

**Broadening Curriculum Coverage of
Queensland Years 3, 5 and 7 Tests
in Aspects of Literacy and Numeracy**



QUEENSLAND
SCHOOL
CURRICULUM
COUNCIL

June 2002

ISBN 07345 2335 1

© The State of Queensland (The Office of the Queensland School Curriculum Council), 2002
Copyright protects this publication. Except for purposes permitted by the *Copyright Act 1968*,
reproduction by whatever means is prohibited.

*In July 2002, the Queensland School Curriculum Council amalgamated with the Queensland
Board of Senior Secondary School Studies and the Tertiary Entrance Procedures Authority to form
the Queensland Studies Authority.*

Further inquiries about this publication should be directed to:

Queensland Studies Authority
PO Box 307, Spring Hill Q 4004
295 Ann Street, Brisbane
Queensland, Australia

Telephone: (07) 3864 0299
Fax: (07) 3221 2553
Website: www.qsa.qld.edu.au
Email: office@qsa.qld.edu.au

FOREWORD

This is a research report that deserves attention. It is well organised, balanced in its approach and detailed in its investigation. It is a credit to the Queensland School Curriculum Council and to those responsible for it.

Ideally assessment and curriculum work as an integrated force seeking with the help of teachers to achieve our educational objectives. Of course, the ideal is just that — an ideal. In practice assessment and curriculum have been all too separate.

Curriculum has been seen in the past as bodies of knowledge. Too rarely has curriculum been clear about what specific standards (skills and applications) are required of students at given grade levels. Those standards have, in the past, been crudely and incompletely operationalised not by curriculum but by examinations and other assessment devices.

In our emerging era, there are substantial changes towards clearer curriculum standards and assessment that mirrors those standards. English is not only Spelling and Reading Comprehension although they are relatively easy to assess. Mathematics is not just having the right answer to arithmetic skills although that is a worthy objective. We want our students to be able to listen attentively and to communicate orally with style and clarity. Assessment needs to provide a wider coverage of a good curriculum.

This research report is refreshing. It shows what can be achieved in terms of clarifying issues, obtaining relevant information and organising that information coherently to provide answers and, inevitably, construct further questions. It is not coincidence that the final section of this report suggests the need for further research. As the person responsible for the initial review (Queensland School Curriculum Council 2000a) that spawned this paper, I am pleased to have become part of the continuing search for improved assessment in Queensland's schools.

Professor Sam Ball
University of Melbourne
Consultant
Ministry of Education and Youth
Dubai UAE

Contents

FOREWORD	iii
ACKNOWLEDGMENTS.....	vi
EXECUTIVE SUMMARY	vii
1. INTRODUCTION	1
1.1 <i>Purposes of research</i>	1
1.2 <i>Background to the research</i>	1
1.3 <i>Research questions</i>	5
1.4 <i>Research approach</i>	6
2. LITERATURE SEARCH AND REVIEW	7
2.1 <i>Contexts for shifting emphases on large-scale tests</i>	7
2.2 <i>Consistency (Reliability and Validity) Implications</i>	8
2.3 <i>Critical positioning</i>	11
2.4 <i>Conclusion</i>	12
3. PROCESSES TO ENSURE VALIDITY AND RELIABILITY OF ASSESSMENTS	13
3.1 <i>Teacher-Assessed Tasks</i>	13
3.2 <i>Moderation</i>	14
3.3 <i>Calibrated Teacher Judgments</i>	15
3.4 <i>Conclusion</i>	17
4. BROADER LITERACY AND NUMERACY ASPECTS	18
4.1 <i>Broader Curriculum Coverage</i>	18
4.2 <i>Broader Coverage of Literacy</i>	19
4.3 <i>Broader Coverage of Numeracy</i>	20
4.4 <i>Conclusion</i>	22
5. VALUED LITERACY PRACTICES AND MULTILITERACIES	23
5.1 <i>Multiliteracies and Valued Literacy Practices</i>	23
5.2 <i>Cross-curricular numeracy and literacy</i>	24
5.3 <i>Conclusion</i>	25
6. SAMPLE MATERIALS	26
6.1 <i>Teacher-assessed Tasks in Large-scale Testing Programs</i>	26
6.1.1 <i>Indiana Statewide Testing for Educational Progress (ISTEP+)</i>	26
6.1.2 <i>Kentucky Department of Education Commonwealth Accountability Testing System</i>	
(CATS)	28
6.1.3 <i>Maryland School Performance Assessment Program (MSPAP)</i>	30
6.1.4 <i>Vermont Statewide Assessment System</i>	33
6.1.5 <i>Australian Capital Territory Assessment Program 2001</i>	36
6.1.6 <i>New South Wales</i>	37
6.1.7 <i>Northern Territory Multi-level Assessment Program</i>	40
6.1.8 <i>South Australia</i>	42
6.1.9 <i>Victoria Achievement Improvement Monitor (AIM)</i>	42
6.2 <i>Appraisal Conclusions</i>	43
6.3 <i>Conclusion</i>	44
7. CONCLUDING STATEMENTS	46
APPENDIX A – INTERVIEW SCHEDULE A.....	47
APPENDIX B – INTERVIEW SCHEDULE B.....	48
BIBLIOGRAPHY	49

LIST OF TABLES

Table 1: Types of Teacher-assessed Tasks and Examples	3
-------------------------------------------------------	---

LIST OF FIGURES

Figure 1: Teacher-assessed task Mathematics Grade 3	26
Figure 2: Scoring Rubric	27
Figure 3: Scoring Exemplar	27
Figure 4: Parent Guidebook on writing portfolio components and marking criteria	29
Figure 5: Example part of a task for Grade 3 on 'Deserts'	31
Figure 6: Marking Criteria and two examples of scored work	32
Figure 7: Example of a marked problem-solving task from Mathematics Problem Solving and Communication Portfolio	34
Figure 8: Some marking criteria for Mathematics Problem Solving and Communication Portfolio Tasks	35
Figure 9: An example of speaking task administration requirements	36
Figure 10: Student record sheet and criteria	37
Figure 11: Example of the PWA teacher-assessed writing task	38
Figure 12: PWA writing task marking criteria	39
Figure 13: Front page of the NTAP student Year 3 literacy booklet showing marking table for the optional common writing task teacher's assessment	41
Figure 14: Example of AIM teacher-assessed mathematics tasks	43

Acknowledgments

The following officers participated in this research: Mr Christopher Dean (research design and project management to December 2001), Ms Judy Gardiner (research design and project management from January 2002) and Dr Terrence O'Brien (research design, interviewing, data collection and analysis, and report writing).

The following members of the Teacher-Assessed Tasks Working Party participated through interviews and advised on the research. Their contribution was much appreciated:

Sue Kloeden, and Jenene Rosser (The Association of Independent Schools of Queensland Inc.),
Amanda Sheridan (Brisbane Catholic Education),
Margo Bampton, Joy Schloss, and Lee Willett (Education Queensland),
Murray Coward (Future School),
Catherine Cushing (Federation of Parents and Friends Association Queensland),
Veronica Plummer (Toowoomba Catholic Education),
Beverly Day (Queensland Independent Education Union),
Claire Clifforth, Jenny Kent, Rhondda Kretschmann, Robin Thomas, Donna Torr, and Julian
Toussaint (Queensland School Curriculum Council),
Jan Hargreaves (Queensland Teachers' Union)

The involvement of Harry Newman (Queensland Catholic Education Commission), Mark Snartt
and Vicki Knopke (Queensland School Curriculum Council) was also valued.

The cooperation of the following people who participated through interviews is appreciated:

Deborah Armitage	Vermont Department of Education, USA
Ronald Bibby	Department of Education, Northern Territory
Jocelyn Cook	Education Department of Western Australia
Maree Crowley	Darra Jindalee Catholic School
Eva DeVries	Queensland School Curriculum Council
Judy Dohrmann	Victorian Curriculum Assessment Authority
Jenny Donovan	Department of School Education, New South Wales
Linda Gough	Sherwood State School
Dianne Grantham	Department of Education and Community Services, Australian Capital Territory
Debbie Harrison	Mitchelton State School
Geof Hewitt	Education Department of Vermont, USA
Tanya Hopgood	Mitchelton State School
Suzette Holm	Queensland School Curriculum Council
Claire Hughes	Queensland School Curriculum Council
Irene Janiszewska	Department of Education, Training, and Employment, South Australia
Christine Ludwig	Queensland School Curriculum Council
Adrienne McDarra	Brisbane Catholic Education
Graham Meiklejohn	Queensland School Curriculum Council
Greg Nelson	Principal, Sherwood State School
Lisa Neville	Sherwood State School
Leanne Riemann	Sherwood State School
Sue Ryan	Mitchelton State School
Greg Skippen	Principal, Mitchelton State School
Dr Elena Stoyanova	Education Department of Western Australia
Melissa Taylor	Darra Jindalee Catholic School
Cheryl Underleider	Education Department of Kentucky, USA
Maureen Whiting	Principal, Darra Jindalee Catholic School

Executive summary

The purposes of this research were to investigate the feasibility and appropriateness of teacher-assessed tasks and using teacher judgment in the Queensland Years 3, 5 and 7 Testing Program in Aspects of Literacy and Numeracy¹. The research also explored the manner in which teacher judgment might be used and the aspects of literacy and numeracy that could benefit from teacher-assessed tasks as one approach to broadening the curriculum coverage of the Testing Program. It also examined the possibility of assessing other valued literacy practices and new multiliteracies. The possibility of a trial of a testing procedure that would be based upon teacher-assessed tasks and teacher judgments was also considered as part of the research implications.

The need to consider teacher-assessed tasks and teacher judgments within the Testing Program was taken up from Professor Ball's review (Review of Queensland Literacy and Numeracy Testing Programs 1995–1999, Queensland School Curriculum Council 2000a). He stated that the current test formats were perceived as too narrow, broadening the test domain would please some curriculum groups, and allow for changes due to emergent English and Mathematics syllabus developments (p. 10). From Ball's review and further literature the major aim of the research emerged. This was to investigate if teacher-assessed tasks could contribute to improvement of the Queensland Testing Program by broadening the literacy and numeracy curriculum coverage.

Improving the Testing Program is part of the Council's cyclical evaluation and review processes. Broadening curriculum coverage of the test would assist the Council in offering the school authorities deeper and/or additional information about student performance in literacy and numeracy.

The following questions were used to focus the research and construct a framework for investigation:

1. What does the literature say about teacher-assessed tasks and teacher judgment in large-scale systemic testing programs?
2. Within the constraints of test development, what processes would be necessary, in terms of validity and reliability, to allow teacher-assessed tasks and teacher judgment in the Queensland Years 3, 5 and 7 Testing Program?
3. Which aspects of literacy and numeracy assessment, if any, would benefit, in terms of accounting for, and contributing to the improvement of, student performance, from the use of teacher-assessed tasks and teacher judgments of student performance within the framework of large-scale testing?
4. What other valued literacy practices, including multiliteracies, might be assessed formally through teacher-assessed tasks within the framework of large-scale testing?
5. What examples are available of teacher-assessed tasks and teacher judgment processes in other statewide or international testing programs? What samples of materials can be gathered for analysis?

¹ Information regarding the Queensland Years 3, 5 and 7 Testing Program in Aspects of Literacy and Numeracy can be found on the Queensland School Curriculum Council website – URL <http://www.qscc.qld.edu.au>

Information was gathered from a literature review and through interviews with literacy and numeracy testing team members from other States, and internationally. Stakeholders' responses to the research were gathered from their nominees to the Teacher-Assessed Tasks Working Party. Queensland School Curriculum Council Literacy and Numeracy testing project officers and Mathematics and English curriculum development project officers, and a sample of teachers and administrators were also interviewed. Samples of teacher-assessed tasks and documents outlining consistency of teacher judgment strategies in large-scale census testing programs were collected and examined. Discussion with other curriculum development teams and Working Party discussions and activities also assisted in forming research conclusions.

Literature was analysed to form a critique, and were analysed to gather supporting evidence. Teacher-assessed test programs were collected and analysed for appropriateness in Queensland contexts and against the literature critique. Interview discourses became jointly constructed texts, which were analysed to reveal significant form and meanings in order to contribute to the general discussion.

A number of conclusions were reached:

1. Literature Search and Review

What does the literature say about teacher-assessed tasks and teacher judgment in large-scale systemic testing programs?

The literature search outlined some national and international trends where shifts in large-scale testing practices to include teacher-assessed tasks and teacher judgments rather than multiple-choice, computer marked tests were made. The literature suggests that not only can the curriculum coverage be broader, tasks more flexible and authentic, but also teacher involvement in scoring has contributed to assisting teachers to value their judgments.

In Queensland, senior secondary and early primary contexts use teacher-assessed tasks and consistency of teacher judgment processes in widespread and large-scale assessment programs. It appears that when assessment results are used for more than reporting to parents, such as for funding allocation, reporting student performance nationally, or for certification, validity and reliability of teacher judgment become larger concerns.

The literature also demonstrated that teacher judgments become more consistent over time when teachers work with assessment criteria, and work with other teachers to refine understandings of those criteria. It explained that moderation procedures, either formal as in senior secondary contexts or informal as with processes associated with *The Year 2 Diagnostic Net*, are important alternative processes to assist in consistency of teacher judgment.

The critical positioning brought about by the literature review suggested four stances that can be used to appraise other testing programs, the samples of test material, and further literature. These stances were about:

- the types of teacher involvement in the testing program, either in marking or test construction
- the implications the testing program has for linking with effective pedagogy
- the gains that could be made in teacher professional development regarding syllabus implications and test construction
- the processes used to assist in consistency of teacher judgment, or validity and reliability of teacher marking.

2. Processes to Ensure Validity and Reliability of Assessments

Within the constraints of test development, what processes would be necessary, in terms of validity and reliability to allow teacher-assessed tasks and teacher judgment in the Queensland Years 3, 5 and 7 Testing program?

Both the literature and the interview discourses imply the need for processes and training to ensure that teacher judgments are consistent, valid and reliable. Administration of the test tasks would have to ensure that variables that affect student performance, such as teachers misunderstanding the nature of the tasks, inappropriate teacher intervention, and/or inappropriate student collaboration be minimised. Marking processes or assessment procedures of the test tasks would require either markers to be trained formally to understand criteria, have their marking checked, and a proportion of test papers re-marked, or processes of consistency in teacher judgment be undertaken.

Moderation processes would ensure consistency of teacher judgments in teacher-assessed test tasks. These processes could produce a positive benefit for teacher professional development, and teachers would profit from the effects in 'calibration' or 'benchmarking' their judgments, although the benefit should not be seen as the sole reason for providing moderation processes in the Year 3, 5 and 7 tests. Constraining moderation processes would be the increased workload to Years 3, 5 and 7 teachers to mark and/or undertake moderation activities. The budgetary requirements for teacher classroom release would also become more of an issue, as would commitment to the Testing Program by the school systems.

3. Broader Literacy and Numeracy Aspects

Which aspects of literacy and numeracy assessment, if any, would benefit, in terms of accounting for, and contributing to the improvement of, student performance, from the use of teacher-assessed tasks and teacher judgments of student performance within the framework of large-scale testing?

Professor Ball (2000a) in his review stated that the test formats were perceived as too narrow and broadening the test domain would allow for changes due to emergent English and Mathematics syllabuses. The Council project officers in testing and in curriculum development supported this statement, although the real impact of the new syllabuses on the Testing Program has yet to be fully investigated.

Broadening the curriculum coverage with the use of teacher-assessed tasks can be viewed as a matter of equity. A greater diversity of the student population could demonstrate their knowledge of literacy as broadening would allow multiple ways of showing performance as well as multiple opportunities for students to do it. In the current formats many students are inhibited from doing this, especially if they do not have test literacy as well.

To account for student learning both **additional** information (for example in aspects of literacy: speaking, being critical with texts, and school-based writing, or in aspects of numeracy: volume, mass, and pattern and algebra) and **deeper** information (for example in other curriculum literacies, use of multi-modal texts, and problem solving) would be gathered by broadening the coverage. To contribute to improvement of student performance, broadening the coverage to allow for consistency of teacher judgment processes would give teachers insights into students' results before the return of reports. This would assist teachers in preparing for the return of reports, and in planning for individual student learning during the school year rather than the following year which, in most schools, involves a different teacher.

Broadening curriculum coverage to include other aspects of literacy may allow for speaking, and critical interpretation of text assessments, through use of multimodal texts, open-ended responses, and school-based writing assessments. It could be difficult, however, to design ways to assess speaking and listening within the context of statewide testing, especially if the focus is on demonstrations of speaking in interactive rather than only oral presentation situations.

Broadening curriculum coverage to include other aspects of numeracy could encompass oral and written responses to numeracy problems, broader areas of measurement, and the use of concrete materials and hands-on activities. Broader tasks would incorporate more diagnostic features in the numeracy test. An analysis of responses would help to ascertain the 'level' at which the student was working, and assist teachers to further program learning activities for individual students.

