

# Evaluation Report of the Pilot of the 2005 Queensland Assessment Task (QAT)

Val Klenowski



2006

**This work was funded by the Department of Education and the Arts.**

**Queensland government  
Department of Education and the Arts (2006)**



# 1 INTRODUCTION

## **Background**

In October 2001 a taskforce of experts and stakeholders was established 'to develop a conceptual framework for assessment and reporting in the Queensland context for Years 1-10; to consider the tools necessary to translate the framework into practice; and to advise on appropriate professional development for teachers and administrators relating to assessment and reporting' (Education Queensland, 2002:1).

## **The Assessment and Reporting Taskforce**

In February of the following year, the Minister for Education accepted and endorsed the strategies that were recommended in the *Report of the Assessment and Reporting Taskforce*. The result was a five year plan for developing teachers' assessment literacy which would involve giving attention to the five key elements of assessment. Namely: the assessment task (derived from the curriculum), student performance (not always written), a judgment of the performance with reference to the standard, feedback to the learning and the teacher/curriculum and moderation. It was emphasized that some teachers and schools might give more attention to the design of assessment tasks while, for others, the professional development required might be a greater emphasis on moderation (ibid: 2).

Funds were committed by the Minister of Education for professional development purposes that involved the development and implementation of assessment workshops addressing areas such as: assessment for and of learning, planning for assessment, designing quality assessment tasks, judging the quality of student work against standards and combining results from different assessments (Education Queensland, 2003).

## **Assessment and Reporting Framework**

Education Queensland's (EQ) Assessment and Reporting Framework Implementation Committee (ARFIC) was established in mid 2002. In addition, an Assessment and Reporting Unit was set up within the Assessment and New Basics (ANB) Branch. It was at this time that EQ advised that there was to be a common frame of reference against which teachers would assess and report student achievement. That is, not only what the student can do following a course of teaching and learning but also how well they can do it and under what conditions. Student achievement was to be reported in terms of certain constructs, at certain junctures in a common format to parents and the system.

The intent was to establish an Assessment and Reporting Framework (ARF) that would comprise the following elements:

- a common frame of reference that includes a set of achievement standards to provide a common language for communication among teachers, parents students and the system about student achievement (Queensland Standards Map);
- a range of assessment methods for collecting evidence of student achievement;
- procedures for collecting statewide data on student achievement;
- processes for reporting achievement and monitoring progress over time and
- professional learning opportunities for teachers (with respect to the elements of the framework and assessment practice in general) (ibid: 5).

### **ARFIC 2003**

The following principles for developing the constructs of the EQ standards map were identified in February 2003 by ARFIC. First, it was stated that the map should be:

- such that it can be readily implemented by teachers;
- related to current work being done by teachers in schools and
- usable in conjunction with teacher-generated and controlled collection of evidence that reflects what is being taught.

Second, there should be a sufficient number of constructs to provide useful system-level data, but kept minimal so that there are no more than five constructs and no more than six levels. The constructs should:

- cover a range of students' knowledges and skills (e.g. cognitive processes, practical skills and dispositions);
- be global rather than fine grained;
- be futuristic (e.g. multi-literacies, ICTs, preferred social future)
- be readily recognized and understood;
- be faithful to, and evolving from, the curriculum (QSA syllabus) eg consider overall learning outcomes;
- cover: domain-specific knowledges; literacy; numeracy; personal, social and ethical development;
- be inclusive of the range of all students' actual performances;
- cover selective content;
- be applicable to all schools, regardless of their curriculum organizers and
- be such that they can be supported by evidence.

Third, the standards should:

- focus positively on what students know and can do;

- be nested;
- be developed from actual student work and standards descriptors should reflect what is seen by the guild of professionals in the student work;
- be inclusive of a range of all students' actual performances;
- be able to show improvement and
- be such that they can be applied to portfolios containing a range of evidence types (pencil and paper, performance/demonstrations, applying ICT).

The standards should be described so that they are expressed generically in that they transcend any curriculum organizer, are qualitative and use language that is meaningful to parents. There should be a range of standards at each juncture and the number of standards for each construct should be:

- sufficient to provide the system with useful data but be kept to the same minimal number as the constructs and
- such that the difference between the standards is obvious.

Fourth the number of junctures at which data are collected should be such that they provide parents and the system with data at student's significant stages of development and should be able to show improvement. It was recommended that the preparatory year be considered a juncture.

Finally, the development of the map should grow out of practice and not be derived from a priori analysis. The map's development should also consider whether constructs are common across junctures and should be 'road-tested' by practitioners. Learners too should contribute to the development of the map through learner-controlled conferences inspired by the question: 'what do you value about your performance?' Another principle included evaluation and a redevelopment period with the map released as a work in progress rather than developed until perfect.

Reporting was recommended to be on achievement rather than development and there should be no reported aggregation of the reported levels/reporting.

In terms of implementation it was recommended that the use of the map should consider:

- mechanisms for accountability;
- accountability in terms of reporting the match between curriculum and assessment to government;
- processes for establishing shared understandings of meanings (moderation, sharing student work samples, sharing judgments about how the standards match the work) and
- should not be linked to funding for schools.

## **Queensland Standards Map**

The Queensland Standards Map (QSM) was conceptualized as ‘a mechanism for teachers to capture the results of their assessments and to report student achievement in terms of the three constructs of knowledges, processing and self and others’ (ibid: 5). It was suggested that evidence of student achievement would be collected from nominated curriculum areas (domains) which would not always be Key Learning Area (KLA)-specific and which would include cross-curricular priorities such as literacy, numeracy, lifeskills and futures perspective. A portfolio of specified student work would provide evidence of student achievement. Achievements of all students in years 3, 6 and 9 (the juncture years) would be reported annually, at the end of the academic year to parents. System level reporting of a sample of these students was also to be reported annually at the end of the academic year (ibid: 6). It was planned that the students’ individual results in a particular domain would be expressed in one of four standards.

Reporting of student achievement was to be based on three constructs. These were, firstly, knowledges that refer to three particular domain-specific knowledge dimensions. These are factual, procedural and conceptual. They include the content, skills and understandings that make one domain different from another. The second construct of processing refers to the cognitive and linguistic processes associated with reflecting, communicating and enacting across domains. These have been described as transdisciplinary skills. Examples of these skills include analyzing and deducing, creating and presenting, expressing and performing. They are used in working with ideas, information, artefacts and texts. The third construct, of self and others, refers to ethics, culture and relationships. This includes: the interpersonal, personal and ethical dimensions of human life, and social and cultural practices in rapidly changing global and local contexts.

The evidence of student achievement in a domain was to be collected through two distinct but related components:

- student responses to standardized tasks in the particular domains and
- student responses to teacher-generated tasks in system-specified categories of the same domain.

It was emphasized that they would be complementary to provide a more comprehensive assessment of the student’s achievement.

## **Queensland Assessment Task**

The Queensland Assessment Task (QAT) was planned to be a collection of standardized tasks. Standardized was interpreted as the same for all the students in the cohort, undertaken in schools according to the same list of task parameters and marked according to a commonly applied marking scheme. The

tasks were to be derived from the intents of existing QSA syllabuses in the eight KLAs. They were to be developed by experts to the school acting under the direction of the EQ's ANB Branch who were responsible for the design specifications.

A Teacher-Generated Task (TGT) was also intended to provide evidence of student achievement within a particular classroom context or school environment. The TGTs were to be designed by teachers within categories specified by EQ's ANB Branch, such as 'the evaluation phase of design in Technology' or 'evidence of appropriate laboratory techniques in Science'. The TGT would complement the QAT in the domain nominated for the corresponding reporting juncture. The TGT would also take into account the richness of the students' classroom experiences, the complexity of students' worlds, the diversity of Queensland's student population, and the imperative for comparability of standards of reported results. Task parameters that were suggested included working individually or in teams, doing a task in paper or electronic format, completing a task over a prescribed period, oral, interaction and demonstration (ibid: 9).

A judgment about the quality of student achievement would be based on a portfolio of evidence, defined as 'a deliberate, strategic and specific collection of student work.' A portfolio of work would be submitted for assessment per nominated domain (that would sometimes be KLA specific but would at times also include cross-curricular priorities). It was recommended that in any year there would usually be two domains, or even three, but no more (ibid). This was planned as a further element of the Assessment and Reporting Framework.

#### *Judging the quality of student work*

All tasks (QATs and TGTs) were to be assessed by teachers with reference to task-specific marking guides. Marking guides would be developed to reflect the intents of the individual syllabuses and to allow for the reporting of achievement in terms of the broader constructs of the QSM (i.e. knowledges, processing, self and others). Teachers were to use the marking guides that contained verbal descriptors of available grades (grades descriptors). To assess student work teachers would determine the grade to be assigned by finding the 'best fit' by matching the qualities of the work with the grade descriptor.

For QATs the grades in the marking guides were underpinned by task-specific grade descriptors, by exemplars of student work illustrating each grade and face-to-face or virtual dialogue associated with task intent and expectations, and the quality of the student work produced. The ANB Branch was to provide marking guides for QATs. Teachers were to be trained to mark student responses to the QATs and in how to make consistent judgments about the quality of student work. While in the marking of TGTs teachers were supported in developing and implementing their own marking guides. Letter grades were recommended and



the number of available grades for marking would be task-specific and range from two to five.

For each domain assessed, a student's results from the QAT and the TGTs would be combined to produce a single result for each of the three constructs in the QSM. (It was indicated that it would not always be possible to provide significant evidence of student achievement in all three constructs.) It was emphasized that when results are combined the single result would give equal weighting to the QAT and the school's TGT.

#### *Procedures for generating data on student achievement*

Data would be collected by the system, annually in September, in Years 3, 6 and 9. It was recommended that data would be collected from a sample of students only and about student achievement in no more than three domains per academic year. However, *all* students would be required to complete the QATs and TGTs for the nominated domains and for a given juncture. The system would be able to generate summary data on the achievement of students statewide in the nominated domains, and report on these data as appropriate, e.g. by gender, by socioeconomic status.

All teachers would have access to professional learning opportunities related to composing and marking QATs and TGTs despite the fact that only some schools would have sample students.

Portfolios would be required for all students in the juncture years in the nominated domains. These domains would be identified well in advance, while notice of the sample of students/schools would be shorter. It was recommended that achievement data from the sample of students would be collected, collated and reported at system level. The moderation process would only apply to the work of sample students.

The two formats for reporting student achievement at the end of the year that were recommended were a school's report to parents/carers; and a system's report. Student achievement at each juncture would be reported against standards on single scales. The three constructs are the basis for developing the scales.

#### *Reporting: Student Level*

It was intended that parents/carers would receive a report that contained information in a standardized format. This would incorporate the domain nominated for assessment in a given year with results reported under the three construct headings of the standards descriptors (one of four). The portfolio of work was to be assessed by expert judges matching the portfolio evidence by evaluating the best fit to one of the standards descriptors.

Strategies suggested for reporting student results included a profile of results (one per construct/ subconstruct) or combination of results across two or more constructs to provide results of a much larger grain size (this requires a second round of decision-making based on an agreed policy combining results) (ibid: 12).

The school report, based on the QSM could contain: academic year, juncture year, curriculum areas, constructs for assessing and reporting (i.e. knowledges, processing, self and others), standard (for each of the three constructs in the two domains), explanation of the constructs and standards awarded, information about the evidence upon which the assessments are based.

### *Reporting: System Level*

Suggestions for how the data for a given student cohort of data might be reported included:

- summary data for each domain in each construct for which evidence is available;
- combined data for each domain across the three constructs to produce an individual student result and then results from the sample students aggregated for reporting systemically or
- results for a cohort could be reported by aggregating from sample students in one of the constructs such as processing so that at system level reporting on students' higher-order thinking skills would occur (i.e. multiliteracies, problem-solving, ICTs).

With the full implementation of ARF longitudinal data would be available to track progress over time at a given juncture and between junctures.

### *Teacher Development*

It was anticipated that the QSM would be implemented in 2006, after a pilot of the QAT in 2003, the development of standards and the voluntary participation by schools in further QATs in nominated domains in 2004, 2005 and 2006. During this period of development professional learning opportunities would be available to teachers.

A selection of standardized tasks would be provided for teacher use in EQ schools. The materials were to include: tasks; administration guides; marking guides and annotated samples of student work.

Models of good assessment practice were to be provided with the opportunity for teachers to compare their students' work with the achievements of students across the State.

### *Scheduled Development*

A schedule of events was planned for the period 2003 – 2006. It was anticipated that all KLAs would be released by 2006, which marked the final year of the 5-year plan for implementation of the recommendations of *The Report of the Assessment and Reporting Taskforce*. The following shifts were envisaged:

- from QATs developed by experts external to EQ schools to QATs co-developed by EQ's ANB Branch and EQ teachers;
- from an available assessment and reporting framework to one that is robust and embedded in the system;
- from a curriculum that consists of KLA syllabuses in various stages of implementation to a curriculum that consists of KLA syllabuses in the final stages of full implementation;
- from New Basics as a concept to the New Basics Framework as a mature approach to curriculum, pedagogy and assessment.

### **The 2003 QAT Pilot (June – December 2003)**

The pilot study was designed to be a 'rehearsal' of the critical stages in the Assessment and Reporting Framework cycle that relate to tasks, students and teachers (ANBB, 2004b: 1). The design specifications for the 2003 QAT will be described and compared with those of the 2005 QAT pilot study. The outcomes will be analysed to understand the implementation of the intended assessment and reporting framework.

### *2003 QAT Design Specifications*

The 2003 design specifications indicated that:

- each QAT would comprise a set of standardized tasks;
- responses to the QAT, with responses to the complementary TGT, would constitute a portfolio of evidence;
- the portfolio was one element in EQ's ARF;
- target achievement of the QAT and the TGT would be in a nominated domain or curriculum area or would be transdisciplinary;
- the content of the QAT and the TGT would take into account cross-curricular priorities such as the multiliteracies and futures perspectives;
- a QAT would be administered annually, in September, to students in at least one of the juncture years (Years 3, 6 and 9);
- a QAT would provide systemic data on the achievement of a statewide sample of students from the cohort of the chosen juncture year and
- experts external to schools were to develop QATs.

The 2003 design specifications indicated that the relationship of the QAT to curriculum and syllabuses would be such that:

- each QAT reflected the broad intents of the KLA syllabuses for Years 1-10 while aligning with the New Basics curriculum organizers for Years 1-9 and
- the natures of the standardized tasks that comprise a QAT were inferred from matching the juncture years to KLA levels as follows: Year 3: Level 2; Year 6: Levels 3 and 4; Year 9: Levels 5 and 6 (except for Languages other than English, where the level is dependent on years of learning).

The structure of the 2003 QAT was specified as follows:

- A QAT would include standardized tasks that could be disparate and distinct but the QAT had to cohere;
- Each standardized task would assess one or more significant aspects of the nominated domain;
- An individual task could contain several small subtasks;
- A QAT for a specific juncture would ideally include a task that is the same as one included for another juncture and
- A 3-section model for a QAT for a domain of Science, say, could be: initial written task concentrating on specific knowledges; laboratory task; task involving synthesis or extension of knowledge and results from the first two.

In relation to the standards map the following was specified:

- Students' individual results in a nominated domain, each expressed as one of four standards (Standards 1-4) were to be determined using a process that involves the use of an overarching standards schema, the Queensland Standards Map (QSM). Standard 1 was to be what students aspire to (the so-called "aspirational" standard) as opposed to what might prove typically to be the work of highest quality;
- Ideally, a QAT was to provide evidence about student achievement within and across the three primary constructs of the QSM for capturing assessment results and reporting student achievement (i.e. knowledges, processing and self & others);
- Ideally, the evidence gathered from a QAT in a construct was to provide information about student achievement in more than one of the three subconstructs (e.g. a QAT designed to assess knowledges should assess in more than one of factual knowledge, conceptual knowledge and procedural knowledge);
- A TGT was to complement a QAT in that it was to assess in constructs and subconstructs not covered by the QAT, or that additionally it could assess constructs also covered by the QAT and
- Student results in QATs and TGTs was also considered to be useful for further reporting to the system on problem-solving, public demonstrations of mastery, ICTs etc.

