions posed in the test, 90 per cent had a verbal component and some were somewhat wordy. This is an area of dispute but perhaps 90 per cent of items embedded in verbiage is over weighted in a numeracy test even if the student is told to ask the teacher if he/she needs help to read a question.
4. Question 6 gives the pattern 6, 12, 18 and then provides the following instruction: “Use your calculator to count on by six (6) and write the next numbers in the counting pattern in the boxes”. Of course, the best students will have no problem getting this correct without the use of a calculator. Indeed being instructed to use it for Question 6, when the top heading for Question 5 to Question 8 says “You may use your calculator for the following questions”, serves only to present unwanted ambiguities to the student.
5. Some questions, and particularly Question 14 and Question 15, could have been taken from an old IQ test (Question 14) or a dated test by Piaget (Question 15) the developmental psychologist. Question 15 is very much dependent upon developmental maturation which is demonstrably difficult for teachers to accelerate through the classroom curriculum. This kind of item makes it a dubious test of classroom teaching and more a test of socio-economic status and general cultural and cognitive development.

6. There is some trickiness to Question 21. “Choose the things that could roll smoothly along a table”. The things are in fact pictorial representations (pictures) of things. A pictured cylinder and ball are the required answers. However, butter (pictured) appropriately cut, ‘could’ roll smoothly. While few students are unlikely to make a pedantic adult’s fine distinction, it remains preferable to word questions in order to avoid unnecessary ambiguity.

7. Question 22 asks: “Which of these is impossible?” There follow four statements each with its illustration. Three of the statements end in the word “today” and each is possible. The required answer does not end in the word “today”.

   There are two problems here: First obvious clues to the right answer should be avoided in the technology of item writing. Second, one might question why this is a numeracy item. It is a logic item expressed solely in words. Verbal comprehension? Literacy?

8. Question 25 is a linguistics question. Psychologists use such questions as this (F is fourth, E is third, M is second, Y is first. Which did E beat?) to assess linguistic and cognitive development (see also above point number 5 re Questions 14 and 15).

9. There is a semantic confusion surrounding Question 30. Pictured are eight carrots, three ice-cream cones, five blocks, seven plants, five apples and four stars. These are interspersed so that hardly any contiguity exists between like objects. The question is: “Draw lines around groups of objects to show enough tens and ones to make 24”. There is no group of ten anything, although you could group five block and five apples together to form one group of ten and you could group seven plants and three ice-cream cones to form a second group. However, it would be virtually impossible to draw lines around these constructed groups. This question has been made unnecessarily complex by the six different pictures. Thirty-two blocks would have sufficed.

10. The words “go to the next page” especially at the bottom of Page 11, would be worthwhile.

11. Question 31 asks for a best estimate and is clothed in verbiage that is “window dressing” and likely, according to item response analysis, to favour boys.

References

Cuttance, P. and Stokes, S.A. (2000) Reporting on School Achievement. DETYA, ACT.

Eisner, E.W. (1991) The Enlightened Eye: Qualitative Inquiry and the Enhancement of Educational Practice. DETYA, ACT

Finger, G. (December, 1999) Queensland Literacy and Numeracy Testing Programs 1995-1999 Issues Paper. QSCC, Brisbane, Queensland.

Keeves, J. (1998) Personal conversations on his research findings based on factor analytic studies of English and Mathematics test results of primary school children. Flinders University, Adelaide.

Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) (2000) 1999 Year 3 Reading National Benchmark Results. Canberra, ACT.

Porter, A.C. (1991) Assessing National Goals: Some Measurement Dilemmas. In T. Wardell (Ed.) The Assessment of National Goals. Proceedings of the 1990 ETS Invitational Conference. ETS, Princeton, NJ.

Queensland Department of Education (1994). Shaping the Future: Review of the Queensland School Curriculum. Education Queensland, Brisbane. (The Wiltshire Report.)

QSCC (March 29, 2000) Invitation to Offer, QSCC, Brisbane, Queensland.

QSCC (1999) The 1999 Queensland Years 5 and 7 Tests Guide to Reports. QSCC, Brisbane, Queensland.

QSCC (2000) Evaluation of 1999 Queensland Years 3, 5 and 7 Testing Program. QSCC, Brisbane, Queensland.