There were questions still asked about who wants this information, and what will it be used for; which impacts on what is assessed. To contribute to improvement of student learning, broadening the coverage and allowing for teacher judgment may present gains for professional development, especially when student performance data is based upon new syllabuses. Through the testing procedure, teachers would develop understandings of the syllabus content, teaching and learning processes, and assessment techniques, as well as in testing processes.

4. Valued Literacy Practices and Multiliteracies

What other valued literacy practices, including multiliteracies, might be assessed formally through teacher-assessed tasks within the framework of large-scale testing?

Formal testing of multiliteracies or other valued literacy practices could prove difficult, even within the processes of teacher-assessed tasks and using consistency of teacher judgment strategies, or the context of key learning area-specific literacies. The newness of the concept of multiliteracies, the diverse social and cultural practices it involves, and its ever-evolving nature make it difficult to explain, or to name the tools which are used to make meanings. Certainly, use of only Standard Australian English would make it beyond the current Testing Program to assess other diverse Englishes.

5. Sample Material Appraisal

What examples are available of teacher-assessed tasks and teacher judgment processes in other statewide or international testing programs? What samples of materials can be gathered for analysis?

There are a variety of ways in which teachers assess test tasks beyond central marking. More extended test responses need more extensive marking processes and some of these include teachers marking their own students' tasks. While Indiana uses central marking processes, Maryland uses local marking of their multiple task and cross-curricular tests at different sites around the state. Both train markers rigidly in understanding marking rubrics and marking processes. Kentucky and Vermont include classroom teachers marking their own students for writing portfolio in Kentucky and writing and Mathematics portfolio in Vermont. Year level teachers concerned attend out of school sessions in training to use assessment rubrics and criteria. Portfolio results are reported separately from writing demand, or in the case of Vermont, Mathematics multiple choice and short answer, test results.

Australian States also show a similar variety of teacher-assessed tasks. Australian Capital Territory employs classroom teacher assessment for their speaking test, as does Victoria in a writing and Mathematics tasks. Both present teachers with marking criteria and exemplars, in video and print format respectively. New South Wales uses local site

marking for their teacher-assessed tests in writing and extended Mathematics response tests, while the Northern Territory offers teachers an option to mark their demand writing task before it is centrally marked. Western Australia uses teacher-assessed tasks in their sample tests of English and Mathematics, but their literacy and numeracy census tests are machine scanned. Tasmania is not using teacher-assessed tasks at the present time and South Australia is trialling teacher-assessed tasks along with processes of consistency of teacher judgment as a way to collect statewide performance data. All use standardisation processes and training of teachers to enhance validity and reliability of test marking. There was little use of consistency processes in as far as marks being moderated through formal procedures or informally through discussion with other markers.

Sample teacher-assessed tasks in testing programs from the USA offer the Queensland Testing Program some avenues to explore for the broadening of the program. As Maryland's cohort is similar to Queensland's the processes used to administer and assess their open-ended tasks could be worthy of investigation. Portfolio assessments similar to Kentucky and Vermont could be adapted for trial in Queensland to include extended response literacy and numeracy tasks related to Queensland syllabuses. The Speaking Test from Australian Capital Territory could also be adapted for the Queensland testing context, as could the processes Victoria uses for teacher-assessed extended mathematics response and teacher-assessed writing tasks.

The processes New South Wales use to locally mark their tests could be further investigated for their use in local site marking. Teacher assessment of demand writing tasks resembling the Northern Territory processes could also be used in Queensland. While not broadening the coverage, both would allow for more teacher involvement in the Queensland Testing Program.

Concluding Statements

This research project was only able to touch the surface of the issues of teacher-assessed tasks and using teacher judgment in the Queensland Testing Program as one way to broaden the literacy and numeracy curriculum coverage. Investigation of computer adaptive testing, beyond the parameters of this research, would also broaden the curriculum coverage but teachers may not be involved in that assessment to a large degree.

Teacher-assessed tasks are a way in which the curriculum coverage of the existing Years 3, 5 and 7 tests could be broadened. Such tasks might involve either formal training of teachers in marking to set criteria, or teachers marking their own students' tasks. Using processes to develop consistency of teacher judgment strategies could be a path to follow if teachers are involved in assessing the test tasks of their own students.

Test items or assessment tasks that require more open-ended or extended responses would broaden and deepen the existing curriculum coverage of the Testing Program. A wider range of literacy and numeracy test tasks would allow for additional or deeper information of student performance in those areas.

Broadening the curriculum coverage may also challenge the purposes of the Testing Program. The impact this would have on the balance between accounting for, and contributing to, the improvement of student literacy and numeracy learning would have to be further investigated, especially as a change in the purpose of the program may raise issues related to sample or census testing. The decision regarding developing teacher-assessed tasks for sample tests would be more to do with the recipients of the results and what they do with them. The school authorities would need to be consulted

regarding this. If the teacher-assessed task were to add to or enhance existing data, then sample tests would be appropriate. In addition, this would need clear communication of the intent of the Testing Program to stakeholders, parent bodies and the wider education community.

Improved accountability, in terms of additional and deeper information being gathered, and improved student performances, in terms of teachers being more able to use the test results to plan for further student learning, would occur from teachers being more involved. The test would be seen more as an adjunct to classroom assessment than at present and used by teachers to plan for improved learning for their students by adding to existing classroom assessment data, instead of being seen as isolated from it.

The more the test items relate to newer syllabus documents, or their marking encourages effective assessment practices, the more teacher professional development occurs. Broadening the curriculum coverage begs a question of the relationship between the Testing Program in aspects of literacy and numeracy and the developing English and Mathematics syllabuses.

More teacher involvement with the test would not only increase workloads for Years 3, 5 and 7 teachers but also the Testing Program budget. This increase would not be valued highly if appropriate teacher release for training and marking was not available, or if the benefits for improving student learning were not obvious, especially with parent unease about teacher absence from the classroom.

1. Introduction

1.1 Purposes of research

Within the context of:

- the Queensland Years 3, 5 and 7 Testing Program, the purposes of which are to account for, and to contribute to the improvement of, student learning in aspects of literacy and numeracy (Queensland School Curriculum Council 2000b, p. 1)
- large-scale literacy and numeracy testing programs interstate and overseas
- views and needs of stakeholders: students, parents, and schools within the Queensland school authorities.

The purpose of this research was to:

- investigate the feasibility and appropriateness of teacher-assessed tasks and using teacher judgment as part of the Queensland Years 3, 5 and 7 Testing Program in Aspects of Literacy and Numeracy
- explore the manner in which teacher judgment might be used and the aspects of literacy and numeracy which would benefit from teacher-assessed tasks
- examine the possibility, if any, of assessing further valued literacy practices and new multiliteracies.

1.2 Background to the research

Background contexts influential upon this research project came from two separate, yet highly related fields. One was the review and evaluation process of the Queensland Years 3, 5 and 7 Testing Program, while the other was the development of Queensland School Curriculum Council (the Council) syllabuses and accompanying assessment principles. Areas that were to confine the research outcomes emerged as the research unfolded. These were the characteristics of testing program development, assessment of aspects of literacy and numeracy, and multiliteracies, debates regarding sample versus census testing, and the purposes of testing. Each context and constraint is discussed below.

In 1999 a review of the Queensland Literacy and Numeracy Testing programs from 1995 to 1999 was undertaken. Dr Glenn Finger of Griffith University prepared an *Issues Paper* (Queensland School Curriculum Council 1999a), which discussed a range of issues regarding the conduct of the Testing Program over the five-year period, and suggestions to facilitate planning for future testing. Subsequently, using the issues paper as a base, Dr Sam Ball from the University of Melbourne prepared an evaluation and review report, which drew conclusions about the conduct of the Testing Program and suggested future directions for its improvement. A number of conclusions were reached, one of which, in answer to a question regarding the extent of the tests to assess parts of the curriculum they are designed to assess, was that:

Conversations might be initiated with relevant groups on the matter of including somewhat greater use of teachers and teacher judgment in the Testing Program (Queensland School Curriculum Council 2000a, p. 11).

This aspect of Ball's review motivated this research project, and in the formulation of research questions considered that:

This could allow a broader scope for the tests but would require teacher co-operation (2000a, p. 11).

The above begs two questions. The first was: to improve the Testing Program how could teachers be more involved? The second was about what exactly could be broadened in terms of the purposes of the Testing Program. Ball's review implied that broadening the test domains in literacy and numeracy was possible (p. 10). They could be broadened to include, for example, media literacy, oracy and science literacy, and more complex assessment tasks, he stated (p. 10). As teacher assessed tasks is a term that could encapsulate the use that Ball suggests, it is necessary for the purposes of this research to describe these activities. Table 1 offers an illustration and examples of some teacher-assessed tasks, and shows that these can range from central or local marking, to teachers assessing their own students' work. The ways in which the current writing component is marked could be seen as a teacher-assessed task, but Ball implies a movement more towards individual classroom teachers marking more complex tasks that require students to give more detailed responses. Therefore, the table is more oriented towards those school-based teacher-assessed tasks and processes to develop consistency of teacher judgment.

In a large-scale testing situation validity of test items or assessment tasks, and reliability of subsequent marking becomes an issue whenever subjective decisions are made regarding student performance, especially when those decisions affect information reported nationally, as in benchmarking, certification, or allocation of funds. The constructs of validity, reliability and consistency are defined in the next section. Table 1 shows how the purpose of the tests affects decisions regarding test item construction and marking processes. Research question 2 allowed investigation into this relationship. The findings are displayed in the table by comparing purposes for tests with types of teacher assessment.

This research project was undertaken within the context of a shift in assessment practices, especially in those practices related to a learning outcomes approach. Queensland Year 1 to Year 10 syllabuses were being developed to reflect an outcomes approach, and a project on *Consistency of Teacher Judgment* (Queensland School Curriculum Council 2000f) was undertaken. Research on the appropriateness, efficiency and effectiveness of annotated work samples was also being investigated (Queensland School Curriculum Council 2001a). It was timely that the feasibility and appropriateness of teacher-assessed tasks and teacher judgment within the statewide Testing Program be examined.

Queensland Years 1 to 10 syllabus documents in each key learning area propose principles of school-based assessment. The syllabuses state that for school-based assessment to be effective it should:

- focus on students' demonstrations of learning outcomes
- be comprehensive
- develop students' capacity to monitor their own progress
- reflect current knowledge of child and adolescent development
- be an integral part of the learning process
- be valid and reliable
- reflect social justice principles (Queensland School Curriculum Council 1999b).

Table 1
Types of Teacher-Assessed Tasks and Examples

Type of Teacher Assessment	Features	Examples of large-scale census assessment (Examples in Section 6)	Purpose
Centrally developed tasks with central marking.	Tasks developed centrally through panelling and trialling processes. Item response theory applied. Markers are trained to understand and use centrally developed criteria. Markers gather in a central location to mark all tasks.	Queensland Years 3, 5 and 7 Testing Program — Writing, spelling and dictation tasks. Queensland Core Skills tests.	Accountability and Improvement. Inform against benchmark standards OP ranking Certification
Centrally developed tasks with local marking.	Tasks developed centrally through panelling and trialling processes. Item response theory applied. Markers are trained to understand and use centrally developed criteria. Markers gather in local sites to mark the tasks of other schools in the locality.	New South Wales Primary Writing Assessment, Writing task for English Language and Learning Assessment, and extended response task for Statewide Numeracy Assessment Program (See figures 11 and 12)	Accountability. Inform against benchmark standards
Centrally developed tasks with teachers marking own students. May be on-line marking.	Tasks developed centrally through panelling and trialling processes. Item response theory applied. Teachers are sensitised to understand and use centrally developed criteria (sometimes through centrally driven professional development or just by publishing criteria), and may undertake consistency processes to understand criteria and/or to mark the tasks.	Victoria Achievement Improvement Monitor — writing and mathematics tasks — consistency processes encouraged. Marks are added to centrally marked writing and mathematics tasks. (See Figure 14) Northern Territory Multi-level Assessment Program — teachers opt to mark the writing tasks which are then centrally marked. Consistency processes not used. Central markers do not see teachers' mark. (See figure 13) Australian Capital Territory Testing program — Speaking task — consistency processes encouraged. (See figures 9 and 10) Queensland Year 2 Diagnostic Net validation tasks — consistency processes seen as highly necessary. Teacher release provided.	Accountability. Inform against benchmark standards Accountability. Inform against benchmark standards Accountability Accountability Funding
School-based teacher developed tasks with central marking.	Individual teachers develop assessment tasks. Products are sent to a central location for marking. Markers are trained to understand and use centrally developed criteria. Markers gather in a central location to mark all tasks	No examples	
School-based teacher developed tasks with teachers marking (rating) own students. Local verification of marking (rating).	Individual teachers develop and mark assessment tasks undertaken by their own students. Products sent to local sites. Markers (panellists) trained to understand and use centrally developed exit criteria and standards. Markers (panellists) gather in local sites to verify the ratings of other schools in the locality.	Senior Secondary Criteria Based Assessment tasks. Expert panellists moderate a sample of marked folios from each school. Central authority monitors student performance in schools.	OP ranking Certification
School-based teacher developed and validated tasks with teachers marking own students. May be on-line marking.	Individual teachers develop assessment tasks. Teachers are sensitised to understand and use centrally developed criteria, and may undertake internal or system-wide moderation processes to understand criteria and/or to mark the tasks, and place students on a learning continua.	Queensland Year 2 Diagnostic Net — teachers make judgments on student performance against indicators of learning on a continua. Consistency processes are seen as highly necessary. Teacher release is provided.	Accountability Funding Diagnosis

The same, or similar principles of school-based assessment could be the case in a statewide testing program, although comprehensiveness, validity and reliability of assessment become larger issues when assessment is used in reporting student performance to State, and federal governments, or when the stakes² rise to include rewarding or punishing schools and teachers for not reaching required performance goals. Moreover, if there is more alignment between classroom-based assessment processes and the processes involved in test development, administration and marking, then one process will not be perceived as having more worth than the other. Both become mutually supporting enterprises (Sanders & Horn 1995).

The Council's syllabus sourcebook guidelines detail suggestions of techniques for gathering information about student's demonstrations of learning outcomes. Techniques of observation, consultation, self-assessment and peer-assessment provide types of information from different assessment or learning situations (Queensland School Curriculum Council 1999c, p. 57). Focussed analysis is also suggested as an information gathering technique, and tests are considered as part of this as they allow for teachers to identify students' strengths and weaknesses through a variety of oral or written tasks. A variety of techniques would ensure that information regarding student performance was more authentic, compared to only using a narrow range. As a focussed analysis technique, tests may form part of the variety of assessment techniques necessary to gather student performance information, as the Queensland Years 1 to 10 syllabus and support materials maintain. Just how stakeholders see an externally developed and marked Testing Program, as another assessment technique along side other school-based assessment techniques is another matter, however.

Knowing that only some aspects of literacy and numeracy can be assessed through pencil and paper, multiple-choice items has been understood for some time. Ball's review (Queensland School Curriculum Council 2000a) concluded that current test domains are basic and vital for future student success. He stated that the perceived narrowness of the Testing Program should not deny the importance of the domain that is being tested (p. 10). Nevertheless, more extended responses to open-ended questions would more broadly assess student performance, and allow for gathering of additional or deeper performance information. This information would both provide better information to systems and could better contribute to student improvement. Research question 3 assisted in gathering information about deeper and additional student performance data.

This research acknowledged the issues involved with the aspects of large-scale testing programs, such as the issues involved in test item (task) development, trialling, administration of test materials, marking of tests, analysis of test scores and reporting of test results. These aspects can constrain the expansion of the test coverage, and the involvement of teachers in assessment of the tasks, as well as the issues of increased workloads for teachers.

The *Literacy: Position Paper* (2000c) and *Numeracy: Position Paper* (2000d), published by the Council, presented a range of assessable areas in literacy and numeracy. For literacy these come from the four resources of coding, semantic, pragmatic and critical practices (Queensland School Curriculum Council 2000c, p. 10) in strands from the trial syllabus documents of reading and viewing, writing and shaping, and speaking and listening (Queensland School Curriculum Council 2002a). Student ability in negotiating multiliteracies and literacies across the key learning areas also offers indicators of

² The term 'High-stakes', used in American testing contexts, describes the influence the test results have in decisions regarding student, school or teacher performance. Test results can also be linked to funding controls, certification of students, entrance or exit requirements, or accountability.

performance. For numeracy assessment areas come from four resources of foundational, linking, pragmatic, and critical practices (Queensland School Curriculum Council 2000d, p. 10) in strands of number, space, measurement, chance and data, and patterns and algebra (Queensland School Curriculum Council 2002b). At present, the literacy test items mostly come from the coding and semantic resources in reading and viewing, and writing and shaping. The numeracy test items mostly come from foundational and linking resources in number, measurement and data, and space. Teasing out these issues also became a focus of this research and formed the third and fourth research questions. The curriculum broadening is discussed more fully later as well as issues involving multiliteracies and its definition.