The technical properties indicated that an individual QAT would:

- Include a balanced range of tasks;

- Be based on material that is generally accessible;
- Contain tasks of varying difficulty so that students have the opportunity to demonstrate the extent of their achievements, and to allow markers to discriminate between achievements of different quality;
- Require responses in a variety of modes (e.g. diagrammatic, verbal, tabular/graphical, pictorial, symbolic) and
- Be cognisant of equity issues, and promote inclusion of students with special needs and students requiring special consideration.

The task parameters were task-specific because of both the variation that is possible in the tasks comprising a QAT and the likely differences among QATs for different domains. Each individual standardised task of a QAT, however, was to be administered under commonly applied conditions. It was specified that except when speed is a critical element, tasks should be unsped (i.e. in the case where time is specified, it should be sufficient to allow most students the opportunity to do their best work). Tasks should also be unrehearsed; which could mean that students were given short notice of task requirements.

The task parameters were specified as follows, they should:

- assist in ensuring that task responses are those of the given student or team;
- make clear whether a task is to be done individually or in teams and, if relevant, any requirements on the size or composition of teams;
- specify any limits on, or freedoms to choose, the medium of response (e.g. pen-and- paper, electronic, oral, demonstration);
- state clearly and precisely what it is that markers will look at and
- specify if students are to need access to resources and equipment (e.g. video camera) and, if so, make clear how.

It was also emphasized that unless otherwise appropriate, a standardised task (i.e. one of the tasks that makes up a QAT) should not have the flavour of a “point-in-time test”.

In specifying the approach to marking the following was stipulated:

- Teachers were to be the markers of student responses to tasks;
- Markers were to assign a grade to a task after referring to a task-specific marking guide containing verbal descriptors of available grades for marking (grades descriptors). The grades descriptors in the marking guides were to be an exemplification of the general standards in the QSM;
- For QATs, the grades were to be described in marking guides devised centrally. For TGTs, the grades were to be described in marking guides devised by the teachers who designed the TGTs to suit particular classroom contexts and to fall within centrally specified categories;
- The grades in the marking guides for QATs were to be underpinned by: task-specific grades descriptors; samples of student work illustrating what is expected for each available grade in the marking guide; face-to-face or

virtual dialogues associated with expectations and with the quality of actual student work;

- Teachers were to decide which of the grades descriptors in the set best matched the explicit features in the student work and were to assign the corresponding grade and
- There was to be a process to monitor the consistency of teacher judgments in the marking of QATs.

In terms of grades, marking guides and annotated samples of student work the following specifications were made:

- Results of assessments of tasks were to be expressed as letter-grades (an interim arrangement);
- Marking guides were to be task-specific;
- The number of grades available for marking were to be task-specific;
- Each response to a task was to be assigned one or more grades, as specified in the marking guides. Individual subtasks did not have to give rise to individual grades. More than one grade could be awarded when a task was clearly a source of evidence of achievement in more than one construct or subconstruct;
- The number of available grades was to range from two to five;
- The grade for work of highest quality was coded as A. The A-grade was to capture responses that have the qualities of the task-developer's vision of an ideal response;
- When students perform in teams, the formula for assigning grades to individual students on a task was to be given in the marking guide;
- Each grade was to be associated with one or more sets of verbal descriptors;
- An indicative marking guide (referring to number of available grades and providing associated grades descriptors) was to be developed contemporaneously with the task;
- The task developer had to provide an indicative A-grade (or ideal) response as an accompaniment to every task and
- Other samples of student responses could be used to illustrate the meanings of grades descriptors in the marking guides and to show how these might be manifested in student work.

A timetable was developed for the four developmental years, 2003—2007, with an indication of the likely domains that would be assessed through QATs and TGTs. Some of the juncture years that were to be involved were also specified. For instance, in 2003 Health and Physical Education were chosen as the KLA strand with the field of knowledge identified as Physical Activity/ Physical Education, the other nominated domain was Science and the field of knowledge identified was Natural and Processed Materials/ Chemistry. The latter was to be assessed in Years 6 and 9 and the former in Years 3 and 9.

The developmental cycle for 2003 involved Australian Council for Educational Research (ACER) and EQ's Assessment & New Basics Branch (ANBB). It was specified that the:

- The validity of all standardised tasks was to be checked at each stage of their development through examination by a panel;
- Each standardised task was to be trialled on a population at the appropriate year of schooling in a school system or population closely analogous to the Queensland Years 1—10 population;
- ACER was to suggest processes and associated methods of analysis, including panelling, trialling, revising, presenting etc. to camera-ready copy and
- ANBB was to suggest timelines for processes of checking validity.

The post-administration analysis would involve decisions that were to be made about the methods of analysis of results from QATs and TGTs. This would involve decisions about:

- individual tasks from QATs, and possibly from selected TGTs;
- whole QATs, and possibly whole TGTs;
- entire portfolios and
- individual or sets of constructs or subconstructs.

It was further specified that:

- The analysis could be by entire sample, entire targeted subgroup of sample, or by school/district (depending on the nature of the sample) and
- Selected results from this analysis, together with commentaries on students' achievements, were to be the subject of a post-administration report.

The following dissemination strategy was made specific:

- The State of Queensland (Department of Education now Education Queensland) would hold the copyright for the QAT. New QATs were to be prepared each year comprising standardised tasks that relate to the nominated domains for that year;
- A prototype QAT was to be made available to all state schools;
- In association with the application of the QSM to various domains, an administration guide was to be circulated to all schools and
- Once a QAT was administered it would no longer be secure and schools might retain it for their own purposes. A retrospective on that QAT for that domain for that year, including model responses where possible, were to be published at the beginning of the subsequent year.

An addendum to the QAT design specifications explained the alteration of assessment domain(s) for 2004/5. The schedule of QAT development to 2007 had nominated Science and Health and Physical Education (HPE) as the syllabuses for assessment in 2003, Languages Other Than English (LOTE) and Studies Of Society and the Environment (SOSE) for 2004, and The Arts and Technology for 2005. It was explained that while a strand from the Science

syllabus (Chemistry/Natural and Processed Materials) and a strand from the HPE syllabus (Physical Activity) were the basis for QATs (and TGTs) in 2003, it was no longer to be the case that LOTE and SOSE would be used in 2004 or that The Arts and Technology would be used in 2005. Rather, a QAT (and TGT) was to be designed for 2004/5 to assess cross-KLA skills; namely, transforming information/ideas from one form to another.

The factors responsible for the change in the choice of the domain and year level were identified as follows:

- ARF pilot study revealed limited uptake of LOTE in schools;
- Final year for implementation of The Arts was 2005 and the ministerial portfolio had now included the Arts;
- QSA consultation on a Year 9 test of literacy and numeracy concluded that it was not desired;
- Ideally, a QAT would provide results under three headings (knowledges, processing and self & others);
- Assessment data from the 2003 QATs would fall mainly into the knowledges category (i.e. facts, concepts and procedures) and
- The processing category would cover generic skills, which include literacy and numeracy.

Finally the change of name from the technical specifications that describe content, construct etc. for the QAT would no longer be referred to as “Design specifications”. They would take on the traditional title of “Task [as in test] specifications”. The term “design” would be reserved for “design brief”.

#### *The 2003 QAT Pilot Study*

The 2003 pilot study involved 20 volunteer schools. The two domains of learning chosen for the common assessment tasks paralleled the implementation of the KLA syllabuses so Science and Physical Education were chosen and in June 2003, teachers were informed that the nominated domains and junctures for assessment would be Chemistry (Years 6 and 9) and Physical Activity (Years 3 and 9).

The Australian Council for Educational Research (ACER) was contracted to provide camera-ready copy of tasks, administration and marking guides. The development of tasks, based on the task specifications, developed within the Assessment and New Basics Branch (ANBB), involved ACER and ANBB staff.

The time for the development of the QAT was reduced to the period between the release of the proposed Assessment Reporting Framework (May 2003) and the preferred administration date in schools of September 2003. The available time was reduced for the processes of panelling QATs and trialling the QAT on another population (Victorian schools) with the implementation of the implications for action and revision.



This model of co-development resulted in differences of opinion concerning what constituted good assessment practice compared with good testing practice. In addition, the KLA syllabuses were reported as insufficiently prescriptive regarding expectations of student performance (Assessment and New Basics Branch, 2004a: 1).

Teachers were familiarised with a prototype QAT, its marking guides and student work representing the highest standard at a two-day conference.

QATs in Chemistry (Years 6 and 9) and in Physical Activity (Years 3 and 9) were administered to students in pilot schools during September. The tasks included written tasks (for both domains), laboratory tasks (Chemistry) and physical performances (Physical Activity). Schools ensured that students completed the QATs under commonly applied conditions.

A week before the tasks were to be completed by the students, materials for the performance aspect of the Physical Activity QAT were sent to schools so that teachers could familiarise themselves with marking guides, training tapes to support the marking guides for game play, and rules of that game. The remaining QAT materials (tasks, response booklets, administration guides, consent forms, student attendance rolls, individual student ID slips etc.) were dispatched to schools a week later in September.

Aspects of the tasks required students to do laboratory work or play a game. Schools were challenged by the organization of these tasks. To illustrate, for Chemistry schools had to purchase iodine, flour, eggs, milk, and utensils such as frying pans (expenses met by ANBB) while for the Physical Activity teachers had to rehearse the game with students and mark this aspect of the QAT on site (after dedicating time to understanding the expected standards by viewing and discussing the training tape).

Student work was returned to the ANBB and assembled in preparation for marking. Sixty-five markers were involved in the marking for five days. In the pre-marking stage, sample student responses were examined and marking guides were further refined. More than 5,000 tasks were marked, over 50 per cent of them twice. The average inter-marker agreement was 0.80.

Teacher-assessors were positive about the experience and commented on how much they had learnt about setting meaningful tasks, developing assessment criteria, and making judgments about the quality of student work.

Comments on student work from marking advisers and teacher-assessors were analysed to reveal:

- a high rate of omitted responses (especially from Year 9 students);
- the poor quality of students' writing, spelling and literacy skills in general;

- students' inability to actually read and/or interpret the demands of the task;
- students' lack of recognition of the Chemistry in the Chemistry task and
- an apparent lack of exposure in classrooms to certain attributes/skills such as analysis, cause and effect, explanation of observations, recording observations, interpreting data and making inferences. (ANBB, 2004a: 2)

Teachers in pilot schools were informed of the design brief for the development of TGTs at a teacher conference in July 2003. The brief required not only the generation of an assessment task but also the determination of task-specific criteria and standards for marking. Teachers worked with colleagues within and across school sites to explore ideas for developing an assessment task that fulfilled the design brief. Four weeks were allowed for this.

In August, teachers recounted their experiences in fulfilling the design brief and received feedback from their peers and ANBB staff on the tasks they had generated. They refined and finalised their TGTs the following week. These teachers then presented their TGTs to ANBB for consideration late August. At this stage, 7 out of 40 tasks were accredited. ANBB staff then worked with teachers/schools in the revision of the tasks. All TGTs were accredited by mid September.

After students had completed the TGTs (which differed from school to school), teachers marked student work using the task-specific marking guides they had developed in October 2003 and forwarded student results to ANBB by mid November. Teacher feedback indicated that the TGT development process was difficult and intensive. Most teachers reported working on their tasks, individually or with colleagues, for a period of at least three of the four weeks available. The TGT development process revealed that, in general, teachers are unable to design good assessment tasks. It was concluded by the pilot study that, in general, teachers possess inadequate knowledge of particular domains of learning. Teacher feedback also indicated that the creation of marking guides was difficult. Several had described it as the most difficult professional task they had faced whereas others stated that they had never before attempted it (ANBB, 2004a: 3).

Information on student performance was communicated to schools in the form of a table of results (grades) for all students on all tasks, with explanatory notes attached.

To conclude the key aspects of the Assessment and Reporting Framework that were included in the pilot study were:

- development of QATs in two domains of learning at three reporting junctures;
- administration of QATs to students under commonly applied conditions;
- marking of student responses to QATs by teachers who trained in the use of centrally-generated task-specific marking guides;

- development and accreditation of TGTs in two domains of learning at three reporting junctures, ultimately for use in conjunction with the QATs and
- marking of student responses of TGTs by teachers using teacher-generated task-specific marking guides (complementing the constructs of the QSM).

Those aspects of the ARF that were not included in the pilot study were:

- procedures for ensuring comparability of standards;
- combining results from different assessments for reporting against the QSM and
- methods for summarizing and displaying data for reporting to the system.

### *Key Findings from the Pilot Study*

There were five key outcomes of the pilot study related to:

- level of intellectual and operational capacity required;
- resourcing the development of authentic, contextualised assessment tasks;
- teacher design of good assessment tasks;
- lack of curriculum coherence and
- cost of “success”.

The task development, administration and marking placed exceptional demands on ANBB staff and pilot school teachers. An analysis of the operational issues, technical matters and logistical problems that emerged during the course of the pilot study was undertaken to inform the next stage of the project.

The development of the QATs and TGTs proved resource-intensive (financially, intellectually and timewise). According to teachers, the resultant tasks were valuable and worthwhile pieces of assessment. It was suggested at this stage that if evidence of student achievement in the form of a portfolio of student work containing QATs and TGTs were to remain a key element of the proposed ARF, the resource implications of this for the Department would require further discussion (ANBB, 2004b: 2).

It was also stated that teachers need intensive support in designing good assessment tasks and related marking-guidelines. This support can be provided through assessment workshops and annotated exemplars that have been part of the future plans for statewide dissemination of materials associated with the ARF. The pilot study revealed that teachers need even more support.

The conclusion was reached that systemic data collection assumes curriculum coherence but as illustrated in the pilot, curriculum in the compulsory years is fragmented. Feedback from schools also indicated significant differences in the way syllabuses are implemented. Many of the concepts within syllabus strands

did not appear in the enacted curriculum. The use of KLA syllabus strands as the basis for describing the domains for assessment (as in the proposed ARF) was based on the assumption that all students have the opportunity to access the intended curriculum (as stated in EQ policy) (ANBB, 2004a: 3). It was concluded that if this was not the case then the use of KLA syllabus content areas for the domain-specific knowledge category in the QSM as central to the systemic data collection had to be questioned (ANBB, 2004b: 3).

The students' responses to the tasks indicated their general lack of engagement with the concepts assessed while the marking revealed low levels of achievement (ibid). Possible explanations were offered, these implied that teachers, in good faith, made choices about what they believed students should and should not be exposed to in different curriculum areas and in so doing it was concluded that teachers were "dumbing down" the curriculum (ibid). This conclusion appeared to be based on possible explanations rather than evidence.

#### *Implications for the 2004 QAT*

At this point in the project it was realized that the ANBB was unable to deliver QATs for 2004 in another two domains at three key junctures. Experienced staff who could work with ACER or in-house on QAT development were committed to delivering assessment workshops, the four stage moderation strategy and developing other aspects of the ARF. Even though funding was available, the recruitment of staff with the required expertise was complicated by the fact that only short-term work could be offered because of the establishment staffing numbers (ibid). A compromise position was offered in that the project would proceed with the in-house development of a futuristic QAT at one juncture only (Year 9) to assess generic skills in literacy and numeracy with an emphasis on the transformation of ideas and information (in the context of SOSE and the Arts) and going beyond the traditional pen-and-paper testing of such skills (ibid). Work was to continue in 2004 with teachers on their involvement in the TGT process.

It was also reported (ibid) that ARFIC had critiqued the QATs that were administered in September 2003 and that not all ARFIC members supported the development of a standards map. It was also made apparent that some senior officers from QSA held views that were in conflict with EQ's stated position on standards-based assessment (ibid).

#### **(ARFIC) Working Parties**

In March 2004 it was agreed at an ARFIC meeting to establish five working parties to advance the refinement of the proposed Assessment and Reporting Framework. The working parties were planned as follows.

### *Working Party on Standards*

This working party was required to:

- consult on and refine the standards descriptors for Years 3, 6 and 9 in two domains of the QSM that were composed in 2003;
- write standards descriptors for Years 4 and 9 in other domains of the QSM. (domains had not yet been confirmed but it was assumed that they would most likely be transdisciplinary);
- progress the standards descriptors through consultation loops (based on the process designed in 2003) and
- test out descriptors composed in 2003 and 2004 against student work.

To inform its work this working party:

- was guided by the set of principles for writing standards descriptors;
- paralleled the process used in 2003 and
- modelled the genre and layout of the draft versions at Standards 1 and 3 for Chemistry at Years 6 and 9 and Physical Activity at Years 3 and 9.