The debate regarding sample versus census testing has more to do with the purpose of the tests, and client needs, than further broadening the curriculum coverage. As the purpose of the Queensland Testing Program is to account for, and contribute to, the improvement of student performance, this drives the decision regarding selecting a sample group for testing, or testing the total population. As this research unfolded sample versus census issues emerged, especially when questions were asked regarding whom the test results would be for, and what was being tested in the broader literacy and numeracy curriculum. If the broader tests are contributing to national benchmark data, then census or sample testing may not be an issue. If, however, the tests are contributing to reports to parents and students, census testing may be more suitable. Wyatt-Smith and Ludwig (1998) add another dimension. Although their work is about testing in literacy, it has implications for testing itself. They suggest three points that serve to address which testing program would be more appropriate. The first is about educational relevance; how 'methods are relevant to educational objectives and ... assessment constructs', the second regarding the respectability that is attributed to certain assessment practices, while the third is to do with the costs of the program. Wyatt-Smith and Ludwig conclude that, although theirs is not an exhaustive list of criteria, the 'relevance-respectability-cost' criteria highlights the complexity in the decision making when choosing the appropriateness of the type of testing program (p. 12). These issues are discussed later especially in terms of the relevance of sample or census testing, if teacher-assessed tasks are 'respectable' processes when marking large-scale test items, and the budgetary requirements needed for teachers to be more involved.

1.3 Research questions

The major aim of this research was to investigate the extent that teacher-assessed tasks could contribute to the improvement of the Queensland Testing Program. The following questions were used to focus the research and construct a framework for investigation:

1. What does the literature say about teacher-assessed tasks and teacher judgment in large-scale systemic testing programs?
2. Within the constraints of test development, what processes would be necessary, in terms of validity and reliability, to allow teacher-assessed tasks and teacher judgment in the Queensland Years 3, 5 and 7 Testing program?
3. Which aspects of literacy and numeracy assessment, if any, would benefit, in terms of accounting for, and contributing to the improvement of, student performance, from the use of teacher-assessed tasks and teacher judgments of student performance within the framework of large-scale testing?
4. What further valued literacy practices, including multiliteracies, might be assessed formally through teacher-assessed tasks within the framework of large-scale testing?
5. What examples are available of teacher-assessed tasks and teacher judgment processes in other statewide or international testing programs? What samples of materials can be gathered for analysis?

1.4 Research approach

Data was gathered from:

- a literature search
- interviews from literacy and numeracy testing team members from other States, and internationally
- collection of samples of teacher-assessed tasks, and documents outlining consistency of teacher judgment processes in large-scale census testing programs
- interviews with the Council's Literacy and Numeracy testing project officers, and Mathematics and English curriculum development project officers
- interviews with a sample of teachers and educational administrators
- discussion with other curriculum development teams, and members of the Teacher-Assessed Tasks Working Party.

The analysis of collected data included the following:

- research literature was analysed to form a critical stance
- articles were analysed to gather supporting evidence
- program and test samples collected were analysed against the critical stance informed by the literature critique.

Interview discourses became jointly constructed texts, which were analysed to reveal significant form and meanings related to issues regarding teacher-assessed tasks and broader curriculum coverage.

Other questions and analyses were raised during the conduct of this research, but the original questions remained the major guide to data collection and analysis. Reporting was achieved through members of the Testing Program Advisory Committee, who were invited to nominate for a Teacher-Assessed Tasks working party. The working party represented the views of stakeholders and served as a discussion group to advise on the research, and to investigate significant questions and issues as they arose.

2. Literature Search and Review

Research Question 1

What does the literature say about teacher-assessed tasks and teacher judgment in large-scale systemic testing programs?

2.1 Contexts for shifting emphases on large-scale tests

Assessment and reporting literature since the 1990s reveals much about teacher judgments of student performance how these judgments can be valid³ and reliable⁴, and how valuable they are in recording student learning. Mostly, the literature argues for a shift away from assessment practices that are narrow, rigid, and decontextualised to more ongoing gathering of work samples that provide demonstrations of learning outcomes. Assessment, or marking, of the work samples involves processes to ensure consistency⁵ of the teacher judgment, rather than computer scanning of responses. Certainly, large-scale testing programs also reveal the same shift (Barton 1999, p. 4), and there are organisations that assess these programs to ensure they are broad, flexible and encompass authentic learning contexts (Fairtest 2001).

There are international and national contexts surrounding assessment techniques, mostly to do with accounting for student performance. In a climate where governments and the general public wish to have education systems that are more accountable, *assessment* and *testing* assume different perspectives, and different purposes. Assessment is generally regarded as school-based and non-standardised and serves to report individual student progress to parents, while testing is regarded as standardised, externally produced, administered and marked, and serves to report student performance to systems for accountability, funding or other systemic purposes. Although teacher judgments and school-based assessments have and do occur in large-scale assessment programs, governments often favour external assessments such as examination and standardised tests to account for student learning.

In Queensland *The Year 2 Diagnostic Net*, more widespread than large-scale, is an assessment program developed within the context of educational accountability such as the above and early literacy and numeracy improvement. Using teacher judgment, data on student achievement, in aspects of literacy and numeracy, in the early years of primary school, is collected and analysed against developmental learning behaviour criteria (continua). Luke et al. in their appraisal report, discussed these broad contexts within which *The Year 2 Diagnostic Net* came about. They maintained:

...an *international* context involving corporate managerialist moves towards outcome-based education, the development of student achievement standards and greater centralised control of schools and increased accountability of teachers for

³ Kerlinger (1986, p. 417) defines validity as epitomised in the question: Are we measuring what we think we are measuring? For the purposes for this research a valid teacher judgment is one that most accurately describes student performance against learning outcomes, standards or marking criteria.

⁴ Kerlinger (1986, p. 405) defines reliability as epitomised in the question: If we measure the same set of objects again and again with the same or comparable measuring instrument, will we get the same or similar results? For the purposes of this research a reliable teacher judgment is one that can be replicated across similar performances of different students on similar tasks, or in different schools and situations.

⁵ Consistency can be defined as: A consistent teacher judgment about student performance is one that is comparable with other teachers in a school, and teachers in other schools, on the same student performance. (Queensland School Curriculum Council 2000f, p. 1)

the quality of student outcomes. ...a *national* context, showing evidence that these ideas travelled to this country and resulted in attempts to develop common frameworks for curriculum, assessment and reporting of school achievement and moves to enhance regular systemic monitoring of student outcome standards (Luke, Land, van Kraayenoord & Elkins 1997, p. 19).

These same international and national contexts shaped the intent of the Council's project on *Outcomes-based Approaches to Assessment and Reporting* (2000e). This report revealed that:

Queensland has developed a unique culture of assessment and reporting over the last twenty years, particularly in the secondary sector. This culture that demands and relies on school-based assessment and teacher judgment is well grounded in research and is highly effective in practice. (p. 19)

Senior secondary schools have established formalised processes that lead to teacher judgments about student achievement that are deemed to be fair and reliable. On the same hand, since 1995, teachers in early primary settings have undertaken less formal, but similar, processes to ensure fair and reliable judgments when using *The Year 2 Diagnostic Net*. An external review of *The Year 2 Diagnostic Net* revealed that teacher judgment consistency was problematic, especially as teachers appeared to confuse concepts of *certainty of judgment* and *consistency of judgment* (Stewart-Dore & Bartlett 1999). Perceived *consistency* was apparent, but teachers wished for their judgments to be more '*certain*'. Stewart-Dore and Bartlett found teacher judgments were consistent and this was achieved through formal and informal peer review and networking, within and across, year levels within schools. Teachers saw this moderation process as an effective means of negotiating agreement about interpretations (p. 40). There was less consistency across schools, and this was where teachers questioned the judgments' *certainty*. Stewart-Dore and Bartlett felt that once teachers had worked more with the continua, their concerns about *certainty* would diminish. At present, moderation processes have moved to only include key teachers in literacy and numeracy from schools meeting to discuss and moderate on samples of work from their schools. It is supposed that the key teachers become sensitised to the criteria and in turn be able to offer 'expertise' on judgments at their base schools. Teachers' interviews revealed that some hold *The Year 2 Diagnostic Net* processes in high regard, although it was mentioned that the performance data is problematic. Because it is sometimes linked to intervention funding, data validity has been questioned by some systems.

2.2 Consistency (Reliability and Validity) Implications

In large-scale testing programs, consistency of teacher judgment is mostly achieved through training of teachers as markers to ensure reliability and validity of work marked. Where there are complex assessment criteria, or test tasks that are more open-ended, 'markers' would negotiate the meanings of the criteria, and 'group mark' the test responses. In some testing programs reviewed during this research markers were often trained formally, and if unreliable were retrained or rejected (Maryland Department of Education 2001). Often there would be 'out of school' meetings to discuss the protocols for assessment of these more in-depth responses (Vermont Department of Education 2001). Administration of the tests was 'standardised', so that variables that would affect individual student performance would be minimised. School-based assessments employ different kinds of techniques to guarantee reliability and validity. See table 1 for examples.

A project undertaken by the Council to identify strategies that support consistency of teacher judgment offers a description. Titled *Consistency in Teacher Judgment* (2000f), the report acknowledged that teachers' professional judgment was fundamental to

assessment and reporting processes which are advocated in the developing syllabus documents. But, the key issue, they stated, was consistency:

A key issue linked to the role of teacher judgment is consistency. Teachers and the broader educational community need to be confident that a teacher's judgments about students' demonstration of learning outcomes are consistent with the judgments of other teachers in a school, and teachers in other schools (p. 1).

Griffin (1997) mentions the tension that exists concerning the need for an accountable system and the best methods for gathering information on student achievement: He maintains:

In the past, governments have valued external assessment practices such as testing, and devalued the information that teachers gather daily about their students' learning. In this context even teachers have themselves learned to devalue their professional judgment and knowledge of their students' learning and development (p. 24).

Yet teacher judgment as Rowe (1997) discovered, after an initial period of working with and understanding the assessment criteria, becomes increasingly reliable. Data gathered from scoring of tasks by original assessors (often the teachers of the students) against expert assessors, to determine correlation, has been undertaken in California as part of an assessment procedure termed *The Learning Record* (Center for Language and Learning 1999). It was found that correlation between assessments was extremely high once the original assessors had been working with *The Learning Record* criteria for over a period of three years. This being so, teachers working with assessment criteria over time, and working with others to refine their understandings of the criteria, assists in making judgments on student performance more consistent.

Griffin (1997) asserts:

Teachers' observations certainly have been open to criticism about bias and lack of objectivity ... but teachers' assessments have the advantage of possessing many more consequences for individual student learning than does external testing.

Maxwell, in a discussion paper on teacher observation in student performance (Maxwell 2001), says that systematic gathering and recording of information from observations on student learning allows assessment to be more comprehensive, connected, contextualised, authentic and holistic. He writes:

It can be argued that unless there is a strong connection between pedagogy and assessment, the assessment will be disembodied and discriminatory, that is, not connected to any means for improving student learning and privileging students with existing cultural capital (p. 4).

Sadler (2001) answers questions regarding the subjectiveness of teacher judgment. Many others are concerned about the lack of objectivity in assessment practices that rely on teacher judgment. Sadler answers:

To question judgments simply because they are subjective would lead to a rejection of most of the decisions made in everyday life. So the real issue is not whether judgments are subjective or objective, but how consistent those judgments are ...the debate is not between subjectivity and objectivity as such, but how credible, how consistent, and how meaningful the assessments are.

This 'quality' of teacher observation in assessment remains a concern for Maxwell (2001). To acknowledge this concern he argues that accountability and verification of observation assures credibility and consistency. He explains further by stating that quality is assured through teachers being able to explain and defend assessment

judgments to students, their parent(s) and other teachers, and being able to revisit the foundations for assessment judgments (p. 6).

Sanders and Horn (1995), in an article that discusses the usefulness of standardised and alternative measures, write that alternative forms of assessment (other than standardised, centrally marked, pen and paper tests) are also viable as long as care is taken to assure validity and reliability, as others, above, would agree. Sanders and Horn question the usefulness of alternative forms of assessment in large-scale testing because they are expensive and difficult to develop, administer and score. The dilemma for stakeholders in large-scale testing programs is the trading off of costs against what kinds and types of information is required for accountability purposes.

Wiggins (1990) makes a point regarding the gains made, when teachers are involved in the procedures of test development and test assessment, in teacher professional development processes. He states:

... while the scoring of judgment-based tasks seems expensive when compared to multiple-choice tests the gains to teacher professional development, local assessing, and student learning are many. ... significant improvements occur locally in the teaching and assessing ... when teachers become involved and invested ...

As discussed previously, reliability and validity of marking of test responses is always considered problematical, especially from test items that require more extended responses. Research from the Council's *Consistency in Teacher Judgment* (2000f, p. 20) identified ways to develop consistency or reliability in teachers' judgments about student performance. These approaches were:

- planning collaboratively
- using a common assessment task
- developing a common criteria sheet
- comparing samples of student work (moderation)
- sharing understandings about the core learning outcomes and their developmental sequence
- sharing understandings about assessment.

Moderation, the report describes, is the process by which teachers meet to compare samples of student work and to discuss and compare the judgments they had made about student demonstration of core learning outcomes (Queensland School Curriculum Council 2000f, p. 11). Griffin and Smith (1996, p. 22) suggest that moderation helps to overcome localisation of standards (expectations of 'normal' student ability levels which teachers draw from practice) and can be an important adjunct of assessment that is based on direct observation. But, moderation is more than that. It is a social practice with accompanying discourses and social behaviours. Typical of any social practice, to allow others access to moderation practices, the discourse constructs need to be examined to discover existing meta-language, behaviours and attitudes. Moderation as social practice is discussed later in this report, as it is an important adjunct to the use of teacher judgment in large-scale testing programs.

In Queensland moderation is used formally in the final years of secondary schooling for the purposes of accrediting work programs and study plans, monitoring standards, verifying and approving levels of achievement (Queensland Board of Senior Secondary School Studies 1999, pp. 7–8), and ensuring that results recorded on Senior Certificates match the requirements of syllabuses (p. 53). Moderation processes are also used, but somewhat differently in early primary schooling with *The Year 2 Diagnostic Net*. These processes ensure that teachers can confirm their judgments about the particular literacy

and numeracy phases in which students operate so that phase allocations are comparable across classes and schools (Department of Education 1998, p. 12). Moderation processes are also used for assessment in Religious Education, the Council syllabuses, English and Mathematics as part of the Consistency of Teacher Judgment Program (Brisbane Catholic Education 2001) in Brisbane Catholic Education schools. Brisbane Catholic Education teachers meet in inter-school groups to moderate on levelled learning outcomes in Religious Education Profiles (Archdiocese of Brisbane 1997), and student performance standards in English (Department of Education and Community Services, South Australia 1996) and Mathematics (Department of Education, Queensland 1994). Brisbane Catholic Education did consider at one time collecting student performance data for accountability purposes, but they had neither the resources nor personnel to do so.

Internationally, States in USA such as Kentucky and Vermont use portfolio assessment⁶ procedures in their statewide testing programs. These are examined more fully later. While Kentucky uses portfolios for writing, Vermont has also developed mathematics portfolios. Freedman (1993), in an article titled *Linking Large-Scale Testing and Classroom Portfolio Assessments of Student Writing*, states that portfolios fit with good writing instruction and are a procedure for thoughtful classroom-based assessment. As well, they can be used for large-scale testing. The problem, however, she states is to make the links: teachers are concerned with instruction, testers with policy and accountability. Links lie in a reciprocal relationship between teachers and test developers. Teachers would need to work together to provide judgments of student ability, have those judgments checked by panels, and to aggregate the results — what Sadler (1995) would call ‘Professional Calibration’, and what Wiggins (1993) would term a process of ‘benchmarking’ their grading. On the same hand, test developers would:

... have to relax their fears that classroom teachers may in some way contaminate test data collected as a natural part of instruction... (Freedman 1993, p. 48).

Both the areas of test development and reliability of assessment, and teacher’s professional conduct as assessors, versus their positions as instructors are discussed later in this report.

2.3 Critical positioning

The literature revealed certain stances that can be used to form a critique. These stances were drawn on to evaluate further literature regarding large-scale testing programs, and sample teacher-assessed task material from interstate and overseas. The critique regarding samples of testing program components follow in section 6 of this report. The following stances emerged from the literature search:

- teacher-assessed tasks are not only about teachers ‘marking’ test papers, either centrally or locally, but also teachers being involved in developing assessment tasks and assessment criteria, and making informed judgments on student performance (see table 1)
- any large-scale testing program has implications for pedagogy. A strong connection between effective pedagogy, effective assessment practices and testing procedures is essential if the testing program is not to sabotage curriculum development and delivery

⁶ Portfolio Assessment is a procedure in which assessors examine a student folio. Queensland School Curriculum Council (1999c, p. 59) defines a student folio as a collection of a student’s work over a period of time. It may include day-to-day tasks, work produced for assessment items or selections of a student’s best work showing effort, progress and achievement. A folio containing a complete collection of a student’s work is often used to demonstrate progress. A folio containing selected items only is more commonly used for summative assessment and reporting.

- if teachers are involved with testing programs, especially those programs which reflect syllabus documents, gains are made in teacher professional development, not only in test development, and in assessment task development, but also in syllabus understandings
- consistency in teacher judgments not only comes from spot checks on markers, double or triple marking, training of markers, but by teachers moderating during the assessment process, and valuing processes for developing 'Professional Calibration'.

2.4 Conclusion

What does the literature say about teacher-assessed tasks and teacher judgment in large-scale systemic testing programs?

The literature search outlined some national and international trends where shifts in large-scale testing practices to include teacher-assessed tasks and teacher judgments rather than multiple-choice, computer marked tests were made. The literature suggests that not only can the curriculum coverage be broader, tasks more flexible and authentic, but also teacher involvement in scoring has contributed to assisting teachers to value their judgments.

In Queensland, senior secondary and early primary contexts use teacher-assessed tasks and consistency of teacher judgment processes in widespread and large-scale assessment programs. It appears that when assessment results are used for more than reporting to parents, such as for funding allocation, reporting student performance nationally, or for certification, validity and reliability of teacher judgment become larger concerns.

The literature also demonstrated that teacher judgments become more consistent over time when teachers work with assessment criteria, and work with other teachers to refine understandings of those criteria. It explained that moderation procedures, either formal as in senior secondary contexts or informal as with processes associated with *The Year 2 Diagnostic Net*, are important alternative processes to assist in consistency of teacher judgment.

The critical positioning brought about by the literature review suggested four stances that can be used to appraise other testing programs, the samples of test material, and further literature. These stances were about:

- the types of teacher involvement in the testing program, either in marking or test construction
- the implications the testing program has for linking with effective pedagogy
- the gains that could be made in teacher professional development regarding syllabus implications and test construction
- the processes used to assist in consistency of teacher judgment, or validity and reliability of teacher marking.

3. Processes to Ensure Validity and Reliability of Assessments

Research Question 2

Within the constraints of test development, what processes would be necessary, in terms of validity and reliability to allow teacher-assessed tasks and teacher judgment in the Queensland Years 3, 5 and 7 Testing program?

3.1 Teacher-Assessed Tasks

In everyday learning and teaching situations, teachers constantly make judgments about student performance. On occasion these judgments are documented, which contribute to reporting of performance to students themselves, parents, schools, and systems. As stated previously, when accountability across the system, State and national level is required, the consistency of these judgments becomes more an issue than when judgments are used to report to parents. In large-scale testing programs validity and reliability are major concerns.

To alleviate concerns, a number of testing programs in Australia and overseas develop procedures by which teachers (or other professionals) are trained to mark test papers. In some cases, as in numeracy test items, and multiple-choice responses, the marker's task is easier, as it requires a simple right/wrong mark or judgment. With more open-ended responses, especially in the case of marking 'writing' tasks, markers need to be sensitised to the differing quality of responses, and what the mark, or score, could represent. If descriptions of a standard are used to indicate a 'level' (as in levels of development, difficulty or sophistication), rather than competency then the task of marking becomes even more complex. Discussed later, formal procedures are used to enhance reliability and validity in the marking of test papers, including formal training and retraining of markers, double or triple marking of tests papers, re-marking of a proportion of test papers, or the use of control papers.

As seen in table 1, marking by teachers can be undertaken centrally (large groups of trained markers work in a central location), locally (trained markers work in clusters nearer to the test paper collection sites), or in the schools where the students sit the test. In each case training is seen as essential, and procedures to ensure reliable scores are considered crucial. The Queensland Years 3, 5 and 7 Testing Program uses central marking processes for the spelling, dictation and writing tasks with 10 per cent of papers re-marked to minimise reliability concerns.

The Council Literacy and Numeracy Testing Program project officers indicated similar concerns for valid and reliable test results. They maintained that tasks assessed by teachers in schools would need to be centrally constructed and accompanied with clear notes of procedures and protocols. They also stated that standards or criteria for marking should be clear, unambiguous, and specific, and the assessment process should include a full range of annotated response exemplars. They thought that teachers would need to try to be as objective as possible in this process. The English project team agreed that if teacher-assessed tasks were to be used in a large-scale testing program the tasks' effectiveness would depend on the training of teachers in task administration and use of marking criteria. They said marking guides would need to be unambiguous and marking and moderating processes would need to be rigorous.

These above comments reflect not only the subjectivity of teacher assessment commented upon by Maxwell (2001) and Wyatt-Smith (1995), but the dilemma teachers have when administering and marking a large-scale testing program. Teachers

interviewed said that they find it inappropriate not to intervene when students are in perceived difficulty. Certainly, literature (Freedman 1993; Wiggins 1993) regarding the differences between the concerns of test developers and teachers would suggest that teachers find it difficult not to intervene in a test situation. Concerns with *The Year 2 Diagnostic Net* also indicate similar difficulties when systems require reliable student performance data from teacher assessors.

3.2 Moderation

Strategies to develop consistency of teacher judgment were investigated as part of a research project for the Council: *Consistency of Teacher Judgment* (Queensland School Curriculum Council 2000f). While the focus of judgments was upon core learning outcomes of the Council's developing curriculum materials, the strategies are still pertinent to the discussion of how to make teacher judgments more reliable. The report established that strategies such as planning collaboratively, using a common assessment task, developing a criteria sheet, comparing samples of student work (moderation), sharing understandings about the core learning outcomes and their developmental sequence, and sharing understandings about assessment, assisted in positive ways to make teacher judgments more consistent.

Presently, within the Queensland Years 3, 5 and 7 Testing Program, common assessment tasks and criteria for marking are developed as part of test item, or task, construction. Practising teachers are not involved in this joint construction, however. Strategies that suggest teachers share understandings of the developmental sequence of outcomes, and of assessment, that they plan collaboratively, and compare samples of student work (moderation), is beyond the Testing Program as it exists. If moderation were to be a process to ensure reliability in teacher-assessed tasks, different processes that at present would have to be developed to ensure teacher judgments were consistent across schools and systems. Some of the processes to ensure reliability are described later, especially those used in other States' and overseas' large-scale testing programs, but moderation, as a unique and discrete process, is further described here.

As discussed previously, moderation processes in assessment are well grounded and appear to be highly effective in senior secondary schooling. Moderation in this context is the formal procedures used to accredit work programs and study plans, monitor standards, verify and approve levels of achievement (Queensland Board of Senior Secondary School Studies 1999). Although more informal, a moderation process of jointly discussing student performance and judgment criteria is understood in early primary settings and in Brisbane Catholic Education schools.

As the developing Council curriculum materials are suggesting informal moderation processes (Queensland School Curriculum Council 1999b, p. 31) to ensure valid and reliable assessments of student performance against learning outcomes, then over time moderation processes, similar to those in early primary settings, would be undertaken in year 3 to year 10 areas of schooling. Similarly, moderation processes necessary for assessment in Rich Tasks of the *New Basics Project* (Education Queensland 2001) will eventually become part of Year 3, 6 and 9 contexts in Education Queensland schools. If the Testing Program in Years 3, 5 and 7 requires teacher-assessed tasks in order to broaden the curriculum coverage, then moderation processes to assess test item or task marks may be developed. There is potential for this development to go hand in hand with curriculum planning and school-based assessment consistency practices.

Moderation serves many purposes, most of which are to do with consistency of teacher judgment of student performance, and in senior secondary schooling, consistency in work programs and verifying levels of student achievement assigned by teachers.

Moderation challenges teachers' views on their own positioning, and can cause reflection on the assessment processes of designing assessment tasks, collecting appropriate samples and making judgments. It often asks teachers to look with unbiased eyes, or different eyes at an assessment task, or work sample. Moderation contributes to teachers' own professional development as it not only values teachers' ability at making judgments on student performance, it can assist in further curriculum planning and contribute to positive conversations about assessment (Queensland School Curriculum Council 2000f).

Maxwell's (2001) discussion paper on *Teacher Observation in Student Assessment* looks at causes for inconsistency in teachers' observations. These causes he grouped as factors that affect the accuracy of teacher judgment. These factors are teachers' prejudgments and prejudices, selective perception, provision of inadvertent clues, and making an inappropriate inference. Maxwell suggests that moderation processes as suggested in the *Consistency of Teacher Judgment* report can be used to reduce the effect of these factors.

If moderation were tied to a large-scale testing program, the process would also challenge teachers to re-frame their judgments in light of the purposes of the program. Therefore, moderation practices for the purposes of large-scale testing may well have to be more formalised, especially as moderation is not yet part of assessment culture in many Years 3, 5 and 7 Queensland classrooms. Moderation discourse constructs would need to be developed and discussed with these teachers. Most vital are those constructs that have to do with knowing and understanding the assessment criteria, putting forward a case for judgment, arguing a particular judgment, negotiating and being able to modify the judgment. In particular, moderation encourages construction of teachers becoming aware that their judgments are appropriate and valued. The professional calibration construct, discussed later, and held in high regard in senior secondary areas of schooling in Queensland, is an important part of moderation processes.

Primary teachers interviewed raised issues about the possible use of moderation in the Testing Program often showed positive acknowledgment of the moderation processes used in *The Year 2 Diagnostic Net*. Teachers commented that they saw value in the time they had away from classroom duties to undertake moderation processes, and speculated that they would require similar time concessions if moderation was to be used in Years 3, 5 and 7 tests. Teachers in the Catholic sector also recognised the worth of having teacher release time to moderate assessments with others. Moderation, which is part of the *Consistency of Teacher Judgment Program* for Brisbane Catholic Education (2001), is considered a professional development strategy.

Teachers also indicated that they would feel their assessments of students' performance more worthwhile if they were asked to assist in the test marking process, as sometimes the current test results conflicted with their own judgments. In response to this type of concern, the Northern Territory Testing Program allows teachers to assess student responses prior to them being centrally marked. This procedure is discussed more fully later in this report.

3.3 Calibrated Teacher Judgments

Although borrowed from measurement of machinery, Sadler (2001, p. 4) uses the term 'calibration' to describe the social practices professionals engage in to fine-tune their individual competency levels in assessment. For teachers this involves collaboratively agreeing to keep themselves abreast of all that is involved in assessment practices, moderation processes, and student performance levels. It also means that teachers keep themselves knowledgeable about the standards and criteria used for judgment, to

ensure judgments are consistent with judgments other teachers would make given the same evidence. Sadler says that as professionals, teachers welcome this collaboration with other teachers, and want their judgments to be consistent with others. He says that they see this as their professional responsibility as a teacher, and the public expects professionals to make sound and consistent judgments. Sadler sees this process as vital to the profession of teachers, and through formal processes with being involved with *The Learning Record* (Sadler 2001) teachers develop and check their notions of being 'calibrated' (p. 4). Different from the processes for *The Learning Record*, calibration ideals are reflected in Queensland senior secondary and in lower primary areas, through teachers' work in moderation processes.

Wiggins (1993) argues that the calibration process is of extreme importance, especially if large-scale assessment programs use teacher judgment when marking test tasks. He insists that for test results to be useful and credible teacher groups must 'benchmark' their grading and work to develop more criterion-referenced procedures and better interrater reliability in their grading. He maintains that the experience of the state of Vermont, when unreliable scores from naïve and untrained assessors were revealed, illuminates the importance of the 'calibration/benchmark' process (p. 22).

Teachers interviewed from the Brisbane Catholic sector and those who had experience in early primary years in state schools shared the view that they possessed skills to reasonably assess their students' work, and those who had been involved in moderation processes were confident in their skills as assessors, and how the moderation process modified and affirmed those skills. Others and some from Education Queensland schools, who had only limited involvement, saw the process as questioning their professional thinking, not only in assessment but also in the activities they use for assessment. It was stated that they felt judgments were being made upon themselves as teachers, both in terms of task development and judgment of student performance, rather than on the work sample being assessed. Moderation meetings were seen as helpful in implementing Religious Education Guidelines by almost 50 per cent of early childhood/lower primary teachers compared with almost 40 per cent of other primary and less than 20 per cent of secondary (Archdiocese of Brisbane 2002). Nevertheless, implications from research have indicated that with continual involvement in moderation teachers begin to view the process more positively. Indication of this improvement is evident in the moderation activities evaluation reports from Brisbane Catholic Education (Archdiocese of Brisbane, Catholic Education 1996–1999).

Calibration adds another facet to the work of teachers, especially for those who may be involved with a testing program that uses consistency of teacher judgment processes. In some ways, the existing Testing Program releases teachers from the professional responsibility of which Sadler (2001) and Wiggins (1993) write, and allows lack of commitment to, or ownership of, it. The Testing Program through use of consistency of teacher judgment processes could assist in the development of teacher professionalism, especially in calibration of, or benchmarking, their judgments. Not only would accounting for student performance be no less consistent than it is at present, and assessment tasks would give deeper performance data, but teachers would have internalised the marking criteria, and know how to respond to student work, at times other than the testing, or data collection periods. Therefore, professional calibration would add another path to improving student performance.

3.4 Conclusion

Within the constraints of test development, what processes would be necessary, in terms of validity and reliability to allow teacher-assessed tasks and teacher judgment in the Queensland Years 3, 5 and 7 Testing program?

Both the literature and the interview discourses imply the need for processes and training to ensure that teacher judgments are consistent, valid and reliable. Administration of the test tasks would have to ensure that variables that affect student performance, such as teachers misunderstanding the nature of the tasks, inappropriate teacher intervention, and/or inappropriate student collaboration be minimised. Marking processes or assessment procedures of the test tasks would require either markers to be trained formally to understand criteria, have their marking checked, and a proportion of test papers re-marked, or processes of consistency in teacher judgment be undertaken.

Moderation processes would ensure consistency of teacher judgments in teacher-assessed test tasks. These processes could produce a positive benefit for teacher professional development, and teachers would profit from the effects in 'calibration' or 'benchmarking' their judgments, although the benefit should not be seen as the sole reason for providing moderation processes in the Year 3, 5 and 7 tests. Constraining moderation processes would be the increased workload to Years 3, 5 and 7 teachers to mark and/or undertake moderation activities. The budgetary requirements for teacher classroom release would also become more of an issue, as would commitment to the Testing Program by the school systems.

4. Broader Literacy and Numeracy Aspects

Research Question 3

Which aspects of literacy and numeracy assessment, if any, would benefit, in terms of accounting for, and contributing to the improvement of, student performance, from the use of teacher-assessed tasks and teacher judgments of student performance within the framework of large-scale testing?

4.1 Broader Curriculum Coverage

Teachers interviewed for this research were critical of the narrow range of competencies being tested, and the perceived importance that was being placed upon the test results by parents. Teachers felt that, if there were to be any inconsistencies between the test result and teacher judgments that occur on classroom-assessed tasks, especially on the end of year report card sent to parents, their own judgment would be seen as inappropriate rather than the test result. Use of teacher judgment, these teachers believed, would alleviate some of the problems of inconsistencies, and also allow teachers insight into student performance on the test tasks over the testing period. This could occur, teachers said, because further individual teaching and learning would result before students leave the year level.

The Council's Testing Program project officers, and members of the teacher-assessed tasks working party had concerns regarding the rationale for broadening the curriculum coverage of the Testing Program. It was felt that teacher-assessed tasks should only be used if the purpose of the test warranted it. Working party members, while appraising some interstate testing program teacher-assessed tasks suggested that some information could have been as easily gathered through a multiple-choice, pencil and paper test, and drew implications from this about the purposes for involving teachers more in the program.

The above responses indicated questions that were being asked, not of this research project so much, but of a change to the Testing Program itself. Some were about what was going to be tested, and whom the information is for, while others were about the tasks themselves. It was asked, 'What else do stakeholders want to know about students' literacy learning? Will the information be additional to what is gathered now, or will it be the same but deeper? How can school authorities, schools and teachers use the data collected from teacher-assessed tasks?' and, 'How will this contribute to improvement of student learning in aspects of literacy and numeracy?' These latter questions, addressed through interviews with project officers, are discussed below and later will form questions for further investigation in recommended follow-up research.

As Ball's (2000a) review of the Testing Program influenced this research project it holds some of the rationale for the need for broadening the curriculum coverage. The review focused upon nine questions, one of which asked about the extent to which the current tests assess those parts of the intended curriculum that they are designed to assess and whether the curriculum areas within and outside literacy and numeracy could be considered by the Council for testing. While it was claimed by Ball that the literacy and numeracy tests were assessing those important areas that they set out to assess, it was also concluded that a greater use of teachers and teacher judgment might be used in the testing process (p. 11). Professor Ball presented caveats to contextualise the conclusions reached. He stated that the current aspects of literacy and numeracy were seen as relatively narrow in terms of the total Queensland curriculum. Ball advocated that broadening the test domain was possible, but believed that even classroom

teachers contributing to the marking of writing tasks, while providing a useful professional development process, would increase budgetary requirements and year 3, 5 and 7 teachers' workloads.

The need to broaden the Testing Program was also referred to in another of Ball's caveats. This caveat had to do with the revision of the mathematics and English syllabuses, and the fact that the revised syllabuses are being developed in terms of learning outcomes. Ball saw that because of this, test development would become a clearer procedure, as the relationship between the assessment task and curriculum outcome would be more direct. He felt that to leave the testing in its current format risks the criticism by teachers that the tests call for a too narrow range of student responses. Ball suggested that the current test items and results students achieve would be especially important when English and mathematics outcome-approach syllabuses for Years 1 to 10 are implemented. The full impact of these syllabuses on the Testing Program, or when the literacy and numeracy test items should be based upon the new syllabus, is yet to be investigated.