The scope of the work completed was narrow given the set timeframe. This working party decided to write only part of a set of standards, to place strong emphasis upon the process for writing standards descriptors and to establish principles that should inform that process.

The standards descriptors that were written were in the form of an overarching template that could be applied according to specific purpose and audience. Additional information was assembled into a document that could constitute parts of the QSM.

Given the limitations of time and resources, the working party did not consult on the 2003 standards, progress the standards descriptors through consultation loops or test out descriptors composed in 2003 and 2004 against student work.

However, the following guiding principles for standards descriptors were developed:

- capacity for schools to relate to the standards;
- grounded in practice and the need for trial;
- student work to be used to substantiate meaning;
- able to be moderated;
- written in positive terms and suitable language for audience and
- to encompass minimum and aspirational performances (ARFIC, 2004a)

### *Working Party on Panelling*

This working party was required to:

- design a mechanism whereby advice on validity is provided during the development of standardised tasks;
- decide how panelling fits into the QAT developmental cycle;
- determine the appropriate composition of panels and
- devise protocols for giving advice and revising tasks.

To inform its work this working party:

- was guided by the literature on panelling as content validation;
- participated in panelling subtasks associated with the 2003 QATs;
- canvassed input from pilot school teachers who were involved in the 2003 marking operation and
- considered how existing processes for panelling other forms of standardized assessment might be applicable to panelling QATs.

This working party adopted the following principles in their approach: the need for panelling to provide quality advice and fulfill the purposes of the Assessment and Reporting Framework of providing the system with data on students' achievements and developing an assessment culture. A further 12 specific principles were adopted to inform the design developed.

The proposed panelling processes were identified for the individual QAT (sub)tasks and for entire QATs. The former involved:

- the development of panelist expertise and an integrated approach;
- individual panels, each expected to deal with a range of validity and technical issues;
- the panelling process to be supported by guidelines that indicate the purposes, the brief, protocols, procedures, expectations;
- the QAT (sub)task and its overview when ready for panelling will involve the flow of the (sub)task through panels and trials to be confirmed by overseers and revised as necessary;
- the flow through panels for a particular QAT to be dependent on its size, nature, level of internal integration;
- each panel to have a trained convener who has some domain knowledge and is trained;
- the taskwriter to be responsible for ensuring that the convener is provided with an overview of the entire QAT, (sub)tasks with statements of assessment intents and (sub)task conditions such as equipment allowed, exemplar responses, marking guide, brief for the panel and curriculum references or justification;
- the convener to select the members of the panel to meet;
- (sub)task panel members to comprise: convener, writer, teachers, working party member, Branch member;
- a system of comprehensive documentation of panellists' suggestions and reasoning;

- prospective conveners and panelists nominated and expertise 'classified' by ARFIC;
- selection criteria determined for prospective panelists, conveners;
- (sub)task panellist and convener training recommended and
- formal accreditation available for those completing panelist and convener training.

The panelling process proposed for entire QATs involved:

- a selection of panelists from the (sub)task panels with some new members;
- this panel to be validating rather than improving the QAT and to focus on issues such as variety, range and balance and
- all these panellists to have been trained and accredited for this second form of panelling.

Other emergent considerations that arose included:

- if QATs are trialled, panellists could as part of the panelling process be shown students' responses on (sub)tasks from trial and the marking outcomes;
- the value of conducting small scale experiments to provide data on students' responses to (sub)tasks;
- in considering equity it was recommended that (sub)task panellists be mindful of the question 'Is this (sub)task as inclusive as it can be without loss of assessment intent?' bearing in mind the nominated equity groups e.g. Special Education, Distance Education, ESL/EFL, ATSIC, rural/remote, boys/ girls and
- a major role of the Branch and the panels that deal with the entire QAT would be to look at equity issues as they relate to the QAT as a whole.

### *Working Party on Intent Matching*

The working party was required to:

- design processes for ensuring alignment of curriculum and assessment by mapping a task (what is demanded of students and what is valued in the marking guide) to the content of the corresponding syllabus document.

To inform its work this working party:

- revisited the 2003 cycle and verified that tasks can be linked to the syllabus;
- documented the techniques used to do this and
- was to apply the techniques to the 2004 cycle.

This working party comprised teachers, members from the teachers' union, QSA, EQ's curriculum branch and ANBB. All had extensive experience with syllabus documents through writing, implementing, supporting or delivering. The methodology adopted to seek alignment of the intent of the domain to the QAT and the marking guides involved mapping and matching. The areas of alignment that were considered were content, level, emphasis and context. To ensure the

alignment of the intent, task and marking guide the following principles were identified:

- define the nominated domain so that there are clear and shared understandings of the intent;
- choose the context(s) for alignment as these need to reinforce the priorities and values of the curriculum intent;
- cue the curriculum intent in the task so that there is clear direction about expectations which will assist students to provide appropriate evidence;
- develop, panel and trial the QAT with teams that include school based personnel to utilize the expertise of those who implement the curriculum and
- include statements of intent in marking guides to reinforce the intent throughout the entire process.

It was reported by this working party that the intent matching for the 2003 cycle 'was done on the run' and would need to be revisited. It was also emphasized that the KLA syllabuses proved to be not sufficiently prescriptive regarding clear expectations of student performance (ANBB, 2004a).

#### *Working Party on Classifying Tasks:*

This working party was required to:

- design processes for deciding how each individual task in a QAT provides information about student achievement in the constructs of the QSM and
- devise a method for combining results on different assessments for reporting purposes.

To inform its work this working party:

- became fully informed of the meaning of the constructs of the QSM (the headings for reporting);
- became immersed in the intent of the 2003 QATs and
- referred to preliminary work on an approach for combining scores (criterion-keyed).

This working party classified each question using the five categories of knowledges (factual, conceptual, procedural), processing, self and others. All questions were classified in terms of the constructs first then the process of assigning relative worth to available grades began.

The working party highlighted a 'grey zone' between processing and procedural knowledge. It was concluded that work that appears initially to be processing could after practice become more familiar and be more appropriately classified as procedural knowledge. This working party also identified that some items that were common to Year 6 and Year 9 Chemistry were being classified in terms of different constructs for the two different year levels.



For any form of assessment, the working party stressed that it was important to make clear links between the intent of the question, the way the question is worded and the grade descriptors designed to capture the desirable features of student responses. This alignment it was emphasised would help avoid instances of confusion over what is being asked, the depth required, or marking guides that value things which are not clear to students from the question.

Finally the working party on classifying tasks recommended that a more suitable balance be found amongst representations of the constructs in future QAT questions and activities to maintain validity. They had identified that the 2003 QAT did not yield data on factual knowledge for either the Year 3 or the Year 9 Physical Activity. The Chemistry components of the QAT had a lower than anticipated representation of processing. This working party concluded that questions should not be included merely to increase coverage of the constructs if they are not relevant. A check in terms of classifications at the time of test development was recommended to highlight anomalies or unanticipated omissions (ARFIC, 2004b).

#### *Working Party on TGT Accreditation*

This working party was required to:

- reflect on and revise the accreditation process used in 2003;
- document the revised process and
- formalise the accreditation process in the 2004 cycle.

To inform its work this working party was to:

- canvass input from pilot school teachers who were involved in the 2003 TGT cycle.

Much of the work carried out by ARFIC working parties came to a halt and was not developed further due to a number of important factors. In the development of the 2005 QAT there was some opportunity to make use of the work of the ARFIC working parties but it appears that due to the changing policy context and limited financial resources available for the completion of the project there was a scaling back of ARFIC developments.

## THE 2005 QAT PILOT (2004 –2005)

### A Futuristic QAT

In May 2004 the Director of the ANBB reported to the Director General of Education and the Arts on Year 9 assessment and/ or testing. In this memorandum (ANBB, 2004c: 2) a potential process for standardized assessment at Year 9 to complement the existing testing program of Aspects of Literacy and Aspects of Numeracy in Years 3, 5 and 7, that would also assess a broader range of skills and knowledges in a forward-looking test format, was offered as the way forward.

The large-scale standardized testing regime that operates in Queensland is directed at students in Years 2, 3, 5, 7 and 12. These ‘tests’ serve different purposes: diagnostic (Year 2 Net), reporting of basic skills (Years 3, 5, and 7), and cross-curriculum skills (group results) for the calculation of Overall Position (OPs) (Year 12). It was reported that there was no mandated common assessment of the content of the curriculum (i.e. the KLA syllabuses) in Years 1-10, and there was no quality assurance of reported results (ibid).

It was recalled that the Assessment and Reporting Taskforce of 2001 had identified two challenges requiring response: the need to build a culture of robust, reflective assessment in schools; and the lack of system-wide data on student achievement. The proposed Assessment and Reporting Framework included mandated assessment of the content of the curriculum through common tasks but at only two (possibly three) junctures and in only two strands per KLA syllabus per year. This was termed a “dip-stick” approach.

The key issues were summarized as follows:

- the desirability of having some form of standardised testing/assessment in Year 9 to complement the existing literacy and numeracy testing program;
- the identification of the knowledges and higher-order skills that such Year 9 testing/assessment might and should measure;
- the capacity of such standardised testing/assessment to be valid, yield reliable results for reporting, make students test-wise, provide good assessment models, and build teachers’ professional expertise and
- the feasibility of the model outlined in 2003 Pilot Study and in particular “[a direction] that would sit comfortably alongside current departmental thinking on the reporting agenda ... to proceed with the in-house development of a futuristic QAT [Queensland Assessment Task, a collection of standardised tasks] ... literacy and numeracy with an emphasis on the transformation of ideas and information” (ibid).

The use of standardised testing/assessment aligned with the general educational and political climates of accountability. It was reported that the trial of the New Basics (2000—03) and the pilot study of the 2003 ARF, had involved the Department in standardised assessment and had thus

contributed to a climate in which standardised assessment/testing was likely to be accepted.

In addition it was anticipated that the perceived resistance to a Year 9 test might be overcome if it were to be more explicitly matched to the curriculum by extending the way evidence of achievement is captured. Reference was made to the 1990 Viviani Report on tertiary entrance in Queensland which required that the Year 12 core skills test be closely related to the skills taught in the senior curriculum, while also testing basic and extended English expression and numeracy and problem solving in mathematics. The process of designing and developing a standardised test that is embedded in the curriculum had a precedent that had a high level of acceptance by Queensland teachers.

Experiences in the development of the QCS Test and the ARF suggested that the assessment should also have the capacity to measure student achievement in expressing and performing, skills that cannot be captured through traditional testing regimes. Such use of authentic performance-based assessment paralleled international trends. In addition, since generic skills, literacy and numeracy were also the foci of middle schooling, it was suggested that there was no substantive reason to suggest that standardised assessment in Year 9 could not be approached in a way similar to the QCS Test, especially if it were to involve more than pen-and-paper/point-in-time testing.

Research conducted during the trial of the New Basics had suggested that increased concentration on multiliteracies was not accompanied by diminished performance in “traditional” literacy and numeracy (ibid). Many students at the upper end (Years 8 and 9) of the middle phase of learning (Years 4-9) do not gain the experience in standardised testing/assessment that they need, to become test-wise to optimise their performances in the high-stakes environment they meet later in Year 12 with the QCS Test (and other external tests) (ibid).

Teachers in New Basics schools had received professional development in identifying standards and grading student work during the four stage moderation strategy for assessing Rich Tasks. Teachers involved in the 2003 pilot study of EQ’s proposed ARF also had received incidental professional development when they were associated with the administration and marking of QATs and with the development, accreditation, administration and marking of teacher-generated tasks (TGTs). The New Basics research revealed that the most effective way to change classroom practice was to have a standardised element in the assessment system. It was suggested that a level of expertise in test/task development (QCS Test, Rich Tasks, QATs) that could be called upon to deliver a new assessment/testing device at Year 9 existed within the Portfolio of Education and the Arts.

However, it was emphasized that substantial additional funds would be needed for the development and quality assurance loops that research and

experience had shown to be necessary in the production of valid and reliable instruments.

In April 2004 the Director General met with Professor Claire-Wyatt Smith (chair of ARFIC) and the Director (ANB). The proposed “futuristic” QAT was portrayed not only as the answer to the Year 9 “non-test” question but also as the opportunity for Queensland to lead assessment policy and practice in this area. It was suggested that it was unlikely that other States would move to testing at Year 9 in such an innovative manner. It was recommended that the following ideas be the subject of further discussion.

- Expose students in Year 9 to standardised assessment with the following features:
  - common assessment tasks, set centrally, marked by teachers, and accompanied by quality assurance loops;
  - capacity to measure literacy and numeracy;
  - capacity to measure generic skills (including expressing and performing) and the multiliteracies (spatial, audio etc.);
  - multi-modal (including the use of ICTs—computer, word-processing program, scanner, digital camera and publishing program);
  - not one-off, single point-in-time judgment (could be selective updating until cut-off date, not unlike the Rich Task model in New Basics).
  
- Design the first such Year 9 assessment to incorporate the following:
  - visual tools that are useful in transforming ideas and information (graphs, charts, maps, lists and sets of information);
  - opportunities for students to demonstrate their literacy skills over a wide range of oral, print and multimedia; aspects of the Arts and SOSE syllabus outcomes.
  
- Allocation of additional funds to ensure high-quality assessment instruments possessing the required technical properties.
  
- Investigation of possible sources of funding through the QSA, in terms of joint development of the Year 9 assessment instrument.
  
- Administration of the first such common assessment in 2006 (end of Year 9 or, possibly, at the beginning of Year 10).

### **The 2005 QAT**

The pilot study of the Assessment and Reporting Framework continued in the development of the 2005 Queensland Assessment Task. As described in the background information the intent was still to capture rich information about student achievement in nominated domains (KLA strands or New Basics fields of knowledge) and across them, in various mediums using a variety of instruments, devices and strategies. This pilot project involved the design and

development of a QAT for Year 9 students. This was administered in August of 2005. In October 2005 trained teachers marked the students' responses and the results of student achievement were ready for reporting in December 2005.

This Year 9 QAT was set in the context of SOSE and the Arts. Students were required to work in multiple modes to transform ideas and information and their achievement was measured in the underpinning repertoires and generic skills.

The 2005 QAT was made up of two separate standardized tasks, although the initial intention had been to have three as follows:

- Computer-based task that requires students to respond to subtasks with a computerized interface;
- Collection of constructed-response tasks in pen-and-paper format and
- Performance-based task.

The performance-based task was not administered in 2005 trial and therefore does not form part of this evaluation.

#### *QAT Design Brief: 2004-2005*

The design brief indicated that a single QAT would be produced for 2004-2005 to be administered to Year 9 students. The QAT was to be made up of three standardized assessment tasks in the different assessment modes of computer-based, constructed response and performance-based. It was also to be complemented by a corresponding TGT for 2004-2005. The QAT was to be intellectually challenging and would have connections to the wide world. The QAT would assess the student's achievements in transforming ideas and/or information and in the underlying generic skills and dispositions. The QAT would draw on SOSE and the Arts and would be pitched at Level 5. The QAT would also incorporate multiliteracies (including ICTs). It had to also provide assessment data on processing (transforming ideas and/or information) and it was suggested that the QAT might also provide assessment data on knowledges (facts, concepts, procedures) from SOSE and the Arts, self and others. Measures of performance in literacy and numeracy were also specified in the brief. The evidence of achievement would be the student's responses to the tasks. Where the assigning of grades was not automated, the quality of student data would be judged by teacher-assessors who would assign grades after referring to centrally-set, task-specific marking guides containing verbal descriptors of available grades.

#### *Financing the Project*

In November 2004 in a memorandum to the Assistant Director-General, Office of Planning, Resourcing and Performance, the Director of the Assessment and New Basics Branch outlined that the initial budget proposal for the Assessment and New Basics Branch included \$1million for Growing an Assessment Culture which included Teacher-Generated Tasks and the development of the QAT. The budget allocation was reduced to \$650 000.

To complete the development of the QAT, which included the trial of the use of computers in a prototype and the involvement of teachers in the development of assurance loops, additional funding of \$200 000 was requested.

The team (Total 6.6 FTEs) consisted of the following overall appointments for the duration of the project July 2004 – December 2005, a proportion of 18 months on the project:

- 1.4 Programmers
- 0.5 Developers
- 0.4 Quality reviewers
- 1.0 Project leader
- 0.8 IT developer
- 0.4 Consultant and
- 2.1 Administrators

The physical resource requirements for this project included: software licences, computer hardware, sorting and storage for student work, venue for marking operation and security for the final form of assessment tasks (ANBB, 2005:6).

### *Rationale for Using ICT*

In 2005 it was decided that the processing construct of transforming would form the basis for assessment and that this transformation of information and ideas was particularly suited to the use of technology. This would also allow for the use of the computer-based capacity for data manipulation.

The decision to include a computer-based medium was also based on the design to:

- Capitalize on the capabilities of technology (computers) to allow students to display knowledge and skills in new dimensions;
- Enable the use of multimodal stimulus materials and
- Widen the scope of stimulus material.

The computer-based task was designed so that students:

- Were provided with help/hints/clues;
- Were provided with opportunities for feedback as they completed the (sub)task;
- Could generate their responses electronically.

The computer-based task was also designed so that student responses:

- Were captured electronically for the marking process;
- Could be computer-marked, for some (sub)tasks, by using a different algorithm for each;
- Could be marked electronically and remotely (as opposed to teacher-assessors at a central location).

Dr Randy Bennett (from Educational Testing Service) an international expert on the use of computers in all aspects of educational testing was consulted. He was provided with the QAT specifications and early drafts of the documentation for the proposed 2005 QAT. In September 2004 he concluded that:

- The concept of a computer-based task was a good one that extended the fairly common practice of testing in multiple-choice format into new territory;
- It would be important to have a clear view of the curriculum domains and content standards pertinent to the assessment task;
- Programs for drawing often merely substitute for what is better done on paper. Spread sheeting was suggested as a useful and powerful mode for this type of approach;
- Care was needed not to overcrowd any one screen and emphasized road testing with students to inform this aspect;
- Field testing with students was recommended early in the development phase.

He also suggested that PowerPoint could be used to facilitate a mock-up of a series of screens to present to students. This form of simulation of the intended task environment would provide feedback, for example, on whether the students know what they have to do. The development process would be iterative and involve personnel who would later write the programs after the simulations had been performed. The rationale for the choices made would also need to be documented.

### *The Prototype*

In developing a fully working prototype of the computer-based task it was intended that the initial feedback would help in the design of the interface of the program. It was also considered essential to trial an assessment in the chosen schools under similar conditions to the intended QAT. This trial would help determine, for example, if the student was able to operate the prototype, follow the instructions and operate the mouse with sufficient dexterity. It would also be useful to ascertain if schools could set-up computers and the testing environment, provide sufficient computers in working order, load software, supervise the computer-based task to handle any emergent problems such as a computer crash, capture data from assessment and clear or re-load computers for multiple sittings if necessary.

The feedback would also provide valuable information about the layout of the computer lab, the approach adopted by students to solving the task on computer (e.g. trial and error), the use of computer based help, the timing for the task, the readability of the computer screen in assessment, ability of the program to cope with all student approaches, the quality of operational documentation and, in addition, provide information about the implications for improvement. In terms of data capture it was expected that the trial would track student work on the computer, transfer data from student computers to a central database, determine the accuracy of the data, enable the analysis of data and trial computer marking.

### *The Design and Production Cycle*

Throughout the design and the development of the QAT the Project Director presented regular reports to the Director General of the Department of Education and the Arts. These reports summarized the progress made in the project against key activities, timelines and outcomes. Any risks or emergent issues were identified and how these were addressed was reported. On several occasions outside expertise was sought to provide advice on technical issues.

Due to the innovative nature of the computer-based task, its design and production cycle were significant components of the developmental phase. Frequent observations of students were essential to investigate, for example, how students interacted with the medium, and to establish their capacity for using computers in assessment. Vital information for development purposes was obtained through the ongoing experimentation using the prototype in the form of a smaller, related but different, subtask and through exploring the stages in the process of developing and refining the computer-based task. The design, production and trial of the prototype informed the final computer-based task design and provided information on related hardware issues, as well as informing administrative procedures related to space and online help.

The specific stages in the development of the computer-based task involved: first the task design, the selection of the most appropriate methods, the development of subtasks, the evaluation and refinement of subtasks, the development of a database with the final version and accompanying documentation completing the cycle of development.

Similarly the constructed-response task involved the development of subtasks, the revision of subtasks, the marking of subtasks, the consequent refinement of subtasks with further revision and refinement following. The layout and proof reading were the final stages prior to the packaging of the materials.

By December 2004 experiments had been conducted with students using the prototype computer interface. The analysis of the data collected informed the restructuring to reduce the trial and error by the student and to increase student engagement with the depth of intellectual rigour of the task. Tests were also conducted to determine the effectiveness of paperless instructions. Information from the initial trial resulted in significant improvement involving more use of the help system by the students when a period of forced reading was required prior to the commencement of the task. The program was developed and tested to automatically capture student results into an Excel spreadsheet.

At this stage two possible constructed response subtasks of the QAT were panelled.

The risks that were identified at this stage included:



- Availability to the project of financial and human resources required to develop a valid, high quality, innovative and multimodal QAT;
- The uncertainty regarding student capacity to interact with a computer-based medium for assessment of this type and their requisite skills;
- Development of cutting edge computer software as a significant component of the design and development phase;
- The uneven distribution of computer confidence across the State and disadvantage to some students if they did not experience a specimen task;
- The capacity (which includes the number and condition of computer resources) within schools to administer a computer-based standardized task to an entire year level and
- The capabilities of the current computer software packages to meet the requirements of the QAT while considering budget and resources.

The development of the prototype was intended to address most of these risks with the latter risk addressed by the capacity of the QAT team. They, together with ongoing research, would ensure that the control of the development would remain with the task developers so that the inadequacies of the software packages would be overcome.

## **Queensland Curriculum and Assessment Reporting (QCAR) Framework**

### *School Reporting*

In a document released by the Department of Education and the Arts in October 2004 it was stated that 'from 2006, parents of every child in every school in Queensland will receive a written report card, either in hard copy or on-line at least twice a year' (Department of Education and the Arts, 2004: 2) with parent-teacher interviews offered each semester. A common framework for school report cards with a consistent five-point results scale forms part of the QCAR Framework agenda.

In addition every child at every school in the state would have a unique student identifier to assist schools to track students who move from one school to another, and to assist in tailoring services to meet each child's individual needs. The Queensland Studies Authority (QSA) manages this system.

Annual reporting by schools is now required. The information includes contextual information such as the curriculum taught, the opportunities for parental involvement and extra-curricular activities, as well as outcomes data such as summary information in the literacy and numeracy tests and retention rates. From mid-2006, it was stated that each school will be required to publish annual reporting information on its school website (ibid:3).

### *Policy Direction of QCAR Framework*

In April 2005, the Department of Education and the Arts in collaboration with the Queensland Catholic Education Commission, the Association for Independent Schools Queensland, Education Queensland and the Queensland Studies Authority were developing the policy direction for the Queensland Curriculum and Assessment Reporting (QCAR) Framework. The intent is that the framework will improve student learning and increase comparability of assessment and reporting across schools. In particular, the framework aims to align: curriculum; teaching; assessment and reporting.

The rationale for the new framework stems from concerns about the lack of clarity around what must be taught across schools and what standards of achievement are expected. It is anticipated that the new framework will also address concerns raised by teachers and the community about the amount of material to be taught in Years 1-10 curriculum that appears to be hindering in-depth learning.

In providing the framework it is expected that teachers will be supported by a shared understanding of, and common language for describing, what students are expected to know, understand and be able to do in the essential curriculum. It is predicted that the identified standards for the essential learnings will help teachers judge how well a student is performing at a particular stage of schooling. A bank of assessment tools, to complement school-devised assessment, aims to provide high-quality assessment resources that teachers can choose to use.

It is further anticipated that 'rigorous comparable assessment against defined standards at key points will allow students to demonstrate a broad range of learning and deep understanding' (Department of Education and the Arts, 2005a: 3). It is also hoped that the framework will help to support school and system-level planning by providing more comprehensive and comparable data on school performance. For parents it is suggested that the framework will deliver easy-to-read reports that show how well their children are performing compared with others and what is expected at each year level (ibid).

The Queensland Studies Authority is developing the materials and tools for Queensland schools which includes: the essential learnings; the standards of student achievement; the bank of assessment tools; the rigorous comparable assessment and reporting framework (ibid:4). Consultation with key stakeholders and trials are planned to take place under QCAR Framework in 2006 with statewide implementation in 2008.

To develop the policy direction two groups were established in April 2005. First, the Policy Steering Committee, which is to provide advice to the Minister for Education and the Arts and includes the chief executive officers of the Catholic, Independent and state education sectors and the Queensland Studies Authority. The second group, the Expert Advisory Group, is comprised of leading academics and expert practitioners. This group is to

provide independent advice on the intellectual rigour of the policy direction for the framework, particularly in terms of theoretical and technical issues.

### *Funding*

In the State Budget Highlights 2005-06, \$8.3 million was allocated to implement the QCAR Framework, \$2.2 million of this amount was specifically allocated for the period 2005-06 (Department of Education and the Arts, 2005b: 3).

### *Rationale and Strategy*

In July 2005 a technical paper, *Background, rationale and specifications: Queensland Curriculum, Assessment and Reporting*, prepared by Professor Peter Freebody, provided the rationale and strategy for the implementation of the framework (Freebody, 2005: 4). The rationale could be summarized as follows:

- Many teachers are 'sampling' the curriculum as revealed in the 2003 Pilot Study;
- Practices of assessment are variable and teachers need support to develop and use assessment practices that are valid and reliable;
- Standards based assessment is an important lever for change and
- Teacher competence and commitment need to be sustained and developed.

The suggested strategy was derived from the understanding that standards can be important levers for change in aligning curriculum, assessment and reporting. The identification of essential learnings was considered fundamental to the processes of establishing and applying standards (ibid:8).

The assessment options included the common assessment approach to provide reliable, system-wide data and inclusion of school-based assessment tasks that should be socially moderated. It was emphasized that the framework should emerge from the empirical and conceptual development of the standards, essential learnings and assessment tasks (ibid: 13). Finally it was reported that the approach adopted for the development of the framework should be evidence-based and incorporate:

- a statement of essential capabilities (learnings);
- a set of standards of student achievement;
- assessment tasks, reflective of the suggested guidelines;
- a bank of assessment tools for teacher use;
- a set of principles to guide how assessment will occur;
- a set of common tasks;
- quality assurance processes for the use of the standards and the development of assessment tasks;
- an accessible framework for reporting and
- with the reporting junctures at Years 4, 6 and 9 (ibid: 13-14).

## *Queensland Curriculum, Assessment and Reporting Framework*

Following the publication of the technical paper (Freebody, 2005) the *Queensland Curriculum, Assessment and Reporting Framework (Department of Education and the Arts, 2005c)* was published.

It was reported that the new framework would:

- make clear statements about the essential learnings that must be taught in Queensland schools;
- provide a common frame of reference and a shared language for communicating student achievement;
- equip teachers with high-quality assessment tools for collecting evidence of student achievement;
- promote teachers' professional learning, focused on good assessment practices and judgement of the quality of student achievement against statewide standards;
- introduce statewide assessment of student learning in Years 4, 6 and 9 and
- provide more meaningful reports of student achievement (Department of Education and the Arts, 2005c: 2).

Essential learnings are to be identified from Queensland's eight KLA syllabuses of The Arts, English, Health and Physical Education, Languages other than English, Mathematics, Science, Studies of Society and Environment, and Technology.

Standards are defined as 'descriptors of student achievement used to monitor growth in student learning and provide information about the quality of student achievement' (ibid: 6) and are to be linked to the essential learnings. To enhance teachers' understanding of these standards, descriptions with exemplars of student work for each standard will be provided.

The Queensland Studies Authority, in collaboration with teacher and professional expertise, is responsible for the development of the standards through analyses of student work, curriculum materials and sequencing of knowledge, skills and attributes of the essential learnings.

To support teachers' classroom assessment practice an assessment bank of high-quality assessment tools for collecting valid and reliable evidence of student achievement will be provided (ibid: 7). In addition there will be support to promote assessment knowledge and skills, models of good assessment practice and resources to support consistency of teacher judgments about student achievement (ibid).

Statewide assessment will take place at Years 4, 6 and 9 to enable a focus on continuity of learning in the middle years and to facilitate comparison with Years 3, 5 and 7 statewide tests of basic literacy and numeracy skills. In 2007 a Year 9 test of basic skills is planned (ibid: 8).

QSA will develop a quality assurance process to ensure that reliable and valid evidence of student achievement is reported. The assessment tasks will involve 'authentic and complex tasks that allow students to demonstrate their breadth and depth of understanding in the essential learnings' (ibid:9). These tasks will be completed under common conditions. These 'common statewide assessment tasks in Years 4, 6 and 9 will be used to collect evidence of student achievement' (ibid). Non-state schools will be permitted to use locally devised tasks that meet quality assessment criteria developed and monitored by QSA.

To maximize consistency of teacher judgments the following strategies will be implemented:

- administration and marking guides will accompany every assessment task;
- for common assessment tasks there will be reporting guides and exemplars of student work to illustrate each standard;
- intra and inter-school comparisons of teachers' judgments of student achievement will be possible and
- QSA will develop processes for validating teacher judgments on statewide assessment.

### *Reporting*

The reasons given for reporting on student achievement include the teachers' and schools' requirements for such data to plan teaching and learning and to monitor the impact of school improvement strategies. In addition it has been emphasized that school authorities need data about particular student cohorts to assist in system-level planning, to identify resource and professional development priorities and for accountability to the community.

A reporting framework will be introduced so that common elements, such as descriptions of the taught curriculum, expected standards and the achieved standard, will be reported for every child.

These reports will occur twice a year and will include a five-point scale for describing student achievement across all areas of learning between Preparatory year and Year 10.

From 2008 in years 4, 6 and 9 student reports will include more information about student achievement in the essential learnings. Parents will be informed of their child's achievements on the assessment and how their child's achievement compares with what is expected for their year level.

QSA will collate the data from the statewide assessment of student achievement in years 4, 6 and 9 and provide summary reports to education sectors and the community on how well Queensland students are achieving and schools are performing (ibid:12).

It is anticipated that all schools will be able to trial the first common statewide assessment tasks in late 2006 with continuing trials in 2007. By 2008 all

elements of the framework will be ready for use in Queensland schools (ibid:14).

## EVALUATION METHODOLOGY

### Context

This evaluation was conducted in the context of the development of the Queensland Curriculum, Assessment and Reporting (QCAR) Framework. This new policy will:

- define the **essential** learnings that must be taught in Queensland schools;
- provide **standards** that describe student achievement in the essential learnings;
- provide a **bank of assessment tools** for collecting evidence of student achievement by methods that require responses that are: short answer, extended, practical, performance-based or oral;
- promote teachers' professional learning related to assessment;
- introduce comparable **statewide assessment** of student learning, in Years 4, 6 and 9, in the essential learnings with a focus on English, maths and science and one other area and
- provide a **common reporting framework that** will describe student achievement using a common five point scale.

This evaluation examined the implications for action from the experience and learning associated with the QATs to inform the implementation of the QCAR Framework. It was designed to meet the needs of the immediate stakeholders (the policy-makers and funders). The focus of the data collection was therefore directed towards implementation issues and outcomes measures. In completing this evaluation the evaluator adopted an independent role.

### Evaluation Standards

In adopting an independent role the evaluator drew on the American Educational Research Association's (AERA) Position Statement Concerning High-Stakes Testing in PreK-12 Education. This statement, based on the 1999 Standards for Educational and Psychological Testing, represents a professional consensus on Standards concerning sound and appropriate test use in education and psychology. These standards are endorsed by AERA, the American Psychological Association (APA) and the National Council on Measurement in Education (NCME) and as such remain the most comprehensive and authoritative statement by the AERA concerning appropriate test use and interpretation (AERA, 2000:1).

In the US, states mandate assessment programmes to gather student achievement data over time and to hold schools and students accountable. When achievement results lead to serious consequences for students, or for educators, they are described as 'high-stakes'. To illustrate, if schools are judged highly according to school-wide average results of their students they can

attract public approval or financial rewards. However, low results can lead to public embarrassment or heavy sanctions. High-stakes assessments are enacted by policy makers to improve education and the setting of high standards of achievement can inspire greater effort on the part of students, teachers and principals. Reporting of results can focus public attention on achievement differences among schools or student groups.