Broadening curriculum coverage using teacher-assessed tasks is also a matter of equity. This can allow students opportunities to demonstrate knowledge of literacies that the current Testing Program does not, and give access to testing practices, and a greater likelihood of success, to a greater diversity of student populations. Making the processes of assessment broader (i.e. using teacher-assessed tasks with a variety of assessment techniques) will allow a greater diversity of the student population to demonstrate their literacies as broadening can allow multiple ways of demonstrating performance as well as multiple opportunities for students to do it. In the current formats many students are inhibited from doing this, especially if they do not have test literacy as well. Further discussion on valued literacy practices and multiliteracies and test literacy follows in the next section of this report.

The Council Principal Project Officer (Mathematics) stated that the new Mathematics syllabus (Queensland School Curriculum Council 2002b) would have a major influence on the Testing Program because of its emphasis on 'thinking and working mathematically' approaches to learning. This may lead to the numeracy test needing to incorporate more open-ended, extended response items. He predicted students would be taught mental strategies that encourage confidence and initiative when dealing with mathematical situations. He suggested that the emphasis on understanding and conceptual awareness might be less suited to the dominant use of multiple-choice items.

4.2 Broader Coverage of Literacy

Interviews with the Council's project officers said that broader literacy coverage and use of teacher judgment could include multimodal texts and questions that would require open-ended responses. It was believed this would mean that reading test items could be less reliant upon responses being within literal understanding of texts; and accordingly, movement away from multiple-choice items could allow for collection of data on student performance in critical interpretation of texts. The project officer could see that there could be use of more school-based writing assessments to complement the centrally marked task. It was stated that if teachers assessed the writing component, there would be opportunities for professional development, and teachers would better understand the assessment/marking process. The project officer thought this could be one way to alleviate the 'surprises', which sometimes exist at the moment between teacher judgment of students' writing and the writing component's test scores.

At present, students' abilities in speaking are not tested, and the project officer could see that changes to the Testing Program could encompass this. She could also see that

videotapes, or on-line visuals, could be used as stimulus (prompt) material, and used to collect information on student performance in reading and viewing. Students were exposed to this type of material out of school, she stated. It was also felt that teacher-assessed tasks would allow assessments to be tailored to students' needs in a broader context than at present. In the situations where the testing dates for some students, for emotional or physical reasons, are inappropriate, there could be alternative arrangements made.

The Council English curriculum development project team indicated that finding ways to effectively assess students' critical use of and work with texts (i.e. text-analysing resources) is still the focus of ongoing investigation. They felt that the assessment of critical understandings would not necessarily be achieved by introducing teacher-assessed tasks to the Testing Program. Similarly, they suggested that even with the introduction of teacher-assessed tasks it would be difficult to design relevant and purposeful ways to assess speaking and listening within the context of a statewide testing program. This would be particularly challenging when focusing on demonstrations of speaking and listening in interactive, rather than only oral presentation, situations.

The project team acknowledged that the addition of teacher-assessed tasks may allow for the collection of more diagnostic information and therefore enable the Testing Program to more directly contribute to the improvement of student performance. It was also argued that ongoing school-based assessment, with a focus on the alignment of pedagogy, curriculum and assessment and the development of relevant and purposeful assessment tasks would remain better placed in classrooms to achieve the improvement purpose of the Testing Program. Working party members agreed, and saw the connections between testing results, school-based assessment practices, and improvement of performance were the diagnostic aspects, and direct feedback that teacher-assessed tasks would create.

4.3 Broader Coverage of Numeracy

The Council's project officer in numeracy testing thought that teacher-assessed tasks and teacher judgments would allow for activities involving more open-ended responses and the use of oral and written explanations in test items. She felt that collection of data on student performance in problem-solving activities, with an emphasis on the process of how students go about solving them, and other mathematical concepts could be undertaken. Numeracy tasks that use an oral language explanation, as in explaining the meaning of symmetry, could also be offered, and individual mental computations could be administered orally with responses recorded by teachers. The project officer could see that 3D material could be used for measuring spatial concepts, and students could demonstrate concepts of measurement, recognising amounts of money, or counting out change, and the like. She could see teachers using concrete materials to assess younger students' understanding. She also thought that a range of test items, to assess an outcome, could be provided, and therefore allow teachers to select appropriate tasks to suit individual needs of students.

Mathematics curriculum development project officers stated that while length and area can be reasonably assessed with multiple choice and pencil and paper, they thought that other aspects such as volume, and mass could be better tested than at present. It was suggested that student ability in pattern and algebra could also be assessed with more open-ended tasks and use of concrete materials. They gave an example of an 'interview task' that could be undertaken with students explaining their strategies, with the student being further challenged by teacher questioning. This, they thought, would give more authentic information of ability. It was mentioned that validation tasks, similar to those

from *The Year 2 Diagnostic Net* tasks, could better cater for individual differences and students' diverse needs.

The mathematics curriculum project officers suggested that teacher professional development would occur when students are asked to solve more open-ended mathematical problems, or when testing conceptual issues. They stated that while teachers observed students at work one-on-one, interpreted their responses, and made suggestions about other strategies or ideas that students could investigate, they would be engaging in mathematical content and thinking. They thought that the test at present doesn't allow teachers to engage much in mathematical content. The project officers suggested that test marking criteria would help teachers to focus their observations, and see students do things that might otherwise go unnoticed. The development of multi-modal test items through the use of computer technology was discussed. It was stated that as this would assist in broadening the numeracy test, especially in the areas of item response and feedback. Students would also be able to be involved in rotating and manipulating shapes, loan calculators and processes for borrowing money.

It was also stated that broader tasks could incorporate more diagnostic features in the numeracy test - where a range of responses could be accepted from the same stimulus question or problem. They stated that analysis of these response ranges would help in providing evidence of demonstrations of learning outcomes and contribute to decisions regarding the 'level' at which the student was working in particular strands, and assist teachers to further program learning activities for individual students.

These project officers mentioned that if numeracy test reports were going to assist in reporting in outcomes, teachers, through test involvement, would gain further knowledge of mathematics processes and the outcomes approach syllabus. It was stated that the aspect of shape is better addressed in the new syllabus, and through teacher-assessed tasks, solid shapes could be used for students to hold and manipulate. Students would be able to design, analyse and interpret their own graphs (especially when computers were available to support their construction), which would allow for more individualised tasks which could be more culturally appropriate.

It was stated that teacher-assessed tasks would help reassure teachers that their judgments were valuable, and assist in teacher knowledge about test development. It was mentioned that, at present, teachers are not aware of how the tests are developed, or how to best interpret and use the results either at the classroom or whole school levels for further improvement in student learning. Certainly, interviews with teachers for this research bore this conclusion out. Teachers were not able to comment on how the test results could be used to improve students' performance because they felt this was constrained by the timing of receipt of test reports.

Teachers interviewed were sometimes confused with the term 'broader'. They stated that they often felt the numeracy test was broad enough as there were items in the test that they considered too difficult and beyond the year levels that were being tested. Once clearer they commented on the need to have broader 'hands on' numeracy tasks or items, especially in the year 3 tests, and more process-oriented writing tasks. Further questioning revealed that these areas were necessary as these were the ones that the test results and their own assessments of students did match as closely. There is an indication here again of teachers being more concerned about instruction than accountability, and that their own assessment practices are often devalued through the test results.

4.4 Conclusion

Which aspects of literacy and numeracy assessment, if any, would benefit, in terms of accounting for, and contributing to the improvement of, student performance, from the use of teacher-assessed tasks and teacher judgments of student performance within the framework of large-scale testing?

Professor Ball (2000a) in his review stated that the test formats were perceived as too narrow and broadening the test domain would allow for changes due to emergent English and Mathematics syllabuses. The Council project officers in testing and in curriculum development supported this statement, although the real impact of the new syllabuses on the Testing Program has yet to be fully investigated.

Broadening the curriculum coverage with the use of teacher-assessed tasks can be viewed as a matter of equity. A greater diversity of the student population could demonstrate their knowledge of literacy as broadening would allow multiple ways of showing performance as well as multiple opportunities for students to do it. In the current formats many students are inhibited from doing this, especially if they do not have test literacy as well.

To account for student learning both **additional** information (for example in aspects of literacy: speaking, being critical with texts, and school-based writing, or in aspects of numeracy: volume, mass, and pattern and algebra) and **deeper** information (for example in other curriculum literacies, use of multi-modal texts, and problem solving) would be gathered by broadening the coverage. To contribute to improvement of student performance, broadening the coverage to allow for consistency of teacher judgment processes would give teachers insights into students' results before the return of reports. This would assist teachers in preparing for the return of reports, and in planning for individual student learning during the school year rather than the following year which, in most schools, involves a different teacher.

Broadening curriculum coverage to include other aspects of literacy may allow for speaking, and critical interpretation of text assessments, through use of multimodal texts, open-ended responses, and school-based writing assessments. It could be difficult, however, to design ways to assess speaking and listening within the context of statewide testing, especially if the focus is on demonstrations of speaking in interactive rather than only oral presentation situations.

Broadening curriculum coverage to include other aspects of numeracy could encompass oral and written responses to numeracy problems, broader areas of measurement, and the use of concrete materials and hands-on activities. Broader tasks would incorporate more diagnostic features in the numeracy test. An analysis of responses would help to ascertain the 'level' at which the student was working, and assist teachers to further program learning activities for individual students.

There were questions still asked about who wants this information, and what will it be used for, which impacts on what is assessed. To contribute to *improvement* of student learning, broadening the coverage and allowing for teacher judgment may present gains for professional development, especially when student performance data is based upon new syllabuses. Through the testing procedure, teachers would develop understandings of the syllabus content, teaching and learning processes, and assessment techniques, as well as in testing processes.

5. Valued Literacy Practices and Multiliteracies

Research Question 4

What further valued literacy practices, including multiliteracies, might be assessed formally through teacher-assessed tasks within the framework of large-scale testing?

5.1 Multiliteracies and Valued Literacy Practices

Multiliteracies are about the growing significance of cultural and linguistic diversity and the influence of new communication technologies upon the social practices of using language (Kalantzis & Cope 1997). Today, English language users negotiate meanings every day in not only local communities, but in increasingly globally interconnected lives. English is a world language, a common language of global commerce, media and politics, and a language broken into multiple and differentiated versions. Migration, multiculturalism, global economics help deepen these differences, and will continue to make English even more diverse in time. Unsworth defines multiliteracies as the multidimensional, multiple literacies used to interact with computer technology, electronic and conventional images, distinctive literacy demands of different curriculum areas, and reproductive and critically reflective literacy practices (Unsworth 2001).

Within a multimodal meaning system (multiple ways of knowing and doing), five major areas constitute the concept of multiliteracies (The New London Group 1996). Termed modes of meaning, these are the linguistic, visual, audio, gestural, and spatial practices used to negotiate meanings with the world. These modes are always interrelated (multimodal) as in the meanings needed to create and understand drama performances, email, or desktop publishing. To assist the language learner, multiliteracies need an open-ended and flexible functional grammar; a metalanguage that describes and explains the patterns of meanings, and illustrates the differences embedded in cultural and linguistic situations and in the multimodal channels used to create and understand meaning.

Student performance in multiliteracy is an ability to understand and create patterns of meaning as products of different contexts - particularly in the changing contexts of communication technologies, and diverse social and intercultural contexts. Students use cognitive tools to make particular language choices within particular cultural and situational contexts. As each mode of meaning has elements that are negotiated with the use of these tools, student ability in utilising and critical understanding of the tools may be assessed. It would be difficult to assess the use of cognitive tools of multiliteracies if they are devoid of the authentic cultural and social situations in which they occur, or if they are embedded in the cultural and social contexts of a standardised test. The standardised test as a social practice is too narrow an instrument to assess full multiliterate performance, although it does test important aspects of literacy (Queensland School Curriculum Council 2000a).

Literacy practices of the test may, or may not, be valued in classrooms, in homes, or in workplaces. The test literacies become valued because of the emphasis placed upon them by the test clients or by governments. Teachers interviewed commented upon the need to know what writing genre was to be undertaken in the writing task, well before the test, and how to pre-teach aspects of viewing. They often felt that a broader coverage would mean more work in sensitising the students to the tests tasks. The literacy practices that are part of the test become the perceived valued practices in at least Year 3, 5 and 7 classrooms.

At present test literacies appear to be practices that are to do with comprehending print-based written and visual texts, writing on demand (one-off genre specific), error-free spelling, and use of Standard Australian English and its construction. Multiliteracies, by definition, could not encompass rigid literacy practices, or a demand for only one version of English. Aspects of speaking and listening are not tested at present, and even when student ability in understanding new technologies is tested, the test items are generally 'web site home page' reading stimulus materials. The test looks at linguistic modes of meaning, but attempts at testing visual literacies remain within a wholly linguistic dimension.

5.2 Cross-curricular numeracy and literacy

Aspects of multiliterate behaviour occur when literacy events are situated within other curriculum areas. These are the subject-specific literacies, or as Unsworth (2001) terms them, curriculum literacies (p. 10). Unsworth sees a need to look at curriculum literacies as the interface between a particular curriculum and its literacies, rather than imagining that there is a single literacy that could be spread across the curriculum (p. 11). Curriculum literacies are about comprehending and composing the modes of meaning unique to particular curriculum areas, and critiquing the perspectives on the way knowledges are constructed in that area. At present, collecting data on student multiliterate performance would be to look further at cross-curriculum literacy and numeracy practices.

The Council's position papers in literacy (Queensland School Curriculum Council 2000c, p. 5) and numeracy (Queensland School Curriculum Council 2000d, p. 8) provide examples of the practices that contemporary societies require of their members for effective literacy and numeracy. These practices are to do mostly with the way the literacy and numeracy codes work, with meanings, with social and cultural function and purpose, and with how users are influenced and positioned by texts, or mathematical problems or investigations. The position papers also describe the sorts of literacy and numeracy practices that occur within curriculum areas other than English.

The issue of what aspects of literacy the Testing Program could assess (eg basic skills, critical understandings) and in which key learning areas was of particular interest to the English curriculum development project team. It was thought that some valued aspects of literacy could not be assessed by point in time large-scale testing. The project team argued that if literacy is valued as a social practice within a multiliteracies agenda then it would be more appropriate to investigate the creation of a framework to support assessment of literacy for breadth, depth and complexity within an assessment program generally, before considering what role a large-scale testing program can play in this.

To assist in developing a framework, curriculum project officers of the Council from all key learning areas met to discuss the concept of multiliteracies. From their own perspectives of curriculum development, it was thought that multiliteracies needed to be centred on the climate where it was born, and as it was born out of globalisation of English, caused by commerce and technology, it has to be future oriented. It was agreed that multiliteracies is about diversity, access, power, and the multiplicity of discourses. It was understood that multiliteracies is dependent upon culture and has metalanguages and grammars to describe it.

The commonalities the curriculum teams saw in the literacy practices of each of their areas were the diverse metalanguages; the language to describe how people create meanings that are specific to a particular curriculum area may be similar. Project officers thought that they should 'name' the literacy behaviours, and the semiotic systems that are unique to their area. It was said that students could not be critical of the practices

unless they knew what the tools were by naming them. It was also thought that issues regarding multiliteracies were moving so fast and in hybrid and multimodal ways that it was difficult to name the practices and the tools, as they are ever-changing.

Once the multiliteracy concept is further understood items may be developed which could assess student performance in diverse literacy contexts, especially in the changing contexts of communication technologies. Administration of the test tasks would have to account for the more diverse social and cultural practices, and the technologies that constitute multiliteracies. Developing protocols, or criteria for marking and analysing tasks would have to recognise the different social and cultural contexts, the demands those contexts have on language choices, and the ways students use tools to negotiate meanings. Reporting back to students, teachers, parents and systems would also have to describe the social and cultural contexts of the test tasks as well as the diverse Englishes required for those contexts, and the social contexts of the test itself.

5.3 Conclusion

What other valued literacy practices, including multiliteracies, might be assessed formally through teacher-assessed tasks within the framework of large-scale testing?

Formal testing of multiliteracies or other valued literacy practices could prove difficult, even within the processes of teacher-assessed tasks and using consistency of teacher judgment strategies, or the context of key learning area-specific literacies. The newness of the concept of multiliteracies, the diverse social and cultural practices it involves, and its ever-evolving nature make it difficult to explain, or to name the tools which are used to make meanings. Certainly, use of only Standard Australian English would make it beyond the current Testing Program to assess other diverse Englishes.

6. Sample Materials

Research Question 5


What examples are available of teacher-assessed tasks and teacher judgment processes in other statewide or international testing programs? What samples of materials can be gathered?

6.1 Teacher-assessed Tasks in Large-scale Testing Programs

There is a variety of large-scale census testing programs that involve teachers assessing tasks both nationally and internationally. These range from teachers being trained centrally to mark test tasks to teachers marking the tasks their own students undertake (see table 1 for examples). Only representative samples of the types of teacher-assessed tasks from the United States of America are discussed here, while all census literacy and numeracy tests from Australian States and Territories are presented. Each is considered below, with an example, and an appraisal based upon the critical positioning from the literature review follows.