The inadequacy of high-stakes assessments, or the lack of sufficient reliability or validity for their intended purposes, has the potential for unintended and harmful consequences. Policy makers can be misled by 'spurious' increases in assessment results that don't relate to improved learning; students may be placed at increased risk of failure or disengagement from schooling; teachers may be blamed or punished for inequitable resources which remains beyond their control; and curriculum and teaching can become distorted if high grades or results per se, rather than learning, become the overriding goal (ibid).

The QATs were never designed to fulfil a high-stakes role, however, in the context of the implementation of the QCAR Framework and the plans for statewide assessment of the essential learnings in years 4, 6 and 9, it is important to consider these conditions for implementation and continuation purposes. Where relevant these standards have been used to guide the evaluation and to help the evaluator maintain an independent perspective. The essential conditions to sound implementation of high-stakes educational assessment programmes are as follows:

- protection against high-stakes decisions based on a single test;
- adequate resources and opportunity to learn;
- validation for each separate intended use of the high-stakes assessment;
- alignment between the assessment and the curriculum;
- the validity of the passing scores and achievement levels;
- opportunities for meaningful remediation for examinees who fail high stakes assessments;
- appropriate attention to students with disabilities;
- appropriate attention to language differences among examinees;
- careful adherence to explicit rules for determining which students are to be tested;
- sufficient reliability for each intended use and
- ongoing evaluation of intended and unintended effects of high-stakes testing (Ibid: 2-5).

## **Framework**

The framework for the evaluation (see Figure 1) took into consideration the contextual factors at international, national and local levels. These contextual factors were at the level of international and national policies on curriculum and assessment; at the state policy level in terms of the levers for change. These include:



- Alignment of curriculum and assessment;
- QSA KLAs syllabus;
- QCAR Framework;
- Common Assessment Tasks (CATs) and quality assured, locally designed assessment tasks (LATs)
- Moderation and
- Reporting using standards.

Another contextual factor was the pilot study of the QATs with specific focus on:

- QAT design brief and specifications;
- QAT technical considerations that include the validity and reliability of student achievement data and
- The implications for the QCAR Framework

At the local professional level it was important that the evaluation seek information concerning the challenges posed to schools and to ascertain the relevance, readiness and resources needed.

The Queensland Assessment Task (QAT) project was described as part of a pilot study of the Assessment and Reporting Framework, and as a pilot, the intention has been for improvement in the overall design, development, implementation and continuation of the use of common assessment tasks.

It is important to stress that the evaluation began while the project was finalizing the assessment and marking of the computer-based and constructed response tasks.

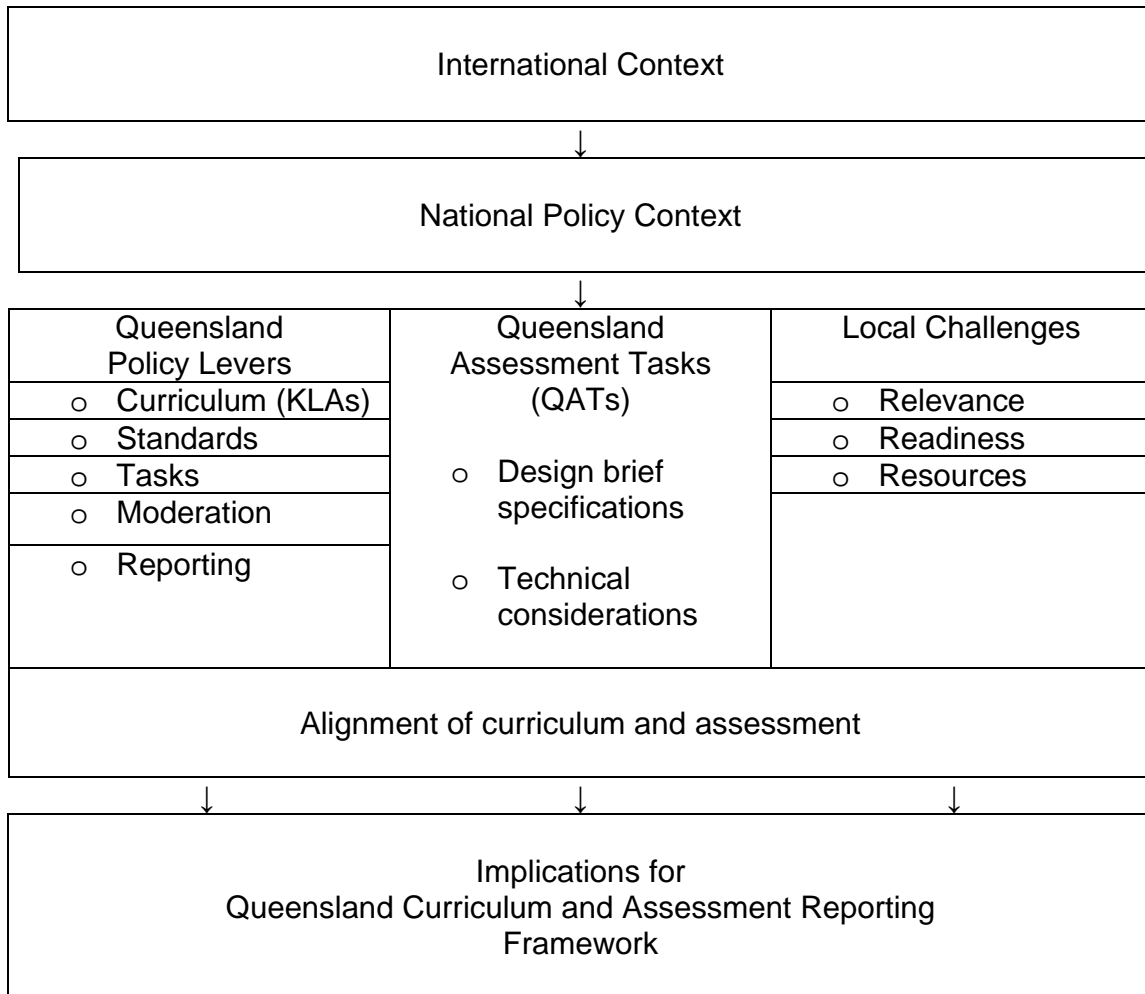
A combination of qualitative and quantitative methods was used to be responsive to the needs of stakeholders as far as possible.

## **Change Process**

Current, international themes that dominate educational change efforts include how to achieve a large-scale reform while sustaining improvement. This evaluation of the Queensland Assessment Tasks (QATs) was contextualized in this educational reform agenda that parallels those at the national level. To illustrate, the Commonwealth Minister for Education, Science and Training is calling for 'national consistency in education' that involves assessing students against national benchmarks and reporting these to parents. Consistent with this is the Queensland Government's planned introduction of QCAR Framework for the purposes of:

- improving student learning;
- increasing comparability of assessment and
- improving consistency of reporting across all Queensland schools from P-10.

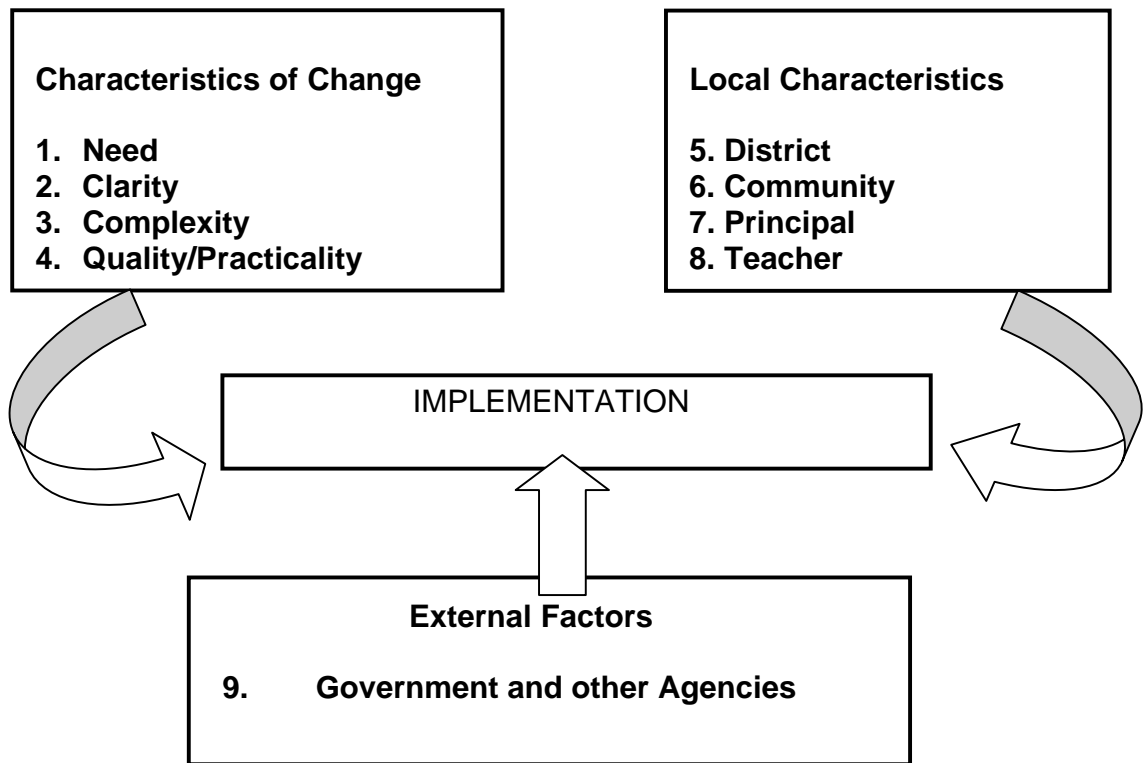
**Figure 1: Framework for the Evaluation of QATs**



The planned strategy is to use a combination of assessment tasks: common assessment tasks *and* locally devised, quality assured tasks. Through the completion of these tasks students are provided with the opportunity to demonstrate their learning that will be assessed using standards, with these results reported to parents. In this way the common assessment tasks operate at the interface between the central political demand for accountability and the local professional level of curriculum planning and development, teaching and learning and assessment practices. Assessment provides evidence to inform decisions at both school and system levels although the data required will be in different formats to fulfil each demand.

In planning the implementation of such a change it will be important to consider the interactive factors (see Figure 2) that will affect the take-up of the intended change in assessment practices.

**Figure 2: Interactive factors affecting implementation**



### **Aims**

The evaluation aimed to assess the design and implementation of the QAT across the following evaluation dimensions:

- design brief and specifications;
- technical considerations including the validity and reliability of student achievement data;
- alignment of curriculum and assessment and
- policy implications.

Some of the standards as described above and endorsed by AERA, the American Psychological Association (APA) and the National Council on Measurement in Education (NCME) as the most comprehensive and authoritative statement concerning appropriate test use and interpretation (AERA, 2000:1) were used to guide the evaluation.

The specific research questions align with the four specified evaluation dimensions of the design brief and specifications, technical considerations, alignment of curriculum and assessment and the policy context of the Queensland Curriculum and Assessment Reporting Framework. Specifically the research questions were as follows.

- Design Brief and Specifications
  - Did the QAT meet the design brief and the design specifications?
    - Identify what the design brief was and how it was determined.
    - Assess the role of teachers in this process.
  
- Technical Considerations
  - How valid and reliable is the QAT?
    - Evaluate the validity and reliability of the key elements of the QAT and the quality of the student achievement data.
    - Identify processes and technical procedures in the development of the QAT, its administration, the trials and their management, the role of schools and teachers
  - How were marking guides developed and handled?
    - Identify how the tasks were assessed and the role of teachers in this process including their role in the decision-making concerning the standards
  - How useful is the experience of the QAT for the development of future CATs for years 4, 6 & 9?
  
- Alignment of Curriculum and Assessment
  - How aligned are the curriculum (SOSE and the Arts) and the assessment tasks (QAT)?
    - Examine the role of the developers, schools in the design, development and implementation
  
- Policy Context
  - What is the value of the QATs given the context of QCAR?
    - Establish if the development of the QATs was part of a planned strategy
    - Identify the conceptual background for the development of QATs and the current focus on QAT as opposed to the TGTs
    - Identify what the key differences are in developing QATs and TGTs for trial purposes as opposed to developing assessment tasks in the policy context of the QCAR Framework

## **Methodology**

To meet the aims and objectives, the evaluation was essentially qualitative, but did include quantitative analysis of new and – where possible – existing data collected via questionnaire responses. The evaluator liaised directly with officials from EQ and QSA before the work started to avoid duplication and to collect useful background.

The qualitative data was used to illuminate, interpret and explain statistical trends in the quantitative data; to explore more detailed aspects and issues associated with the implementation of the QAT; and to present more subtle and complex perceptions of teachers and students' perceptions and experiences of the QAT. The school sample was demographically and geographically representative, including state and independent schools.

Qualitative and quantitative information was gathered after the development, implementation and during the assessment and marking of the QAT. Data was collected via analysis of documents, observations, interviews and surveys.

### *Scope of the Research Project*

To meet the principal evaluation objectives priority was given to:

- Review and analysis of documents, reports, relevant literature and data and information available from the development and trial phases of the QAT;
- Classroom observations of the administration of the QAT from two schools and the inclusion of teachers/principals/QAT co-ordinators from these schools in the interview sample;
- Focus group interviews with students from these two schools (30 interviewees);
- Survey of students' immediate responses after administration of the QAT in two different schools from those where the observations and focus group interviews took place (227 total responses);
- Questionnaires to 20 students from each of the 56 schools (794 out of a notional total of 1120 total responses).
- Telephone interviews with 28 principals and teachers from 27 of the sample of 56 schools (Total 28 interviews).
- Focus group interviews with teachers from 8 schools during the time of the marking of the QAT, these were different schools to the 27 selected for the telephone interviews. (Total 12 interviewees) and
- Focus group interviews of the technical support team during the time of the marking of the QAT. (Total 12 interviewees).

Questionnaires to principals/QAT coordinators (see Appendix 1) and teachers (see Appendix 2), from the sample of QAT schools, were sent out electronically by administrative and computing support from Brisbane. The questionnaires were distributed to 56 principals and to 56 teachers at each school. These questionnaires sought qualitative and quantitative responses and were essentially 'attitudinal' and experiential in nature. The function of the questionnaires was to lead the evaluation to key questions and issues which could be further explored during the interview phase. For instance, principals' views of the effects of the QAT on the implementation, manageability and effects on curriculum and teaching practice were investigated.

The evaluator was located in Townsville and found it difficult to ensure that the electronic surveys were sent out in a timely manner and followed up with reminders by the support staff in Brisbane. The colleagues who were offering support were not full time employees and therefore could not commit to completing these tasks throughout the period of the evaluation. Consequently the response rate was inconsistent and those respondents who did complete the questionnaire were too few to make the numbers significant. This constitutes one of the limitations of the study.

### *Selection of Schools*

The schools chosen to take part in the study were selected with the help of the Manager of the QAT team and her team's own sampling-data and mechanisms.

In order for findings to be as generalisable as possible, but at the same time to take full account of local differences (including the impact on curriculum and assessment reforms of such matters as location and intake of school, or the approach of the principal, community leader or district to the QATs), schools were chosen with the aim of covering a wide demographic and geographic spectrum – in particular, including schools in urban, suburban, rural and remote locations.

Schools from both state and independent education systems including those with students with special educational needs were chosen. The fifty-six schools involved in the pilot were included in the sample, including schools from different parts of the state (see Figure 3).

**Figure 3: Sample of Schools\***

	State	Independent	New Basic	Total
Urban	0	1	0	1
Suburban	8	4	2	14
Rural/Town	23	5	6	34
Remote	4	0	3	7
Totals	35	10	11	56

\* One school administered the QAT with provision for special educational needs.

### *Sample*

Each sample of teachers was as varied as possible in terms of age, gender, ethnic background, length of time in teaching, position in the school and specific teaching specialisms. This was achieved through consultation with the Manager of the QAT team and team members. Twenty seven of the fifty-six participating schools were selected and teachers or principals interviewed.

Students were selected for two focus group interviews after their experience of completing the QAT. It was particularly important to interview and survey the Year 9 students, who at this middle stage of their secondary-school careers had a great deal to tell us in terms of the success or otherwise of the QATs in aligning curriculum content and assessment practices.

To meet the requirement of evaluating the validity and reliability of the assessment of the QATs on issues of inclusion and equal opportunities, the research tried to ensure that pupil groups included a range, including, where possible, bilingual students, students from urban and rural locations, students with special needs or with emotional and behavioural difficulties.

Interview and survey responses were supported by two days' classroom observations of the administration and student completion of the QAT: the computer-based task that required students to respond to subtasks with a

computerized interface and the collection of constructed-response tasks in pen-and-paper format.

Classroom observations in two sample schools included observations of the completions of both tasks and the administration and implementation of the tasks by the teachers, in order to monitor the impact of the QATs. These observations were followed up with focus group interviews with the students and discussions with those teachers who administered the tasks and were responsible for the technical and ICT considerations of the task.

Observations took place at the two schools from which two -teacher and principal/QAT co-ordinator- interview samples were drawn and where the student focus group interviews were conducted. Each observation lasted for the duration of the completion of the QATs.

### *Questionnaires and Interviews*

The study used student questionnaires (see Appendix 3) and interviews involving teacher markers (see Appendix 4) and QAT developers (see Appendix 5), and teachers/principals/ QAT developers (see Appendix 6). These methods were specifically designed to generate two kinds of data:

- data relating to the impact of the QAT on teachers themselves, including details of their responses to their role in the administration of the QAT, the impact of the QAT on teaching of the curriculum, teachers' views concerning key elements of the QAT and any suggestions they had for improvements or modifications for the development of future Common Assessment Tasks and their implementation;
- data relating to the impact of the QAT on students' performance, motivation and attitudes to the assessment tasks; from students' own responses to the QAT and from teachers' and principals' responses to their roles in the administration and assessment of the tasks and the decision-making processes concerning the establishment of the standards, their views regarding the alignment of the curriculum and the assessment tasks and the value of the QAT for development and implementation of common assessment tasks.

The emphasis on teacher (rather than pupil) interviews reflected the nature of the central questions being addressed, several of which students were in no position to comment. By way of triangulation, however, and to take account of the possibility of over-subjectivity in the teacher and principal/QAT co-ordinator responses in relation to the interview data, it was also necessary to interview students at some of the schools and to observe the administration of QATs in these schools. While the time-scale of the project inevitably limited the size of the sample, difficulties were minimised by selecting two schools from the sample that were varied in terms of location, school type and intake.

Given the scale and timespan of the project and the time and cost entailed in transcription, interviews were taped but not all were transcribed fully. Rather, a summary-sheet with relevant headings was devised for completion at the time of some of the interviews. In the case of group interviews (of students) the summary findings were read back to the group for validation purposes.

### *Analysis*

SPlus software was used for analysis of quantitative data collected from the questionnaires. The evaluator with the help of the Manager of the QAT team and her team members created approaches to data analysis that were consistent across the team and that facilitated qualitative description of the sample as a whole.

### **Code of Practice**

A code of research conduct was agreed with the schools, reflecting the ethics procedures of James Cook University and that of Education Queensland. All require parental permission for student interviews.



## DESIGN BRIEF AND SPECIFICATIONS

### The Construct of Transforming

In 2005 it was decided that the processing construct of transforming would form the basis for assessment and that transformation of information and ideas and the underlying generic skills and dispositions were particularly suited to the use of technology. SOSE and the Arts were the chosen disciplines in which the task was contextualised, this coincided with the roll out of these syllabuses from QSA. The QAT aimed to incorporate more than just a pen-and-paper test and did not take the form of a traditional exam of these disciplines.

The guiding construct for the item writing for the 2005 QAT:

“was not factual and conceptual, but focused on kids’ capacity to process. There was a need identified to move towards stronger links with the purposes of education for the 21<sup>st</sup> century and repertoires of practice while drawing from the disciplines and constructing assessment tasks in a transdisciplinary way”.

The QAT development team decided that the assessment task needed to be “thoughtful but adventurous”. The intention was also to gather student achievement data in terms of literacy and numeracy.

The reasons given for the adoption of this approach are many and most have been outlined in Chapters One and Two of this report. What is important at this point to emphasise is that a more generic assessment, achieved by focusing on the construct of processing, provided the solution to the existing curriculum context in Queensland schools. Teachers use a wide-range of curriculum materials and approaches that provide students with a range of learning experiences. However, if the developers were to assess curriculum content in a standardized way they could not assume that students had learnt the same content in SOSE and the Arts throughout the state. A generic assessment approach as opposed to a curriculum content approach was therefore chosen for the development of the assessment tasks.

### Did the QAT meet the design brief and the design specifications?

It was the Project Director, in consultation with the Director General of Education and the Assessment and Reporting Framework Implementation Committee, who took responsibility for the design brief and specifications. As explained in Chapter Two the design brief in 2005 was part of the specifications that had been written prior to the 2003 QAT pilot. However, given the findings of the 2003 pilot and the funding situation the brief was developed particularly for 2005 to fit within this context, the overall specifications and the changing policy context.

## Transformation

The decision to focus on Year 9 and pitch the QAT at level 5 aligned with the federal agenda for more nationally consistent schools and the then Minister for Education and Minister for the Arts' priority to improve students' literacy and numeracy particularly in the middle phase of learning of Years 4-9.

Using the vehicle of SOSE and the Arts the development team focused on the generic skills underlying transformation of ideas and information, numeracy, literacy, ICTs, multiliteracies and the four types of transformation that had been developed and categorized by one of the developers in consultation with other key team members. The classification used is described in Table 1.

Table 1 Types of Transformation (Education Queensland, 2004b)

Type 1 <i>Genre</i> changes, but not the <i>medium</i> . The natures of <i>information and ideas</i> remain highly similar.			
	Information and ideas	Medium	Genre
Subject of the transformation	Scientific report on connections between a prescription drug and potential side effects	Written text	Scientific report
Product of the transformation	Leaflet to accompany the drug warning of possible side effects	Written text	Informational leaflet
Type 2 <i>Medium</i> changes, and hence <i>genre</i> . The natures of <i>information and ideas</i> remain highly similar.			
	Information and ideas	Medium	Genre
Subject of the transformation	Data and analysis section of a scientific report on connections between a prescription drug and potential side effects	Written text	Graphical and tabular data analysis
Product of the transformation	Series of posters for campaign informing the public of possible side effects of the drug	Visual art	Poster
Type 3 There is a substantial shift in the nature of <i>information and ideas</i> (the nature of the shift might be thought up by the students themselves). <i>Medium</i> changes, and hence <i>genre</i> will need to change.			
	Information and ideas	Medium	Genre
Subject of the transformation	Scientific report on connections between a prescription drug and potential side effects	Written text	Scientific report
Product of the transformation	Dramatic exploration of the impact of exaggerated reactions to drugs on patients and their lives.	Drama	Comic play
Type 4 There is still a substantial shift in the nature of <i>information and ideas</i> . (The transformation is not Type 3, though, because the <i>Medium</i> remains the same.)			
	Information and ideas	Medium	Genre
Subject of the transformation	Scientific report on connections between a prescription drug and potential side effects	Written text	Scientific report
Product of the transformation	Personification of drugs which allows them to pursue their own agendas (via the side effects)	Written text	Drama script

The identified underlying skills involved in the processes of transformation included the following: translating, presenting, designing, imagining, writing, tabulating, devising, ICT skills, synthesizing, performing, mapping, interpreting, evaluating, modifying, deciding, contrasting, analyzing, summarizing, graphing, exemplifying and sketching. The identified dispositions included: a comfort with using technology, sensitivity, logic, self-expression, inclination to experiment and a striving for accuracy.

### **The Standardised Assessment Tasks and the TGTs**

A single QAT was produced for 2004 –2005 and was administered to Year 9 students. The specifications stipulated that three standardised assessment tasks in the different assessment modes of performance-based, computer-based and constructed response would be developed but only the latter two were implemented and evaluated.

#### *Performance-based Task*

Interview data and document analyses suggest that the performance-based task was not implemented due to the lack of funding and resources. Given the context of the Arts, and the construct of transforming, the performance task was planned and designed to demonstrate the ‘achieved curriculum’ (Suter, 2004) as apparent in the student performance. This assessment dimension, while recognized as important in providing the opportunity for the student to demonstrate their artistic skills, would also address the need to increase teachers’ repertoires of assessment skills in the context of building an assessment culture. The planned performance task required students to produce an audition tape that captured their particular skill or talent. The student was also required to include a covering letter or document that explained and reviewed the audition after it had been recorded. As expressed by members of the QAT team:

“ So they (the students) have the right to say what really worked, what really didn’t and you would get a richer experience of what the performance was. For example, if the student recorded his playing of the violin, he should have the opportunity to say what he would have done better or differently if he had the chance and that would give you a richer appreciation of everything he knows about playing the violin. So these three elements (that is, the performance, constructed response and computer based task) could be assessed. But the assessment would be beyond this project so it (the performance-based task) was published but parked on the side.”

“The time and the funds were lacking to follow through on how to assess the performances. The amount of time required to listen to the tapes or to view the performances would make the assessment of performances prohibitive, it would have had to be modified to the first five minutes of the performance and a blunt measure applied. ”.

### *Computer-Based Task*

Given the lack of funding a decision was taken to concentrate efforts on the development of the computer-based task. The rationale for this strategy of how to use the available funds was based on the belief that this would provide the most valid use of the available funds and result in the best value for money. An opportunity existed for the team to be innovative in achieving a balanced approach to assessment but to also push the boundaries. The 2005 QAT aimed to be at the cutting edge of assessment development. As expressed by one team member it was about:

“... having a commitment to pushing the boundaries and having a go! Because if you don't push the boundaries you never know where you are going to get. We never thought we would have got the computer-based task to this.”

The decision to include a computer-based medium was based on the aim to capitalize on the capabilities of technology (computers) to allow students to display knowledge and skills in new dimensions and to use the computer-based capacity for data manipulation. The computer-based task was more than the completion of multiple choice type questions. The task would incorporate multimodal stimuli to widen the scope of the materials used and to allow students to display knowledge and skills in new dimensions. It was also possible for instructions to be provided in different modes. This was task-based assessment that was to incorporate these key qualities into the design of being: futuristic, innovative, rigorous, authentic, intellectually challenging, multi-modal, capable of providing valid assessment and reliable data. The marking of the computer-based task was introduced and was to be completed in two modes by the computer and electronically by teachers.

The central objectives identified by the team were to:

- Identify knowledges and higher order skills that Year 9 assessment should measure;
- Develop a standardized assessment task that was valid and yielded reliable results for reporting;
- Provide good assessment models and
- Build teachers' professional expertise.

The computer-based task consisted of seven tasks that were related to an imaginary central Queensland tourist attraction called QAT National Park. For a more detailed description of the computer-based task, the various steps involved in the development, the issues, the task itself and how well the solution worked and what might have been done differently with the benefit of hindsight refer to the *Report on Computer Based Task (QAT 2005)*, completed by Richard Owens and the *2005 Queensland Assessment Task Reflections* report (EQ, 2005).

The seven questions incorporated multiliteracies (including ICTs) which is apparent from this analysis of the computer-based task, the marking guides,

and observation of students completing the task. To illustrate, the computer-based task required the student to:

- Construct a timeline of milestones in the history of the park's development.
- Place four icons (representing flat ground, gentle slope, steep slope and unmarked peak) on the most appropriate area of the map.
- Drag six descriptive markers to an appropriate section of the road. The markers were 'attached' to the road with a level indicator providing real time feedback.
- Use the paint/draw tools to create pictorially a road sign on one of the three preset backgrounds (no words allowed).
- Create a profile of a section of the contour map by dragging height indicators up or down fixed lines. Only ten points from a total of thirty could be used.
- Describe the contrasting landscape directly east and west from a fixed point (P) positioned on the contour map. The description was typed into a text box at the bottom of the screen.
- Create an SMS message to a friend who recommended a visit to the park, giving an opinion of the park with the computer generated 'photo' to be sent.

The development team explained that evidence of achievement was the students' responses to the tasks and where the assigning of grades was not automated, the quality of student data was judged by teacher-assessors who assigned grades using centrally-set, task-specific marking guides containing verbal descriptors of available grades.

The computer-based task was designed so that students:

- Were provided with help/hints/clues;
- Were provided with opportunities for feedback as they completed the (sub)task;
- Could generate their responses electronically.

The computer-based task was also designed so that student responses:

- Were captured electronically for the marking process;
- Could be computer-marked, for some (sub)tasks, by using a different algorithm for each and
- Could be marked electronically and remotely (as opposed to teacher-assessors at a central location).

### *The Constructed-Response*

The constructed-response consisted of three parts. The first part was called "Last of their Kind" and was based on a passage that described a major heritage site in Western Australia that is also rich in natural resources. A map and two photographs, one of a panel of petroglyphs (rock engravings dating back to the last ice age) and the other of a mound of granite were included. There were fifteen questions (9 multiple choice and 6 requiring short answers). The first ten questions were about the different aspects of the area

of Western Australia with its rich natural resources and its importance to Australia's cultural heritage.

A second passage which was adapted from a newspaper article from *The Sydney Morning Herald*, 24 November 2001 was about thylacines (Tasmanian tigers). Questions 11 – 15 were based on the topics of thylacines and petroglyphs.

A third shorter passage about the extinction of the thylacine and petroglyphs was included before question 15. Students were required to complete the text on an incomplete poster that depicted a photograph of a stamp of a thylacine and a statement next to it that read "Fifty years after the thylacine was wiped off the face of the earth, we brought out a stamp to say that it was endangered." Students were asked to consider the possible destruction of the art described in the first passage by the expansion of heavy industry into the area in Western Australia. They were instructed to write short punchy text to complete the poster and a slogan at the bottom, to make people stop and think.

The intents of part one of the short answers of the constructed response (Questions 10 – 15) were to assess:

- Literacy: deconstruction of meaning, generalisation of a metaphorical statement (Question 10);
- Numeracy: recognition of the length of an interval (of time) as the difference between the values at its endpoints, and correct calculation of this difference (Question 11);
- Arts (media)/literacy: knowledge of the nature of the components of a newspaper (Question 12);
- SOSE and literacy: knowledge of, and recognition of allusion to a defining episode of the 20<sup>th</sup> century (Question 13);
- SOSE and literacy: connection of textual allusions with historical details (Question 14);
- Construction of attention-getting text fulfilling a given purpose (Question 15);

Part two was called "An Australian traveller in Egypt". There were six questions requiring short answers. The first three questions required the student to match the numerals used in Egypt with those used in Australia. The other three questions related to the cost of a souvenir with a price tag in Egyptian pounds. A combination of photographs and illustrations were used to present the numerals.

The intents of this part of the constructed response were to assess:

- Numeracy: visual discrimination (Questions 16 and 17);
- Numeracy: performing a calculation in context (Question 18);
- Numeracy: working with numerical symbols (Question 19);
- Numeracy: working with numerical symbols and conversion of currencies (Question 20);
- Numeracy: conversion of currencies (Question 21);

The third part was called “American Gothic-not!” and was comprised of eight short answer questions. The first five questions were about the parodies of the painting American Gothic. There were three more questions about the meaning of the painting American Gothic, the pitchfork motif and a painting called Australian Gothic. Photographs of the painting American Gothic and parodies of this painting were included. The painting Australian Gothic was also incorporated.

The intents of this part were to assess:

- Recognition of significant differences in visual images (Question 22);
- Recognition and explanation of what a single component of visual text contributes to its overall meaning (Question 23, 24 and 27);
- Inference of purpose of visual text (Question 25);
- Justification on the basis of evidence (Questions 26, 25 and 26);
- Identification of pictorial symbols (Question 28);
- Understanding of how visual texts, individually and together, comment on significant socio-cultural issues (Question 29).

#### *Teacher Generated Task 2004-2005*

It was always intended that the QAT would comprise the three tasks outlined above and would be further complemented by a corresponding teacher generated task. However, the funding did not permit further development and full implementation of the TGTs.

As explained by a team member the teacher generated task process involved the development of three examples of tasks for Years 3, 6 and 9 drawing on the disciplines of SOSE and the Arts and exemplifying the construct of transformation. Teacher conferences were held which aimed to assist teachers develop tasks by providing and explaining to them the structure and process. Once teachers had attended these conferences they were asked to plan and write-up the potential task using the given outline and processes. Once the TGTs had been constructed teachers sent them to the QAT TGT team. An accreditation process was implemented and during the first two terms of 2005 the TGTs were assessed and sent back to the teacher with feedback for improvement.