6.1.1 Indiana Statewide Testing for Educational Progress (ISTEP+)

The Indiana State Board of Education developed ISTEP+ (Indiana Education Department 2001). One component is a criterion-referenced applied skills assessment for Grades 3, 6, 8 and 10 in English and Mathematics. The applied skills test is marked by trained qualified 'readers', recruited publicly, but are teachers or those who hold a teaching degree. Readers must successfully complete a formal training program and are monitored heavily — unacceptable readers are either retrained or replaced. Figure 1 is an example of a task in Mathematics and the scoring rubric is seen in figure 2. Figure 3 shows a scoring exemplar. While not an extensive open-ended task example, it is representative of the sorts of teacher-assessed tasks that occur in ISTEP+.

II  Use your punchout coins to solve this problem.

Laura found 3 coins in her backpack. Two of the coins are the same and one is different. What is the LEAST amount of money the 3 coins can add up to?

You MUST show your work.

Answer _____ c

Figure 1
Teacher-Assessed Task Mathematics Grade 3 (Indiana Department of Education 2002a, p. 22)

Rubric:	
2 points	Exemplary response
1 point	Correct answer only OR Correct complete process; error in computation
0 points	Other

Figure 2
Scoring Rubric (Indiana Department of Education 2002b, p. 67)

Session 2—Item 11
Score Point 2

This response matches the exemplary response contained in the rubric. The student gives the correct answer of 7 cents and shows the correct work demonstrating the process used. The response receives a Score Point 2.

SCORE POINT 2

II Use your punchout coins to solve this problem.

Laura found 3 coins in her backpack. Two of the coins are the same and one is different. What is the LEAST amount of money the 3 coins can add up to?

You **MUST** show your work.

Answer 7 cents c

Figure 3
Scoring exemplar (Indiana Department of Education 2002b, p. 67)

ISEP+ applied skills test is an example of teachers assessing open-ended tasks. Criteria for marking are rigidly set, and markers must pass scoring examinations before they can be accredited. Teacher involvement is limited to marking only, and most of the teachers involved are not working in classrooms. There would be limited professional development for classroom teachers, and few links to effective pedagogy. There is little use of consistency processes, except formal training of markers.

6.1.2 Kentucky Department of Education Commonwealth Accountability Testing System (CATS)

The CATS (Kentucky Department of Education 2001) was developed through a collaborative process between teachers, parents and education advisors. Students are assessed in national, core content, writing portfolio, and demand writing tests. The writing portfolio is the only part of CATS that is teacher-assessed. The rest of the test, both open response and multiple-choice items, are scored by a test contractor for the state. Teachers, administrators, specialists and members of education organisations undertake assessment design and writing of item and scoring guides. The writing portfolio is assessed at the 4th, 7th and 12th grade. Teachers are trained to use an established scoring rubric (marking criteria) and random student portfolios are reviewed by the state for accuracy in scoring. Classroom teachers are encouraged to collaboratively mark the portfolios. Portfolio results are reported separately from the centrally marked demand writing tests.

Portfolios are a collection of a student's best writing. Students, with their teachers, choose pieces produced in their classes over a period of time. Grade 4 students make four selections of genre: reflective, personal, literary, and transactive. Grade 7 and 12 select the same as Grade 4, but make another selection on a genre of their own choice. Figure 4 shows part of the Parent Guidebook that explains the writing portfolio components and the marking criteria.

WHAT ARE THE REQUIRED PIECES IN THE 7TH GRADE WRITING PORTFOLIO?

The student includes a total of **5** pieces of writing in the portfolio. Any of the following portfolio entries may come from study areas other than English language arts, but a minimum of one piece of writing must come from another subject area.

- **Reflective Writing** in the form of
 - **Letter to the Reviewer** — discussing the student's growth as a writer and reflecting on pieces in the portfolio. *(Student must include one.)*
- **Personal Expressive Writing(s)** in the form of
 - **Personal Narrative** — focusing on one event in the life of the writer
 - **Memoir** — focusing on the relationship of the writer with a particular person, place, animal, or thing
 - **Personal Essay** — focusing on a central idea supported by a variety of incidents in the writer's life
(Student must include one or two.)
- **Literary Writing(s)** in the form of
 - **Short story** • **Poem** • **Script**
(Student must include one or two.)
- **Transactive Writing(s)** for a variety of authentic audiences and purposes in real-world forms (e.g., letter, article, editorial, proposal, brochure, review).
(Student must include one or two.)

In addition to the 5 pieces of writing, each portfolio must include the following:

- **Table of Contents**
- **Student Signature Sheet** — states ownership of the portfolio and may give permission to use the portfolio for training (optional)

The Scoring Guide

Unlike a grade of A or B, your child's score on the Writing Portfolio can give you information about the characteristics most often observed in your child's writing. When you and your child know what to look for, you also know what needs improvement. The *Kentucky Holistic Scoring Guide*, below, lists the qualities of effective writing under "Proficient," the goal for all Kentucky students.

NOVICE

- Limited awareness of audience and/or purpose
- Minimal idea development; limited and/or unrelated details
- Random and/or weak organization
- Incorrect and/or ineffective sentence structure
- Incorrect and/or ineffective language
- Errors in spelling, punctuation, and capitalization disproportionate to length and complexity of writing

APPRENTICE

- Some evidence of communicating with an audience for a specific purpose; some lapses in focus
- Unelaborated idea development; unelaborated and/or repetitious details
- Lapses in organization and/or coherence
- Simplistic and/or awkward sentence structure
- Simplistic and/or imprecise language
- Some errors in spelling, punctuation, and capitalization that do not interfere with communication

PROFICIENT

- Focused on a purpose; communicates with audience; evidence of voice and/or suitable tone
- Depth of idea development supported by elaborated, relevant details
- Logical, coherent organization
- Controlled and varied sentence structure
- Acceptable, effective language
- Few errors in spelling, punctuation, and capitalization relative to the length and complexity

DISTINGUISHED

- Establishes a purpose and maintains clear focus; strong awareness of audience; evidence of distinctive voice and/or appropriate tone
- Depth and complexity of ideas supported by rich, engaging, and/or pertinent details; evidence of analysis, reflection, insight
- Careful and/or subtle organization
- Variety in sentence structure and length enhances effect
- Precise and/or rich language
- Control of spelling, punctuation, and capitalization

Figure 4

Parent Guidebook on writing portfolio components and marking criteria (Kentucky Department of Education 2002, pp. 6 & 9)

Portfolio assessment is considered to be an effective technique in writing assessment (Freedman 1993). Its use in a large-scale test would contribute to teacher development and to effective writing pedagogy. Teachers are involved in most aspects of test design and in the marking of the portfolios with scoring rubrics. These processes would contribute to teacher development. As teachers are encouraged to share their marking with others, only some consistency processes are used, but it is not known how this sharing contributes to moderating the assessments. As schools are rewarded for increased student performance the stakes for CATS are reasonably high, and if

classroom teachers are involved in selecting samples of writing, and in marking the samples, it could be difficult for them not to offer intervention, or 'mark up' the work. As the portfolio results are reported separately any differences between these scores and writing demand, as they comprise two different writing situations, may be explained, but this issue could be problematic from an accountability position.

6.1.3 Maryland School Performance Assessment Program (MSPAP)

MSPAP (Maryland Department of Education 2001) consists of criterion-referenced performance tests in reading, mathematics, writing, language usage, science and social studies for students in grades 3, 5 and 8. Tests are based on learning outcomes developed by Maryland educators. These specify what students should know and be able to do. The tests emphasise higher order skills to solve problems, make decisions and understand information. It utilises short and extended response items and individual and group performance tasks.

Teachers write the tasks (approximately 140) and mark the tests using state-developed rubrics. Approximately 650 teachers are involved in marking 185 000 tests. All answer books, for a given grade and cluster, are marked at the same time, and at different sites around the state. Scorers mark the open-ended responses and assign a score point on a scan sheet. Quality of scorers' marks is maintained by check sets, accuracy sets, spot checks, and retraining. Figure 5 is an example of part of a task for Grade 3 on 'Deserts'. This part involves writing to persuade and is at the end of a five-day test sequence. All test tasks are related to the theme, and test performance in reading, language use and Social Studies as well as writing. Figure 6 shows the marking criteria and two examples of scored work.

Wednesday, Task 1
Title: Deserts

WRITING PROMPT: WRITING TO PERSUADE

You have heard of someone who is thinking about traveling across the Sahara Desert, the way Geoffrey Moorhouse did. However, this person is not sure whether to take the trip. Write a letter to the traveler to persuade him or her either to go on the trip or to stay home. You may use information from your reading to help support your point of view.

PRE-WRITING

- Think about the problems Geoffrey Moorhouse had on his trip.
- Think about things another person could do differently today to avoid those problems.
- Think about whether or not the traveler should take this trip.

As you write, you may try making a list, a web, or a diagram on the lined paper provided to arrange the ideas you want to share in your letter.

DRAFTING

Use your ideas as you write a first draft of your letter on the lined paper. You will have 30 minutes to plan and to write your first draft.

Thursday, Task 1

Title: Deserts

REVISING

Yesterday you wrote a first draft of a letter. Today you will take 5 minutes to read your draft and think about what you have written. Imagine that you are the traveler reading the letter. Think about the answers to the questions below.

1. Does the letter persuade the traveler to do what is best?
2. Does the letter give reasons that support that advice?
3. Does the letter make sense?

After you have thought about how well your letter answers these questions, you will get some ideas from a partner to help improve your writing.

Thursday, Task 1

Title: Deserts

PEER RESPONSE

You have had the chance to ask yourself questions about how well you have composed your writing. In order to determine if your writing says what you want it to say, it is usually helpful to get someone else to react to your writing. This is called "peer response." You will work with your partner to do your peer response. Your Peer Response Form is on page 36 of your Answer Book.

1. Decide with your partner who will go first.
2. Follow the instructions on the Peer Response Form, and be sure to allow enough time for both of you to read and take notes about the answers to the questions.

Figure 5

Example of part of a task for Grade 3 on 'Deserts' (Maryland State Department of Education 1996a, pp. 30, 34 & 35)

WRITING PROMPT: WRITING TO PERSUADE

You have heard of someone who is thinking about traveling across the Sahara Desert, the way Geoffrey Moorhouse did. However, this person is not sure whether to take the trip. Write a letter to the traveler to persuade him or her either to go on the trip or to stay home. You may use information from your reading to help support your point of view.

0 points

- 0 **Development:** The writer identifies an ambiguous position with little or no relevant personal and/or factual information to support that position; or, the writer fails to identify a position.
- 0 **Organization:** The writer presents an argument that is illogical and/or minimally maintained.
- 0 **Attention to Audience:** The writer does not address the needs and characteristics of the identified audience.
- 0 **Language:** The writer seldom, if ever, uses language choices to enhance the text.

Dear, Ola,
If you go on the trip, you can get lost. You might miss your family or pets. You might get sick.

Score = 0

1 Dear explorers
I do not want to go because I do not want to get in trouble so I'm going to stay home. I read a little paragraph about you maybe meet you but explorers were in it. It must be a good job to be an explorer and hard job to and finish. Long does it take to be an explorer? Was it fun? Is there? I wish I could help you but I'm not going to go but I would like to but I'm not.

Score = 0

Confusing, irrelevant information that is not persuasive.

Figure 6
Marking criteria and two examples of scored work (Maryland State Department of Education 1996b, pp. 19 & 20)

Maryland encourages 'teaching to the test' as the test items are about higher order thinking skills (Maryland Department of Education 2001 — What is MSPAP?, p. 1). Teacher involvement is high, with strong links to pedagogy — especially when teachers are encouraged to teach students how to respond to broader response tasks. MSPAP contributes to professional development because of the large numbers of teachers involved in task development and marking. Consistency issues are dealt with through training of markers, and checking strategies. Maryland has received some publicity recently over its Testing Program (Center for Education Reform Newswire 2002). Some sites' test marks indicated that schools in the area were not able to show required improvements. Parent groups in Maryland have criticised the tests and asked for their suspension while a tool could be found that will help the state comply with the Federal 'No Child Left Behind' Act (House Education and the Workforce Committee 2001). This is an indication of the American large-scale test discourses, and the high stakes that accompany the test processes.

6.1.4 Vermont Statewide Assessment System

Vermont assessments (Vermont Department of Education 2001) include the Vermont Developmental Reading Assessment (DRTA), Written Language Portfolio (WLPA) and Mathematics Problem Solving and Communication Portfolio (MPSCPA) assessments that are teacher-assessed. All results from these tests are reported separately from the centrally computer-marked tests.

The DRTA is a standards-based assessment in reading. It is administered to all Grade 2 students. Teachers mark oral reading for accuracy and retelling for comprehension, and results are analysed centrally. It is reported back to schools the percentage of students who score in the highest two levels — *achieved the standard* and *achieved the standard with honours*. DRTA was adapted from the original Developmental Reading Assessment published by Celebration Press.

The WLPA is a standards-based assessment administered to students in Grades 5 and 8. Students prepare six pieces of writing that have gone through the entire process of draft to final edit. Teachers assess the portfolios using a rubric which scores writing on a two-point scale. The six pieces are about a response to literature, a report/expository piece, a narrative, a procedural piece, a persuasive argument and a personal essay.

The MPSCPA, aligned to state framework of standards, is administered to all students in Grades 4, 8 and 10. The portfolio is a compilation of the students' best problem-solving work on assigned, complex tasks. In addition to maths work, students are asked to describe how they approached the problem. Teachers mark the portfolio by rating seven areas for Grades 4 and 8, and five areas for Grade 10, on a six-point scale. Areas assessed are approach and reasoning, connections, accuracy of the solution, mathematical language, representation and documentation in Grades 4 and 8, and approach and reasoning, execution, observation and extensions, mathematical communication and presentation in Grade 10. Sample portfolios are selected for re-scoring for state data, but schools use the locally scored portfolios for their school reports.

Figure 7 shows a marked task from the Mathematics portfolio, while figure 8 gives some of the criteria used for marking. The marked task shows that the student has identified an underlying mathematical concept — Level 2 'connections' criteria — second bullet point.

Identified an underlying mathematical concept of pattern in her/his solution

Connections

Level 2 Bullet 2

The student identified the pattern that the days decrease by .25" when he/she states "the pattern is subtract .25" each time". The student also comments that the amount 2.166666 will be 2.2 yards.

Zeno The Xylophone Maker

There are 12 keys on a xylophone (C, D, E, F, G, A, B, C, D, E, F, G).

The first key (C) is 7.5 inches long.

The second key (D) is 7.25 inches long.

The third key (E) is 7 inches long and so on continuing this pattern is also.

Zeno the xylophone maker was about to begin construction on a new xylophone when he realized he had run out of wood striping. This wood striping comes in 1 yard pieces. How many 1 yard pieces does Zeno need to buy in order to make 1 xylophone.