The TGT took the form of a group task, with the focus on SOSE and the Arts and the transformation of ideas and information from one form to another. In this particular teacher’s case, students as a team had to develop a kit comprising their own letter, diary entry, speech and board game. The students were given the two scenarios of organising an ANZAC ceremony or considering how to avoid terrorism. Using the basis of tasks completed previously the students developed the kit. This teacher who was interviewed at a school where observations were conducted, stated that the:

“feedback from the TGT team consisted of comments like, ‘Feel it was too much busy work’, ‘Applied the item (specifications)’, ‘Review of the

student outcomes', 'Not enough challenge', and yet only minor changes for improvement were recommended by the TGT QAT team."

Once the TGTs were accredited the teachers implemented them and they were marked on site. School-based moderation occurred and in this particular case a minimum of six assignments were sent to another teacher for assessment. Teachers were recommended to consider a sample from the higher, middle and lower ends of the continuum of acceptable standards for student achievement.

Due to the decision taken to focus on the constructed response and the computer-based task the TGTs and the emergent grades and standards were not analysed and do not form part of this evaluation.



## TECHNICAL CONSIDERATIONS

In this chapter the technical dimensions of the validity and reliability of the 2005 QAT will be evaluated but first it is important to define these terms. Validity is the extent to which an assessment, test or examination does what it was designed to do, or measures what it claims to measure. Reliability refers to the consistency of measurement. To be meaningful measurement must be replicable, comparable and consistent.

### Key Determinants of Validity

All assessments are based on a sample of behaviour or performance in which we are interested and it is from this sample that we generalise to ‘the universe of that behaviour’. The ‘fidelity of the inference drawn from the responses to the assessment is what is called the validity of the assessment’. This is why the specification of the domain of behaviour in which we are interested is critically important. “... validation is a difficult process ... and is impossible if we do not have a clear idea of what it is we are trying to assess” (Nuttall, 1987: 110-111).

The 2005 QAT drew on SOSE and the Arts and specified the construct of transformation as the focus for the assessment tasks. As a QAT team member explained:

“I was involved with what we might make the subject of the assessment and that’s where the idea of transformations came in. I categorised transformations and developed an actual typology of those so that when we looked at assessing the students’ performance in transformation we could do it across a different range of transformations, and I also had a role with coming up with a construct that said what we were assessing were transformations but also the underlying generic skills that are involved in doing transformations.”

To demonstrate the validity of the inferences marking guides were provided for the teachers who were assessing the tasks. These guides detailed the intent of each of the questions of the constructed-response and the computer-based tasks. Teachers were also provided with some student responses that had been graded by the marker advisors to help them gain an understanding of how the inferences were made. That is, teachers were able to appreciate how the results support the inferences that the marker advisors had drawn from the student responses.

The QAT developers indicated that they critically scrutinised and analysed the stimulus materials, carefully considering the implications of each question asked. In focusing on the construct transforming and contextualising the transformation skills in SOSE and the Arts the developers had to map back to the syllabus. This is how the choice was made to use contour maps in the

design of the computer-based task which a team member explained would allow the assessment of the task to be multi-modal.

“We wanted to exploit certain modes of assessment at the same time. I traded quite a few possible tasks that they (the developers) had to select from but I actually ended up doing a whole lot of revisions of similar tasks that had come from other people ... I came up with the idea of contour maps ... there is a sort of interchanging allocation of people to duties ... the development process is not linear it is multi-phased and multi-faceted...”

In further explaining how validity was addressed a team member commented:

“We had plans for external panels of the experts and classroom teachers but in the end we didn’t have the time or the resources because of all the things that happened, so we had to do away with those plans. We had experts who were part of the staff, who I used as a surrogate panel and also because we were doing transformation we really didn’t have to have experts in the disciplines of SOSE and the Arts. The contour maps are part of the content of the syllabus of SOSE. Whether it was being taught in schools raises the issue of knowledge. It is not consistent across schools because of the way that teachers interpret the syllabuses and the way they are implemented. So the validity of content was an almost impossible task because of the situation in Queensland ... we weren’t testing content, it was context that was important ...”

The tasks and conditions that reveal a student’s best performance include:

- Tasks that are concrete and within the experience of the student;
- Tasks that are presented clearly;
- Tasks that are perceived as relevant to the current concerns of the student and
- Conditions that are not unduly threatening.

Improving the tasks and conditions that bring out the best in our students will not automatically improve validity (ibid: 115-116). This is achieved by improving the sampling of the tasks and providing a range of contexts for student performance. It needs to be remembered that the 2005 QAT had been conceptualised as comprising not just the constructed response and the computer-based tasks but also the performance task, all of which were to be complemented by the TGT. By increasing the range of contexts and providing more extensive sampling of the skills in this way the selected constructs were more likely to be assessed thereby enhancing validity.

However, in 2005 it was decided to focus on a futuristic QAT that would comprise the computer-based task and the constructed response. Given the shortage of funds and limited capacity to develop, administer and mark all the intended tasks the quality and appropriateness of the QAT for the assessment of the selected construct appears to have taken precedence over the total

quantity of assessment; an important principle for maximising validity (Crooks, 2001).

The nature and format of the assessment tasks should align with the learning intentions that can lead “to substantial variety in tasks, with benefits in versatility of approach and development of transfer skills” (Crooks, 1988: 470). The tasks involved in the 2005 QAT provided students with considerable diversity. All the questions of the constructed response and computer-based tasks were designed to assess transformation skills. As a team member suggested:

“Well contour lines have implicit transforming of information. You have to work out where you are from the contour lines ... that sort of information. That includes the cross section task, which is a major transforming exercise. It maximises the capacity of the computer and encourages the kids to make a decision about which points will most correctly represent a cross section.”

Any assessment instrument does not possess validity in isolation but rather the validity depends on the way in which the result is interpreted. Messick’s (1989: 13) notion of validity is apt here:

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.

Validity in this case refers to the evidence available for assessment interpretation and the potential consequences of assessment use. For the 2005 QAT the potential consequences of the use of the assessment results are uncertain and therefore it is difficult to determine if these will be used in a valid way. Each student received a statement of results on the construct of transforming ideas and information. On the certificate sent to students the following description was given:

Common, standardised assessment of transforming ideas and information employing underlying generic skills and dispositions in the context of Study of Society & the Environment (SOSE) and the Arts. (QSA, 2005)

The students received a grade for the computer-based task and the paper-based task with an overall grade. Grades A-E were used, however, only grade A was defined on the statement of results.

Collectively, A-grade students:  
Exhibit knowledge of key aspects of history, geography and media.

Extract information from prose, diagrams, maps and symbolic text;  
clarify it and transform it to display meaning in multiple media.

Discern patterns and relationships in verbal, pictorial and symbolic text (alone or in combination); make significant decisions and judgments, operationalise these into accurate representations and products. (QSA, 2005)

Further information provided in graphical format in the statement of results was the percentage of grades awarded based on all task items (overall), computer-based items and paper-based items.

This use of these grades is an important factor for consideration in terms of validity. Students have completed the tasks and while their performances may be a reflection of their skills on the day of completion, the results may not be valid for the purposes used. It is not clear how the results will be used and interpreted at the classroom level and by the teacher, student or parent/carer.

### **How valid was the 2005 QAT?**

There was a balanced range of tasks incorporated in both the computer-based and constructed response to enable the students to demonstrate their transformation skills in the four types identified for assessment. This important determinant of validity was met. In relation to the computer-based task a member of the team commented:

“... a wide range [of tasks] that are demanded of the student ... Skills involved estimation, calculations, contour calculations, creative writing, creative imaging, numeracy, evaluation, decision-making. These are some of the skills of transforming.”

For the constructed response the following marking guide of the second unit illustrates how transformation types were identified in questions 16-21 and how they together with the intent of each question were communicated to the markers.

*Intents and desirable features*

<i>Q#</i>	<i>Task Identifier</i>	<i>Trans. Type?</i>	<i>Intent(s): to access</i>	<i>Desirable features</i>
16	Similar numerals	Underlying skill: compare	Numeracy: visual discrimination	The Egyptian numerals identified are ١ and ٩
17	Confusing numerals	Underlying skill: compare	Numeracy: visual discrimination	Numerals that are sources of potential misinterpretation of the intended number are identified and explained.
18	Distance between signs	I: IF a double conversion is made to perform the subtraction	Numeracy: performing a calculation in context	The correct difference in the shown distances is given using Egyptian numerals.
19	Ticket conversion: Australian numerals	I: 'script' replaces 'genre'	Numeracy: working with numerical symbols	The price is given correctly in Australian numerals.
20	Ticket conversion: Australian price	I: 'currency' replaces 'genre'	Numeracy: working with numerical symbols and conversion of currencies	The correct price is given in Australian dollars, to at least the required order of accuracy.
21	Money conversion: Oz to Egypt	I: 'currency' replaces 'genre'	Numeracy: conversion of currencies	The conversion is given in Egyptian pounds, to the closest pound.

For the purposes of planning and assessment, outcome levels typically relate to year levels. Usually then students at the end of Year/Grade 9 would be demonstrating Level 5 outcomes. The stimulus material was generally accessible although there were a number of teachers, markers and marker advisors who commented that the constructed response was pitched at a higher level than Level 5. To illustrate, the first comment below indicates that the teacher thought the tasks were intellectually challenging, however, questioned the level at which these challenges were pitched:

“Any kid who could have managed that test (constructed-response) in Grade 9 would be well set to go on to senior studies. I have only marked a small representation of the sample but that small sample has revealed some glaring gaps in the expectation of what’s happening in Grade 9. But this is a very small representation of the sample and that may be the papers that I have seen so far.”

“... not all questions were aimed at Level 5 some were way too hard for Year 9 students ...”

“... maybe pitched at higher level than 5”

“... the nature of the items and in particular the language used is too difficult for the average Year 9 student ...” and “... conceptually it was difficult ...”

The responses in both the constructed response and the computer-based tasks incorporated a variety of modes (such as diagrammatic, graphical, pictorial and symbolic). Some markers thought that the “number of bits” in the

constructed response such as the multiple choice questions, short answers, completion of a table, identifying pictorial symbols, was expecting too much of Year 9 students. The setting out of some questions also caused problems.

“ I think the questions are ... (pause) even the three markers had difficulty with a question where we thought we had to pair the words in the table. However the marker advisor indicated that this was not the case. So there are some problems with the wording of that task.”

“The language ... I think is also ambiguous at times. For example, with the photos of the painting depicting the pitchfork and one without the pitchfork, the caption reads that the picture (view 1) celebrates American history and the one without (view 2) condemns it, and the way this is written out it isn't stated at the top. ... it then asks the students to state how the picture supports view one and view two and the kids have got confused because the two pictures are there ... had they restructured the way those questions were put together because the kids tend not to go back to the exam (question) to look. They just focused on the pitchfork and missed the point of the question by the time they got to the end of it. This will be highlighted by the poor responses.”

Some teacher markers also commented that they thought some students might not have taken the constructed response task seriously. This determinant of validity has much to do with the way in which the task was administered. Observations were conducted in two schools, one an inner city, independent school and the other a country, state school. It was observed that staff followed the administration guidelines, making sure that the task was undertaken according to the same set of conditions or task parameters. Students in these schools were given the same instructions and were given assistance according to the guidelines. This may not have been the case in all schools as one teacher suggested:

“I don't know how it was presented to schools but from our school's perspective the HOD said we have to do this, ... she said I know you have a computer room. Can I use that room? It was all the stress of trying to get that done. For a lot of people they thought that the Education Department has given us another thing to do, let's get this thing over and done with. Some of the teachers were cranky about losing their class time ... I don't know how seriously it was presented to the kids but the HOD in charge of it is very professional. ... when she wasn't there I don't know how seriously it was taken. From some of the answers it doesn't appear that some of the students did take it seriously.”

This view in part was associated with the choice of stimulus material. For example:

“ I don't know whether it was the stimulus material in the constructed response, or why they used ... American Gothic, because it is Year 9,

and there is no explanation. It is beyond their experience. I don't think the students understand the relevance to the tasks."

"As soon as kids see the word Gothic they have a particular association with the word and they have said this isn't a Goth and because this has happened they haven't really read what the question expected of them. That's limited them."

"... students have a stereotypical view of Gothic ..."

### **Quality of Student Data**

As explained in the report, *Queensland Assessment Task Reflections 2005*, grading was defined as (EQ, 2005: 68):

"... the process whereby a candidate's 'raw' score is translated into a letter grade. A student's individual result, expressed as one of five letter grades (from A-highest, to E-lowest), is determined using a process which reflects the distinct nature of the QAT. The method used to determine the student's grades incorporates traditional scoring techniques together with appropriate aspects of criteria-based assessment as practised in Queensland.

Each student's total score on the QAT is determined by combining their individual results on questions, having first established the relative worth of each question."

Multiple-choice questions were marked according to a key, with each response coded as 1 or 0. All other questions on the QAT had contributory grades (A to D) and these were assigned a numeric value, the non-contributory grades (N and O) had a coded value of 0.

The relative worth of the grades for a question had two dimensions. The first was the maximum possible which was the numeric value assigned to an A-grade. For one question, coding of the A-grade took into account its relative worth in relation to all other questions on the QAT.

The second dimension was the numeric gaps between pairs of contributing grades on a question. After the coded value of the A-grade was determined, numeric values were assigned to the remaining grades of the question based on the extent of the performance in the particular domain.

Decisions about the cut-offs were informed by the desirable features of an A-standard performance and the calibration of the set of questions in the QAT. The calibration provided information about the sort of skills high scoring students have over lower scoring students.

The range of grades achieved both on the computer-based task and the constructed response suggest that the tasks were of varying difficulty so that the students were given an opportunity to demonstrate the extent of their

achievements and markers were able to discriminate between achievements of different quality. For the computer-based task a team member explained how he was involved in the validation of the algorithms used to mark the students' responses.

“Making sure that the algorithms are logical and that the raw results I get from processing translate into the correct grades from the algorithms. ... there are two stages: student data to the raw data [processed data in the context of the algorithm]. For example, take the timeline question the raw data would be the number of events in the proper sequence, which is one of the indexes we used. That's a basic one. And then there is data needed for the relative positioning of events. That provided a lot of results. Once you have those raw results you can determine the cut off points within those raw results and then the student can be given a grade.

There is a whole range of results from A to O. Some of the questions it seems that the students still didn't get the concept and some it appeared too easy. That's not a quantitative measure it is my judgement from what I have seen. Definitely a wide range that are demanded of the student in the computer-based task.” (Interviewee, 11)

The overall results achieved by students for the 2005 QAT are presented in Table 2.

Table 2 Overall 2005 QAT results (EQ, 2005: 66)

	A	B	C	D	E	TOTAL
Number of female students	210	512	933	431	22	2108
% of female students	10.0	24.3	44.3	20.4	1.0	100.0
% female of total students	5.0	12.1	22.0	10.2	0.5	49.7
Number of male students	151	391	866	493	25	1926
% of male students	7.8	20.3	45.0	25.6	1.3	100.0
% male of total students	3.6	9.2	20.4	11.6	0.6	45.4
Number of students – gender unrecorded	17	54	94	37	2	204
% of students – gender unrecorded	8.3	26.5	46.1	18.1	1.0	100.0
% unrecorded of total students	0.4	1.3	2.2	0.9	0.0	4.8
Total number of students	378	957	1893	961	49	4238
% of total students	8.9	22.6	44.7	22.7	1.2	100.0

### Ensuring Reliability

The essential reliability question is: would the assessment of the student's response result in the same or similar assessment on two occasions if assessed by two assessors? In evaluating the reliability of the 2005 QAT the consistency of approach to the assessment task, as well as the consistency of standards of assessment, have been evaluated.



Forster & Masters (1996:43) have identified the following major methods for ensuring reliability:

- Documented, field tested marking guides;
- Specified criteria;
- Annotated examples of all score points;
- Ample practice and feedback for raters;
- Multiple raters with agreement prior to marking;
- Periodic reliability checks throughout;
- Retraining if necessary and
- Arrangements for the collection of suitable reliability data.

These methods of ensuring reliability have been evaluated for the 2005 QAT.

## Reliability

The reliability of the QAT focused on the consistency in measuring achievement in the generic skills associated with transforming ideas and information in the context of SOSE and The Arts along with aspects of literacy and numeracy.

The QAT was administered on one occasion, so the important fact of reliability pertinent to us is how reliable are the results on the test so that they can be generalised to the point that the students' results would be consistent if the same design criteria were used and other items were on the test.