Key	Length in inches
C	7.50
D	7.25
E	7.00
F	6.75
G	6.50
A	6.25
B	6.00
C	5.75
D	5.50
E	5.25
F	5.00
G	4.75

Handwritten notes:
 The pattern is subtract .25" each time.
 This is an arithmetic sequence.
 This is 9 original keys.
 2.166666... = 2.2 yards
 2.2 yards = 2.2 * 36 = 79.2 inches
 79.2 / 7.5 = 10.56 → 11 keys
 11 keys * 7.5 = 82.5 inches
 82.5 / 36 = 2.291666... → 3 yards

Figure 7
 Example of a marked problem-solving task from Mathematics Problem-Solving and Communication Portfolio

Problem Solving Criteria			
Approach and Reasoning			
START HERE			
Level 1	Level 2	Level 3	Level 4
<ul style="list-style-type: none"> - Approach wouldn't work or - No approach evident 	<ul style="list-style-type: none"> - Approach would lead to solving only part of the problem¹ or reaching a partial solution or - Approach would work but there is some flaw in the reasoning 	<ul style="list-style-type: none"> - Approach worked or would work for solving the problem, and reasoning, if evident, is not flawed <p><i>(Note: Use of a formula is an approach that worked or would work.)</i></p>	<ul style="list-style-type: none"> - Approach worked, and at least one of the following 3 additional aspects of good problem solving is evident: <ul style="list-style-type: none"> - Justifying the application of a known formula or rule used to solve all or part of the problem or - Making a formula or rule used to solve all or part of the problem or - Describing verification of her/his solution²
Connections			
START HERE			
Level 1	Level 2	Level 3	Level 4
<ul style="list-style-type: none"> - Response stopped without including a mathematically relevant observation with respect to her/his solution 	<ul style="list-style-type: none"> - Made a mathematically relevant observation about her/his solution or - Identified an underlying mathematical concept or pattern in her/his solution or - Solved the problem and (s)he recreated³ the problem and found a new solution or - Solved the problem and (s)he used a different mathematical process to solve the same problem 	<ul style="list-style-type: none"> - Related this problem to a similar problem(s) to a real world phenomenon by expressing the mathematical relationship(s) or - Analyzed the relationship among elements in her/his solution (e.g. among similar or different mathematical topics in her/his solution) or - Tested and accepted and/or rejected an hypothesis or conjecture about her/his solution or - Identified a formula or rule, while solving the problem, that worked or would work in solving all or part of that problem 	<ul style="list-style-type: none"> - Solved the problem, discovered a general rule⁴ about the solution⁵, and demonstrated understanding of the generalization either through explanation of the derivation, or through application to more than one other case or - Solved the problem, and (s)he evaluated her/his solution in a more complicated situation or - Evaluated the reasonableness or significance of her/his solution
Solution			
START HERE			
Level 1	Level 2	Level 3	
<ul style="list-style-type: none"> - No work is present or - No part of the solution⁶ is correct or - Some work is present, but the work doesn't support the answer given 	<ul style="list-style-type: none"> - The solution⁶ is correct for only part of the problem, and there is work to support those correct parts) or - The solution⁶ contains mathematical errors which lead to an incomplete or incorrect answer 	<ul style="list-style-type: none"> - The answer is correct, and the work in the solution⁶ supports the answer 	
<p>¹ Would: An approach that would work for solving the problem addresses all aspects of the mathematical situation presented in the task. An approach that would work may contain mathematical errors, an incorrect solution, or may be incomplete.</p> <p>² Part of the Problem: Within a problem, there may be several mathematical components that need to be addressed, or there may be multi-parts, if not all of the mathematical components of the problem are addressed, or not all of the parts of the problem are addressed, then the student only found an approach to solve part of the problem.</p> <p>³ Solution: All of the work that was done to solve the problem, including the answer.</p> <p>⁴ Recreated: The student substituted different numbers in the same problem and found another solution, or used the same procedure in a different circumstance.</p> <p>⁵ General Rule: A rule that can be used no matter what the numbers in the problem are, either expressed in algebraic notation or in words.</p>			
September 1997			

Figure 8
Some marking criteria for Mathematics Problem-Solving and Communication Portfolio Tasks

Fairtest (2001) states that Vermont has nearly a model testing system. The assessment burden is reasonable as are the stakes. Teacher involvement is high due to the local marking of the three tests. Teachers are trained to use the portfolio rubrics. Marking meetings are held across the state twice yearly. Fairtest states that the portfolios were also intended to improve teaching. Independent reviews by the RAND Corporation confirmed that this intent was met; pedagogy links are high. Other grade areas include portfolio assessments, but as the program is comprehensive, selection of writing samples and problem-solving activities may dominate the curriculum in the years the test is administered. To assist with consistency, teachers are encouraged to collaboratively mark the portfolios, but Fairtest has stated that reliability of the assessments will need


continued attention. To help in this, the Vermont Department of Education has reduced the marking criteria from five points to three.

6.1.5 Australian Capital Territory Assessment Program 2001

The speaking component (Australian Capital Territory Department of Education and Community Services 2001) is the only component of the ACT tests in Literacy and Numeracy that is teacher-assessed, and only for Grades 3 and 5. Teachers assess on a four-point scale for content and performance — Grade 3, 1–4; Grade 5, 2–5. ACT provides a video for professional development when marking students' talk. Teachers introduce the task, and students talk about the task with a partner — instructions also include reminders about formal speaking. Figure 9 shows these instructions. Each student in the class then speaks for one minute while the teacher scores from a marking guide. The second task involves rehearsal time before the formal talk. Grade 5 are asked to prepare the talk before rehearsal with a partner. Palm cards and prompts are encouraged. Teachers are encouraged to mark in pairs, but this is not mandatory.

YEAR 3 - SPEAKING

ACT Assessment Program 2001 Year 3 SPEAKING



TASK 1: THE BEST TIME I EVER HAD

Speaking on the topic *The Best Time I ever had* involves three steps:

- Introduction
- Interview with a partner
- Individual presentations.

REQUIREMENTS

Time
approximately 1 to 1.5 hours

Materials
copies of *The Best Time I ever had* stimulus sheet (one for each student)

Students will need
adequate space to work with a partner
one *The Best Time I ever had* stimulus sheet for each student
pencil
palm-sized paper or card

For marking teachers will need
2B pencil
Marking Guide
Student Record sheets for 2001 ACT Assessment Program Year 3

Figure 9
An example of speaking task administration requirements

**ACT ASSESSMENT PROGRAM
2001 SPEAKING
STUDENT RECORD SHEET
YEAR 3**

INSTRUCTIONS:

- Use pencil only, preferably 2B
- Do NOT use any other ink color
- Please mark as fully
- Mark to show marks

← **CONFIDENTIAL!**

Please MARK LIKE THIS:

STUDENT ID: STUDENT NAME:

SCHOOL NAME:

TASK 1			TASK 2		
NON-PARTICIPATION	CONTENT	PERFORMANCE	NON-PARTICIPATION	CONTENT	PERFORMANCE
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 10
Student record sheet and criteria

For this component the classroom teacher is the task administrator and the task marker. Figure 10 shows the record sheet and marking criteria. There are some pedagogy links as the test is close to real-life public speaking situations and strategies. There is a small amount of Professional Development through a video to support marking requirements, criteria to mark speaking on content and performance for a range of spoken texts, and room for moderation. Consistency is achieved through clear instructions, criteria understanding, practice marking with video support, and marking in pairs with moderation processes only encouraged. Speaking assessment does not contribute to National Benchmark data.

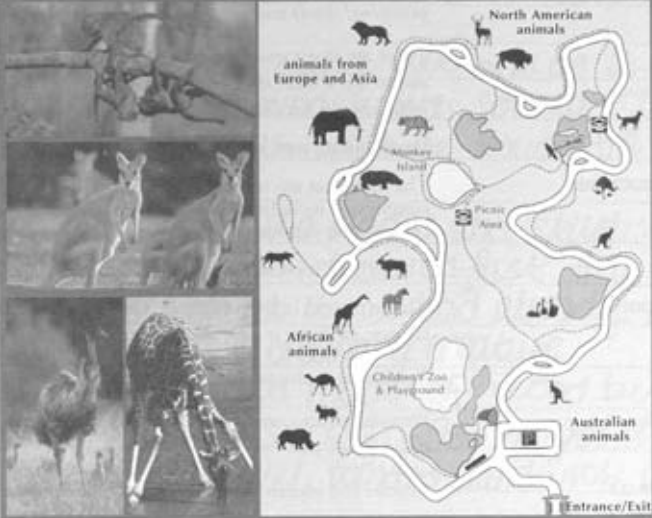
6.1.6 New South Wales

The Primary Writing Assessment (PWA) is administered to Years 3 (New South Wales Department of Education and Training 2001a) and 5 (2001b) students, at around the same time as the centrally marked multiple-choice Basic Skills Test. Students complete two writing tasks — one literary, one factual. Teachers mark these using set criteria for each task. The English Language and Learning Assessment (ELLA) is administered to Years 7 and 8 (New South Wales Department of Education and Training 2001, School Assessment and Reporting Unit 2001). The writing task for this program is teacher-assessed. The extended response task for the Statewide Numeracy Assessment Program (SNAP) (New South Wales Department of Education and Training 2001c) for Years 7 and 8 is teacher-assessed. Teachers who nominate for ‘in school’ markers receive two days of classroom release — one for training (after administration of the test) and one for marking the tasks. Figure 11 shows an example of the teacher-assessed writing task with figure 12 showing marking criteria.

D. MARKING WRITING TASK TWO: ZOO EXCURSION

Writing Task Two

Below is a map of the Western Plains Zoo. Visitors to the zoo can ride bikes, drive or walk around the zoo and look at the animals. The animals are not kept in cages. They can move around freely at a safe distance from the people.



Imagine your class went on an excursion to this zoo.
Write a recount of the excursion for the school newsletter.
Use the photographs and the map to give you some ideas about the things you might have seen.

- You should begin your recount of the excursion with an introduction.
- You should include details about the things you saw.
- You should write in sentences and use paragraphs to help organise your writing.
- You should pay attention to spelling and punctuation.

4

Figure 11
Example of the PWA teacher-assessed writing task

1. TASK LINKS TO THE CURRICULUM

Text purpose – to recount a school excursion

This task requires students to produce a piece of writing that recounts an excursion for the school newsletter. Students are asked to provide an introduction for readers, and then describe what they saw on the excursion. The purpose is for students to write about their observations on a school-based excursion to members of the school community.

Documenting educational experiences and events is part of all KLAs. The focus of the assessment of students responses is on their ability to present a logical organised text with elaborations about what they observed.

2. TASK-SPECIFIC MARKING CRITERIA

The specific marking criteria, and marks to be allocated for each criterion in Writing Task Two, are represented by the columns marked 1 to 15 on page 3 of the test books. These are identified below and expanded with examples in the table which follows.

Non-Attempt	TP										WL						
<input type="radio"/>	1	2	3a	3b	3c	4	5	6	7	8	9	10	11	12	13	14	15
	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯	⑰
	①	①	①	①	①	①	①	①	①	①	①	①	①	①	①	①	①
M																	
S																	

Writing Task Two — Zoo Excursion

Text	1	=	Text Function
Processes	2	=	Theme
	3a	=	Text Structure - Orientation
	3b	=	Text Structure - Elaboration
	3c	=	Text Structure - Conclusion
Text	4	=	Paragraphs
Features	5	=	Pronouns
	6	=	Conjunctions
	7	=	Sentence Structure
	8	=	Tense
Sentence	9	=	Clause Pattern
Level	10	=	Agreement
	11	=	Prepositions
	12	=	Articles/Plurals
	13	=	Punctuation
Word Level	14	=	Verb Form
	15	=	Spelling

Figure 12
PWA writing task marking criteria

Teachers are formally trained to mark the tasks, which are assessed against criteria. There are pedagogy links as the tasks are both closed and open-ended, and linked to syllabus documents. It is suggested that from the test results, teachers can adjust teaching programs to meet the needs of individual students. Professional development occurs through training of markers, and through their subsequent conversations upon returning to schools. Consistency is only achieved through formal training and markers being able to contact a coordinator by phone during the marking process. Only PWA and ELLA are used to contribute to national benchmark data.

6.1.7 Northern Territory Multi-level Assessment Program

The reading, spelling and numeracy tests are teacher-assessed. The Common Writing Task (CWT) is assessed centrally (Northern Territory Government 2001). Trained teachers mark the common writing task, and the criteria for marking are published in the administration guide. Teachers can mark their own students' work; complete the optional marking table on the cover of the test booklet, before returning scripts for central marking. The marking procedure is to assign a numerical score (based upon learning outcomes) against criteria of subject matter; ideas and vocabulary, and textual features; generic structure, cohesion, punctuation and spelling. Figure 13 shows the optional marking table on the tear-off front page of the student test booklet.

Reading item responses are reasonably precise. Students work through multi-levelled reading stimulus, recording responses using a variety of response techniques. Reading material is levelled against NT assessment profiles, and scores indicate which level students are working within. Descriptors explain reasons for each response in terms of learning outcomes. Numeracy responses are varied, but reasonably precise. Teachers mark with a key for right or wrong responses. Items are explained by referring to syllabus document page numbers. Scores indicate which level within which students are working.

The Northern Territory Board of Studies
Year 3 030211
Literacy Tasks, MAP 2001

Student Background Information
(To be completed for every student including those who did not do the test)

Teacher's use only

School Code
Student No
ESL
Absent
Exempt
Special Needs

School Name _____
Student Name _____
Outstation Name (if applicable) _____
Date of Birth _____
Day Month Year

1. Are you a boy or girl? Boy Girl
2. Are you an Aboriginal or Torres Strait Islander? Yes No
3. Does everyone at home speak to you in English? Yes No
4. How often do you speak English at home? Never Sometimes Usually Always
5. What is the main language spoken at home? _____
6. In which country were you born? _____
7. In which country was your father born? _____
8. In which country was your mother born? _____
9. Do you need English language support in an ESL program? Yes No

Teacher's use only

Please tick if this is a special needs student	<input type="checkbox"/>		
	Writing	Reading	Spelling
Please tick if this student was absent	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Please tick if this student is exempt	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Please attach explanation for exemptions			

OPTIONAL Common Writing Task: Teacher's Assessment

Subject Matter	Textual Features	Spelling
Quality of Ideas <input type="checkbox"/>	Cohesion <input type="checkbox"/>	Spelling in Context <input type="checkbox"/>
Vocabulary <input type="checkbox"/>	Punctuation <input type="checkbox"/>	
	Generic Structure <input type="checkbox"/>	

Based on school program and assessment measures indicate where on the continuum this student is currently working in Reading and Writing.

English (NT Profile) Reading Level
Writing Level
ESL (NT Profile) Reading Level
Writing Level
First Steps Reading
Writing
English NT Curriculum Framework Band Reading
Writing
Other Reading
Writing

Northern Territory Government

Figure 13
Front page of the NTAP student Year 3 literacy test booklet showing marking table for the optional common writing task teacher's assessment

Teachers would have to engage with marking criteria if taking the option to mark the CWT. They would have to indicate where the student is currently working within school programs based upon NT profile levels, First Step phases, or NT curriculum bands. There are obvious links here to NT syllabus, which indicates the program provides some positive professional development. Teachers are encouraged to let students know the results as soon as they have completed marking, and to identify beneficial teaching points for the remainder of the school year. There was little evidence of consistency or moderation processes, except for contacting a relevant government officer for further explanation. The CWT is double marked, where both marks and the teacher optional mark are not disclosed. A moderator rectifies any discrepancies. There is no data available regarding correlation between teacher and marker assessment, but anecdotal evidence suggests it is high. In 2002, the Northern Territory Government is offering

professional development for teachers in marking the writing task. All tests contribute to national benchmark data.

6.1.8 South Australia

All tests are centrally marked in South Australia, but they are investigating using teacher assessments and judgment. In 2001 South Australia published a product termed Calibrated Assessment Tasks (Out of Print). This was designed for teachers to work individually with assessment tasks, and then to trial their own judgments against criteria and exemplars in the package. This has been discontinued as the tasks were limited, their use was dubious, and there was little durable change to teaching practice or assessments that could have been collected. This year South Australia will trial a process of professional development where teachers will come together to discuss assessment tasks, assessment criteria against the Curriculum Frameworks and Accountability Framework. They will trial the tasks and meet for moderation processes. This trial will be documented as a possible path for State collection of data on student performance using broader tasks and teacher judgments.

6.1.9 Victoria Achievement Improvement Monitor (AIM)

Teacher-assessed writing and mathematics occurs as part of the AIM program in years 3 and 5 (Victorian Curriculum and Assessment Authority 2000). Students write on a topic that fits in with normal classroom activities, or selected from three that have stimulus material supplied. Students discuss the stimulus, prepare a draft (using all resources in the classroom, including discussion with peers) and revise their work. Final versions are written individually and within a set time. Maths involves students performing applied mathematics tasks, with a timed teacher demonstration and timed task completion.

Writing is marked against criteria that relate to the Victorian outcomes approach Curriculum Standards and Framework support documents and accompanying elaborations. Teachers are given exemplars to explain criteria, and trial pieces to practise marking. Teachers assess mathematics by marking responses with a numerical score if responses are right or wrong. Maths tasks are hands on; encompass explanation and a variety of responses can be given to the tasks. Figure 14 gives an example of teacher-assessed mathematics tasks.

2. Look at the spinner:

If you were to spin a paperclip on this spinner, which number would it be most likely to land on?

Use the spinner with the paperclip as your teacher showed you. Spin it 20 times and record your result for each spin.

Number	Tally	Total
1		
2		
3		
4		
5		
6		

(b) Draw a bar graph (column graph) of your tally table.

(c) A pizza shop sells several different types of pizzas. Below is a list of the most popular types of pizza and how many of each type were sold in one week.

(i) Round each total to the nearest 10 units and write this value in the rounded amount column.

Type of pizza	Number sold in one week	Rounded amount
Vegetarian	80	
Supreme	187	
Hawaiian	118	
Mexican	106	
Chicken	95	

(ii) Using the rounded amounts and the key given, draw a pictograph of the number of each type of pizza sold. Please be sure to put a title on the graph and label each axis.

(d) What number did your paperclip land on most often? _____

(e) Would you expect this result? _____

Why? _____

STOP YEAR THREE STOP YEAR FIVE CONTINUE WORKING.

Teacher Assessment

Figure 14
Example of AIM teacher-assessed mathematics tasks

Writing prompts are a little lacking in breadth, purpose or audience requirements. Because of the 'process' focus, this task links to writing pedagogy. As the marking criteria are related to Victoria Curriculum and Standards Frameworks documents, this would assist in teacher professional development. Consistency in marking would be an issue, as moderation is not overemphasised. The purpose was for the teacher-assessed tasks could also be questioned, as the tasks are often too similar to standardised components.

6.2 Appraisal Conclusions

Discussed earlier, four stances were drawn from the literature, which were used to form a critical position to critique the samples of test materials and the above testing programs. Questions, based upon the stances, were asked of the materials and programs to see what they had to offer. Questions asked were:

- what kinds of teacher involvement are there in the program
- what are the implications the program has for pedagogy
- are teachers professionally developed by the program, and
- what consistency processes are used to assist in validity and reliability of marking?

There was a variety of ways teachers were involved in these tests. They ranged from teachers being formally trained to mark tasks, to teachers being involved in planning and developing tasks and marking tasks that their own students undertook. In programs where the stakes were higher, as in budgetary rewards for test results, formal training of teachers as markers resulted. Table 1 shows some of these trends.

Positive pedagogy links were reported when testing programs were related to local syllabus documents, effective assessment practices, or if the results could directly affect changes to teaching and learning programs for individual students. A number of the programs were about testing to outcome statements or performance standards, with marking criteria developed from these. 'Teaching to the test' was encouraged in these instances.

Professional development occurred when teachers were involved in task writing or marking. If teachers were only involved in administration of the test, gains for professional development was dubious. Where there was not a positive pedagogy link, professional development was limited to understanding only the test processes, which in itself, is helpful for teachers in understanding the Testing Program. The test assisted further in teacher development of effective assessment practices if these practices were valued in the Testing Program.

There were very few moderation practices occurring with the tests that were reviewed. Often there were chances for teachers to meet and share the marking process, but these were only encouraged, and not seen as vital to the process, nor was it stated how this sharing contributed to moderating the marks. Mostly, training of markers, especially when perceived test stakes were higher, was the way consistency was realised. In the collected sample it appeared that the more reliant the test stakeholders are in valid, reliable and objective assessments (as when performance data is used to report to National Benchmarks), the more test contractors turn to formal training of markers and central task construction (see table 1).

6.3 Conclusion

What examples are available of teacher-assessed tasks and teacher judgment processes in other statewide or international testing programs? What samples of materials can be gathered for analysis?

There are a variety of ways in which teachers assess test tasks beyond central marking. More extended test responses need more extensive marking processes and some of these include teachers marking their own students' tasks. While Indiana uses central marking processes, Maryland uses local marking of their multiple task and cross-curricular tests at different sites around the state. Both train markers rigidly in understanding marking rubrics and marking processes. Kentucky and Vermont include classroom teachers marking their own students for writing portfolio in Kentucky and writing and Mathematics portfolio in Vermont. Year-level teachers concerned attend out of school sessions in training to use assessment rubrics and criteria. Portfolio results are reported separately from writing demand, or in the case of Vermont, Mathematics multiple choice and short answer, test results.

Australian States also show a similar variety of teacher-assessed tasks. Australian Capital Territory employs classroom teacher assessment for their speaking test, as does Victoria in a writing and Mathematics tasks. Both present teachers with marking criteria and exemplars, in video and print format respectively. New South Wales uses local site marking for their teacher-assessed tests in writing and extended Mathematics response tests, while the Northern Territory offers teachers an option to mark their demand writing task before it is centrally marked. Western Australia uses teacher-assessed tasks in their sample tests of English and Mathematics, but their literacy and numeracy census tests are machine scanned. Tasmania is not using teacher-assessed tasks at the present time and South Australia is trialling teacher-assessed tasks along with processes of consistency of teacher judgment as a way to collect statewide performance data. All use standardisation processes and training of teachers to enhance validity and

reliability of test marking. There was little use of consistency processes in as far as assessments being moderated through formal procedures or informally through discussion with other markers.

Sample teacher-assessed tasks in testing programs from the USA offer the Queensland Testing Program some avenues to explore for the broadening of the program. As Maryland's cohort is similar to Queensland's the processes used to administer and assess their open-ended tasks could be worthy of investigation. Portfolio assessments similar to Kentucky and Vermont could be adapted for trial in Queensland to include extended response literacy and numeracy tasks related to Queensland syllabuses. The Speaking Test from Australian Capital Territory could also be adapted for the Queensland testing context, as could the processes Victoria uses for teacher-assessed extended mathematics response and teacher-assessed writing tasks.

The processes New South Wales use to locally mark their tests could be further investigated for their use in local site marking. Teacher assessment of demand writing tasks resembling the Northern Territory processes could also be used in Queensland. While not broadening the coverage, both would allow for more teacher involvement in the Queensland Testing Program.

7. Concluding Statements

This research project was only able to touch the surface of the issues of teacher-assessed tasks and using teacher judgment in the Queensland Testing Program as one way to broaden the literacy and numeracy curriculum coverage. Investigation of computer adaptive testing, beyond the parameters of this research, could also broaden the curriculum coverage but teachers may not be involved in that assessment to a large degree.

Teacher-assessed tasks are a way in which the curriculum coverage of the existing Years 3, 5 and 7 tests could be broadened. Such tasks might involve either formal training of teachers in marking to set criteria, or teachers marking their own students' tasks. Using processes to develop consistency of teacher judgment strategies could be a path to follow if teachers are involved in assessing the test tasks of their own students.

Test items or assessment tasks that require more open-ended or extended responses would broaden and deepen the existing curriculum coverage of the Testing Program. A wider range of literacy and numeracy test tasks would allow for additional or deeper information of student performance in those areas.

Broadening the curriculum coverage may also challenge the purposes of the Testing Program. The impact this would have on the balance between accounting for, and contributing to, the improvement of student literacy and numeracy learning would have to be further investigated, especially as a change in the purpose of the program may raise issues related to sample or census testing. The decision regarding developing teacher-assessed tasks for sample tests would be more to do with the recipients of the results and what they do with them. The school authorities would need to be consulted regarding this. If the teacher-assessed task were to add to or enhance existing data, then sample tests would be appropriate. In addition, this would need clear communication of the intent of the Testing Program to stakeholders, parent bodies and the wider education community.

Improved accountability, in terms of additional and deeper information being gathered, and improved student performances, in terms of teachers being more able to use the test results to plan for further student learning, would occur from teachers being more involved. The test would be seen more as an adjunct to classroom assessment than at present and used by teachers to plan for improved learning for their students by adding to existing classroom assessment data, instead of being seen as isolated from it.

The more the test items relate to newer syllabus documents, or their marking encourages effective assessment practices, the more teacher professional development occurs. Broadening the curriculum coverage begs a question of the relationship between the Testing Program in aspects of literacy and numeracy and the developing Years 1 to 10 English and Mathematics syllabuses.

More teacher involvement with the tests would not only increase workloads for Years 3, 5 and 7 teachers but also the Testing Program budget. This increase would not be valued highly if appropriate teacher release for training and marking was not available, or if the benefits for improving student learning were not obvious, especially with parent unease about teacher absence from the classroom.

Appendix A – Interview Schedule A

The following schedule was used for interviews with the Council project officers in testing and curriculum development. This schedule was used as a base to develop a shared interview text — other questions arose as these issues were discussed.

Issues to think about

Broadening the Curriculum of the Testing Program is about gathering more authentic, and contextual information about student learning for the purposes for accountability and improvement of student performance.

Question: How could the Testing Program enable more authentic and contextual gathering of information?

The literature suggests that if Testing Programs are more aligned to curriculum development and assessment practices, there are spin-offs for professional development.

Question: Can the Testing Program be more aligned?

Once assessment becomes high-stakes (certification, early years intervention, accountability, distribution of funding, and benchmarking) reliability (consistency) and validity (authenticity) of results become major concerns, especially in teacher-assessed tasks.

Question: Can we gather more authentic information, and remain valid and reliable?

Some valued literacy practices, and multiliteracies are unable to be assessed with the Testing Program in its present format.

Question: Can we assess valued literacy practices and multiliteracies?

Appendix B – Interview Schedule B

The following schedule was used to interview teachers and school administrators. This schedule was used as a base to develop a shared interview text — other questions arose as these issues were discussed.

Issues to think about

Broadening the Curriculum of the Testing Program is about gathering more authentic, and contextual information about student learning for the purposes for accountability and improvement of student performance.

Question: What sorts of information do you think the Testing Program could gather if it was broadened?

The literature suggests that if Testing Programs are more aligned to curriculum development and assessment practices, there are spin-offs for professional development.

Question: Can the Testing Program be more aligned to the assessment practices in this school?

Once assessment becomes high-stakes (certification, early years intervention, accountability, distribution of funding, and benchmarking) reliability (consistency) and validity (authenticity) of results become major concerns, especially in teacher-assessed tasks.

Question: How could we ensure teacher-assessed tasks would be reliably assessed?

Some valued literacy practices, and multiliteracies are unable to be assessed with the Testing Program in its present format.

Question: How can we assess more valued literacy practices and multiliteracies?

Bibliography

- Anastasi, A. 1997, *Psychological Testing*, Macmillan, New York.
- Archdiocese of Brisbane, Catholic Education. 1997, *Religious Education — a curriculum profile for Catholic schools*, Brisbane: Catholic Education Office.
- 1996–1999 *Curriculum Updates*, Brisbane: Catholic Education Office.
- 2002 March, Curriculum Update No 49, *Report on the Implementation of the 1997 Religious Education Guidelines — Summary of Research Findings*, Brisbane: Catholic Education Office.
- Australian Capital Territory Department of Education and Community Services, 2001, *ACT Assessment Program 2001 Year 5 Speaking: Administration Guide*, Author.
- Barton, P., 1999, *Too Much Testing of the Wrong Kind; Too Little of the Right Kind in K-12 Education*, Educational Testing Service, URL <http://www.ets.org/research/pic>
- Brisbane Catholic Education, 2001, *Consistency of Teacher Judgment: Information Kit*.
- Center for Education Reform Newswire, 2002, Vol 4, No 5, February 5, URL <http://www.edreform.com>
- Center for language and Learning, 1999, *Connecting Classroom and Large-scale Assessment*, cited 15 November 2001, URL <http://www.learningrecord.org/LRs/moderations/reports/99/ModReport99.html>
- Department of Education and Community Services, South Australia, 1996, *Using Student Performance Standards in English August '96 draft, supported by Parent Descriptors, English*, Author.
- Department of Education, Queensland, 1994 *Student Performance Standards in Mathematics for Queensland Schools*, Author.
- , 1998, *The Year 2 Diagnostic Net Handbook*, Author.
- Education Queensland, 2001, *New Basics Project — New Basics — The Why, What, How and When of Rich Tasks*, Author.
- Fairtest, cited 9 November 2001, The National Center for Fair & Open Testing, URL <http://www.fairtest.org>
- Freedman, S., 1993 Linking Large-Scale Testing and Classroom Portfolio Assessments of Student Writing, in *Educational Assessment: A critique of Current Policy* Vol 1, Issue 1, winter.
- Griffin, P. & Smith, P., 1996 *The Implications of Outcome-Based Education for Teachers' Work*, Assessment Research Centre, University of Melbourne.

- Griffin, P. 1997, Assessment Principles for OBE, in Griffin, P. & Smith, P. *Outcome-Based Education: Issues and Strategies for Schools*, Assessment and Research Centre, University of Melbourne.
- House Education and the Workforce Committee, cited 10 November 2001, URL <http://edworkforce.house.gov/issues/107th/education/nclb/factbushtest.pdf>
- Indiana Department of Education, cited 8 November 2001, URL <http://www.doe.state.in.us>
- cited 11 March 2002, Applied Skills Assessment Book 1 — English/Language Arts and Mathematics Grade 3, URL http://www.doe.state.in.us/istep/pdf/40494_WEB_03_ApSk_INf00.pdf
- cited 11 March 2002, Scoring Guides — English/Language Arts and Mathematics Grade 3, URL http://doe.state.in.us/publications/pdf_istep/03elam_SG_INf00.pdf
- Kalantzis, M. & Cope, B., 1997, *Multiliteracies: rethinking what we mean by literacy and what we teach as literacy in the context of global cultural diversity and new communications technologies*, Centre for Workplace Communication and Culture.
- Kentucky Department of Education, cited 1 November 2001, URL <http://www.kde.state.ky.us>
- cited 11 March 2002, Sharpen Your Child's Writing Skills: A Guidebook for Kentucky Parents, URL <http://www.kde.state.ky.us/oapd/curric/portfolios/SharpenWritingSkills.asp>
- Kerlinger, F., 1986, *Foundations of Behavioural Research*, CBS College Publishing, New York.
- Luke, A. Land, T. van Kraayenoord, C. & Elkins, J. 1997, *Report of an intrinsic critical appraisal of the Year 2 Diagnostic Net continua and associated teacher support materials undertaken for the Queensland School Curriculum Council*, Queensland School Curriculum Council.
- Maryland State Department of Education, 1996a, cited 14 March 2002, URL http://mdk12.org/share/publicrelease/deserts_task.pdf
- 1996b, cited 14 March 2002, URL http://mdk12.org/share/publicrelease/deserts_sg.pdf
- cited 31 October 2001, URL <http://mdk12.org/mspp/mspap>
- cited 31 October 2001 — What is MSPAP? URL <http://mdk12.org/mspp/mspap/what-is-mspap/index.html>
- Maxwell, G., 2001, *Teacher Observation in Student Assessment: a discussion paper*, Queensland School Curriculum Council.
- New South Wales Department of Education and Training, 2001a, Year 3 Primary Writing Assessment 2001: Writing Task Marking procedures, Author.

- New South Wales Department of Education and Training, 2001b, Year 5 Primary Writing Assessment 2001: Writing Task Marking procedures, Author.
- New South Wales Department of Education and Training, 2001c, Secondary Numeracy Assessment Program Year 7 and Year 8 Extended Response Marking procedures, Author.
- New South Wales Department of Education and Training, 2001 School Assessment and Reporting Unit, 2001 English Language and Literacy Assessment Years 7 and 8: Writing Task Marking procedures, Author.
- Northern Territory Government, 2001, Multilevel Assessment Program Teacher Guidelines for Administration, Author.
- Queensland Board of Senior Secondary School Studies, 1999 *The Moderation Handbook*, Author.
- Queensland School Curriculum Council, 1999a, *Review of Queensland Literacy and Numeracy Testing Programs, 1995–1999*, Issues Paper, Author.
- , 1999b, *Health and Physical Education: Years 1 to 10 Syllabus*, Author.
- , 1999c *Health and Physical Education: Years 1 to 10 Sourcebook Guidelines*, Author.
- , 2000a, *Review of Queensland Literacy and Numeracy Testing Programs 1995–1999*, Author.
- , 2000b, *Queensland Years 3, 5 and 7 Testing Program: Report to the Minister for Education*, Author.
- , 2000c, *Literacy: Position Paper*, URL
http://www.qscc.qld.edu.au/p-10_framework/literacy.pdf
- 2000d, *Numeracy: Position Paper*, URL
http://www.qscc.qld.edu.au/p-10_framework/literacy.pdf
- , 2000e, *Outcomes-based Approaches to Assessment and Reporting: Project Report (Unedited Draft)*, Author.
- , 2000f, *Consistency in Teacher Judgment*. Research Report, Author.
- , 2001a, *Annotated Work Samples: Research Report*. URL
<http://www.qscc.qld.edu.au/research/index.html#annotatedwork>
- , 2001b, *Position and Guidelines on Assessment and Reporting, Years 1 to 10*. URL
http://www.qscc.qld.edu.au/research/pdf/enposition_assessrtp.pdf
- , 2002a, *English Years 1 to 10 Draft Syllabus*, Extended Trial 2002, (Implementation 2004)
- , 2002b, *Mathematics Years 1 to 10 Draft Syllabus*, Extended Trial 2002, (Implementation 2004)

- Rowe, K., 1997, Factors affecting students progress in reading: key findings from a longitudinal study, in Swartz, S. & Klein, A., (eds) *Research in Reading Recovery*.
- Sadler, R., 2001, *Conversations about the Learning Record*, *Learning Record Online* cited 23 October 2001 URL
<http://www.cwrl.utexas.edu/~syverson/olr/sadler.html>
- Sanders, W. & Horn, S. (1995) *Educational Assessment Reassessed: the usefulness of standardized and alternative measures of student achievement as indicators for the assessment of educational outcomes*, cited 28 September 2001 URL <http://olam.ed.asu.edu/epaa/v3n6.html>
- Stewart-Dore, N. & Bartlett, L., 1999, *External Review of the Year 2 Diagnostic Net*, Queensland School Curriculum Council.
- The New London Group, 1996, *A Pedagogy of Multiliteracies*, *Harvard Educational Review*, Vol 66, No 1, Spring, p. 83
- Unsworth, L., 2001, *Teaching Multiliteracies across the curriculum: changing contexts of text and image in classroom practice*, Open University Press.
- Vermont Department of Education, cited 7 November 2001 URL
<http://crs.uvm.edu/schlrpt/perform.htm>
- Victorian Curriculum and Assessment Authority, 2000, *Achievement Improvement Monitor 2001* Guide for Principals and teachers: English and Mathematics testing Component Year 3 and Year 5, Author.
- Wiggins, G., 1990 *The Case for Authentic Assessment*, in *Practical Assessment, Research and Evaluation*, cited 29 October, 2001, URL
<http://ericae.net/pare/getvn.asp?v=2&n=2>
- Wiggins, G., 1993, *Assessing Student Performance: exploring the purpose and limits of testing*, Josey-Bass, San Francisco.
- Wyatt-Smith, C., 1995, Teachers' Reading Practices: The interplay of pre-specified assessment criteria and other factors, *Literacy Learning: Secondary Thoughts*, Vol 4, No 2.
- & Ludwig, C., 1998, Teachers' Roles in Large-Scale Literacy Assessment, *Curriculum Perspectives*, Vol 18, No 3.