The form of reliability that is reported here is internal consistency. The consistency of performance on each test item is considered. The measure used is called "Cronbach's Alpha".

Coefficient alpha,  $r_{\alpha}$  is given by:

$$r_{\alpha} = \left( \frac{N}{N - 1} \right) \left( 1 - \frac{\sum \sigma_i^2}{\sigma^2} \right)$$

where

$N$  is the number of items

$\sigma_i$  is the variance of the other  $i$ th items

$\sum \sigma_i^2$  is the sum of the variances over all items

$\sigma^2$  is the variance of the total scores.

Cronbach's Alpha for the 2005 QAT is 0.85.

## Marking Guides

The QAT team responsible for the development of the marking guides, and field testing them, was experienced in the marking operation. It was a considered decision to recruit this team.

“You could only do this because of the nature of the team. The members were highly experienced and had a wealth of knowledge, background, expertise that made this team effort possible. ... this team was able to complete this task .... ”

Where possible the team members drafted the marking guides at the same time as they developed the tasks so that the guides could feedback into the task development. In this inter-relationship, of task and marking guide development, the team borrowed from their experiences of the 2003 QAT and their experiences of assisting with the Queensland Core Skills (QCS) development of marking guides.

The marking schemes for the constructed response were developed with the understanding that some of the short answer responses needed rewarding of higher order skills. Consequently an innovative format was designed to match this identified need. This required a more sophisticated approach that enabled the assessment of the skills underlying the construct of transformation. A QAT team member explained:

“I have created those new types of marking guides which even though different ...[were] motivated by work I did on the grading master for the rich tasks ... the ones that had the graphical component ... was an important development to ... overcome some difficulties ... in the sense that they are not unidimensional ...[I was] trying to cope with the fact that you’ve got this interplay of different skills and you want to be able to reward that interplay and not just treat it with (pause) the answers are correct or half correct. So I designed the format. ”

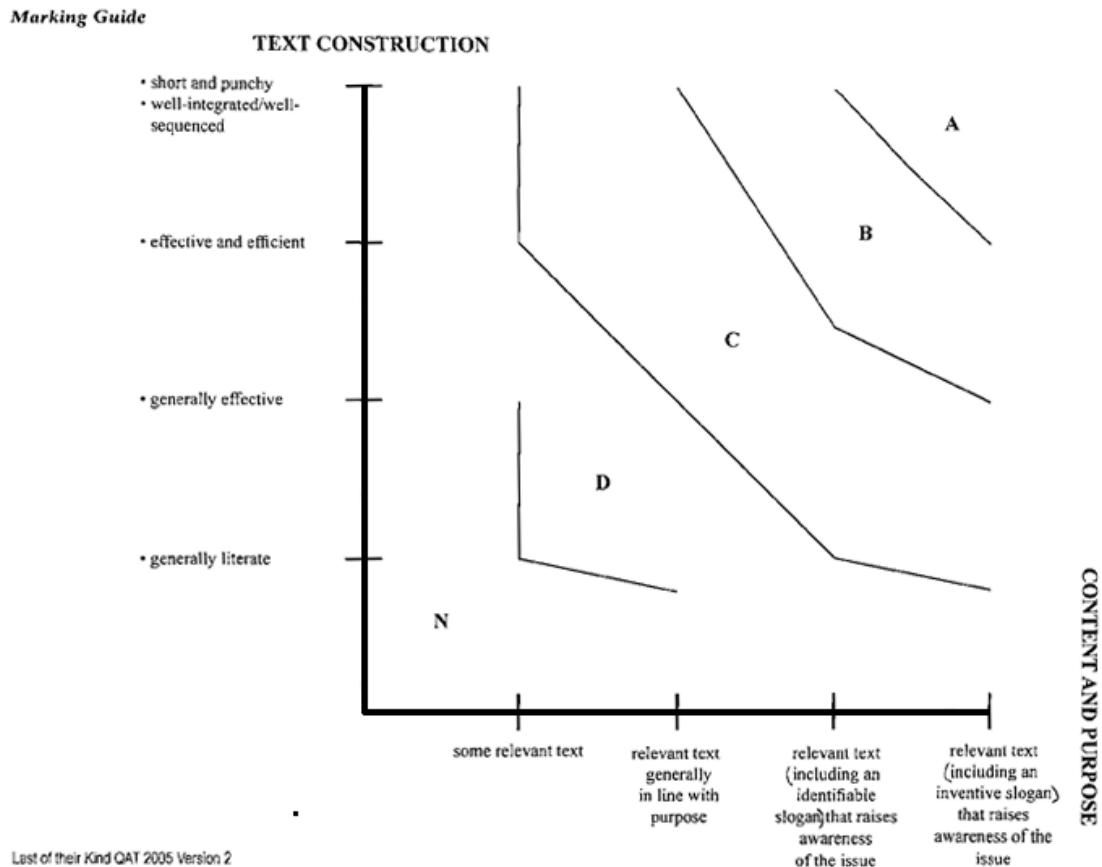
There followed a continual revision process both before and after the administration of the tasks to ensure that they were markable and to evaluate whether the guides assessed what they set out to assess. Criteria were identified, specified and refined and the marking guides were field tested.

An inhibiting factor in the development of the marking guides was again a lack of funds to trial the items satisfactorily and to evaluate how well they fulfilled their function. To illustrate:

“If you follow true process when you develop a unit and put it into the possible pile you should develop the marking guides at the same time and if we had ‘ve trialled you would have known if the marking guide worked ... we didn’t have the capacity to follow all those procedures. But before it went to print we had to have those marking guides developed so that we knew at least it was markable. ...That was when

... developed the graph for marking the multi-dimensions of the open-ended questions. ”

The following example of this graph illustrates the dimensions of text construction and understanding of content and purpose for responses to Question 15 of the constructed response. Students were asked to complete a poster with an attention getting text to fulfil a given purpose.

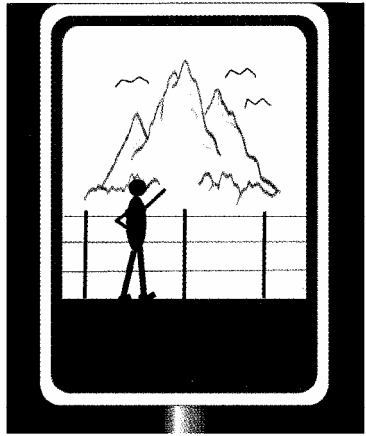


Once the tasks had been administered, although most guides did prove effective, there was still a need for some further refinement in terms of the language and to ensure there was alignment with tasks that had been completed. This was intended to facilitate the use of the guides for the markers by making them more recognisable and illustrating how manifestations of the standards were captured. The standards, themselves, did not change during this period of refinement.

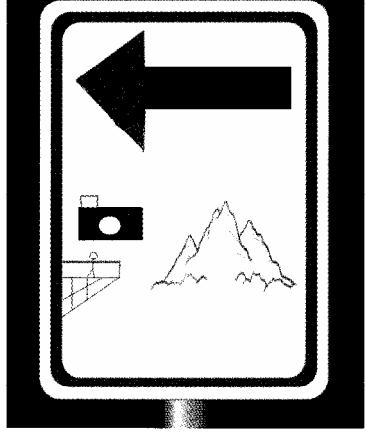
It was also at this stage that annotated examples of the standards were written up by the marker advisors using the students' actual responses. To illustrate, two examples are given here to show what the teachers who marked the computer-based task (Question 4) remotely used. The following

exemplars illustrate a B standard (Example 6) and an A standard (Example 7).

Example 6

 <p>SCENIC LOOKOUT TO THE LEFT</p>	<p>Sign shape matches message? Yes Was text included? No</p> <p><b>B</b></p>
	<p>This work captures attention, is visually effective, shows appreciation of the context and shows effective use of the paint program. It clearly shows a scenic lookout. What it doesn't show is that the lookout is to the left! So while a driver may grasp the thrust of the message about a lookout, s/he won't know which way to turn off. Had an arrow to the left been included this would have been awarded an A-grade.</p>

Example 7

 <p>SCENIC LOOKOUT TO THE LEFT</p>	<p>Sign shape matches message? Yes Was text included? No</p> <p><b>A</b></p>
	<p>This sign conveys not only that there is scenery (in this case represented by mountains) to the left, but also that there is a lookout, shown by the observation deck and reinforced by the camera. The sign captures attention, is visually effective and is probably obvious to a typical driver. The student has made effective use of the paint program.</p>

## Standards

The development of standards is a task that cannot be completed cheaply. The markers should not have a role in the development of standards for standardised tasks. Teachers complete the units and surmise how the students will respond. The task developers then analyse these and determine the standards. The draft marking schemes are used to mark student responses and if there are problems then they are redrafted until agreement is reached that the marking scheme is reliable. It is the test constructors who complete the marking scheme. It is important to understand that the standards emerge from the student work not from what the markers think the standards should be.

When teachers were asked to describe their role in determining the standards used in the marking of the QAT, they commented:

“... only in terms of clarifying ... with the marker advisors. It's been pretty much directed to what we do. ... this is what you have to mark and that's it. I don't think the teacher's input is expected. The limitations and the boundaries are made clear.”

### **The Marker Advisors and Markers**

The three marker advisors for the 2005 QAT were observed fulfilling an important role during the marking operation, which took place from October 10 to 14, 2005 at a central location in Brisbane. They were described as an 'expert team' that had been carefully selected because of their experience in marking and also their expertise in training markers. Some marker advisors were described as 'master teachers'. These qualities enabled the project co-ordinator to trust the marker advisors to get on with the job without supervision.

The marking was described as an operation because it is not a straight, forward linear process; it is multi-faceted. The marker advisors are responsible for giving information about marking guides, criteria and standards used. They also monitor the entire operation.

Markers received a marker manual that included information about the venue, administration and marking procedures and the marking guidelines. They were informed that the QAT was "...a new approach to standardised assessment incorporating ... sophisticated approaches to task based assessment ... including a computer-based component employing the latest technology." (QAT, 2005: 1)

### **Marker Training Programme**

Observations and interviews were conducted during the week that the marking took place. The marker training programme for the constructed response consisted of a full day of input, practice and initial monitoring and evaluation of the markers' consistency in judgement by the QAT team, manager and the three marker advisors. There were three groups of teacher markers, including the 'veterans' who had a wealth of experience gained from marking QCS papers in previous years. Each marker advisor took responsibility for training a group of markers in a particular marking unit. These were colour coded as follows yellow Parts 1 & 2 and purple Part 3. This coding was important for the mechanics of the marking procedures.

To begin the manager explained to all markers that the domain of the QAT was 'processing and in particular transforming ideas and information in the context of SOSE and the Arts'. The manager explained that the grades would be allocated for transforming, and information gathered on literacy and numeracy. The rationale for the 2005 QAT and the reasons for focusing only

on the computer-based and constructed response tasks were given. The computer-based task consisted of seven questions: four of which were computer marked algorithmically with questions 4, 6 and 7 marked by teacher markers remotely.

The training session which was observed by the evaluator was for the constructed response, a pen and paper task. The marking model was outlined with the following emphasised:

- markers would read a variety of responses
- each response would require the marker to assign one of a limited set of meaningful grades and
- no impression marking to occur.

It was explained that the marking guides were developed before the constructed response paper went to the printers and revised after students' responses were reviewed. The point was strongly made that all markers in each unit should have the same understanding and interpretation of the marking guides. The mechanics of the marking procedures were explained and the markers were told that their marking would be monitored by close scrutiny of the marks that they were awarding. This role would be carried out by the QAT team. Once a marker was identified as being out of step with the others there would be a need for that marker to refocus. This monitoring process would help to ensure that there was consistency of judgement.

To further ensure reliability marker advisors and those experienced in the marking operation indicated that it was crucial that the markers 'accept the marking scheme and work with this'. Markers need to understand that the marker advisors provide 'the single source for the explanation of the grades' and facilitate the shared understanding of the standards used. They also need to appreciate that their role in the operation requires them to be consistent and reliable.

Throughout the operation it is possible for markers to become vulnerable due to exhaustion or stress. Marker advisors, together with other team members, monitored each marker's consistency and comparability in assessing the tasks. This was achieved by comparing the individual marker's performance with that of the group to identify whether an individual was out of synchronisation. This was carried out at the beginning of the operation because if inconsistencies are not identified early it could cause some instability to the whole operation. Often such inconsistency of judgment is due to a misconception or a lack of understanding with the particular marker 'contaminating' the results. At this point the marker advisor advises the individual about her or his decision-making. It might also require some retraining of the particular marker if revealed by the marker monitoring data. To rectify this situation it might require a third marker to mark the tasks again. Generally such referred marking is used when noteworthy differences occur between different markers of the same tasks.

## The Management of the Marking Operation

The use of an expert team, who had been carefully selected, was crucial for the 2005 QAT pilot. Given the circumstances for the QAT team it was not possible to train new staff or to implement the intended professional development model. As explained by a team member

“Originally it was going to be an experienced person teamed with an inexperienced person and a teacher ... the ideal model ... didn't have the time, money or capacity ... In July when we were taken over by QSA it was about getting the job done and getting the job done well. And it wouldn't have been done any better if ... the other model [was] in place but the other model *is* about teacher professional development of other people and training for the future.”

As a consequence the marker advisors trained the teachers to do a very specific, complex, multi-layered task. Teachers had to accept this singular role of marker in completing this task. The marker advisor had to be sure that the markers understood the information given.

Marker advisors were there to monitor the teachers' progress to ensure that what was being asked of them was within their expertise. The marker advisors were confident with the procedures involved in marking operations and were alert to such processes as randomised movement of the scripts between markers from the outset.

The marker advisors decided how to manage their group of markers. For the constructed response, there were units and each marker advisor took the responsibility for advising their particular group of makers. One marker advisor organised the markers into two groups:

“... no one can capture all the dimensions to the task and keep focused at the one time. There was a need to break down the marking task into two groups. ... you look at the marking decisions that need to be made, assessing all the intricacies that the markers have to keep in mind. You have half the group doing each part of the unit. One of the two groups was moving faster than the other because the decisions were easier in this group. It doesn't matter about the quantity of the marking it is the qualitative difference in the decisions that need to be made. So ... some of the people in the group that had the easier decisions [were] retrain[ed] to help the other group so that you keep the pace the same. It's important to have the decisions being made at the same time otherwise we get into trouble because it costs money if people are left waiting...”

The marker advisors make these management decisions. Later during the marking operation the opportunity did arise for those markers who were consistent to be retrained to complete the other half of the unit. This strategy helped to alleviate the boredom factor and refocused the markers so that they

were refreshed, it was also a recognition and appreciation of their consistency and reliability.

As the marking nears the end there can be times when other markers are left waiting for others to finish. The maximum wait time for any of the markers involved in this operation was forty-five minutes. Markers are requested to wait until the operation is finished because there may be some scripts that need to be remarked. The team completed the operation one hour after the estimated time.

Some factors hindered the smooth operation of the marking operation and these were largely due to the lack of resources. For example, a scanner was not available on site so some of the team worked back at the office while the manager was at the marking venue. Consequently some new staff did not realize the detrimental consequences of not processing the data quickly.

The marker advisors who did not split their groups into two at the outset, in retrospect regretted their decisions, and after two days it was decided to split the second unit into two. Another reason for splitting the groups was because this was comparatively a small marking operation involving 4646 scripts. In total there were 41 markers, making the actual percentage that one marker is responsible for quite large in terms of the potential consequences for actual error to occur. Splitting the group helped to reduce this risk.

## **Reporting**

How the results were to be reported was determined during the period of data collection for this evaluation. Consequently, teachers and markers were asked about their perspectives regarding how they thought the reporting should occur. Some teachers thought that the results should not be reported student by student but rather that teachers should be provided with student data that indicates the range and the mean of their particular students' results. There was the fear that if individual students' results were reported it could lead to early labelling whereas if the reporting was done class by class it could provide the teacher with important insights.

Some thought that the students would get the results question by question not an overall result.

I think it needs to be reported carefully if it is aligned with the QCS because if the students haven't taken it seriously and they have done poorly and you compare them to other people in the state it could leave a negative impression about what they have achieved or haven't achieved ... teacher[s] ... don't trust the statistics because ... can be manipulated.

## **Equity Issues**

Equity issues appear to be an important determinant of validity that was not implemented due to the lack of resources and time.



“Special needs was planned for but we did not have the capacity to fully implement this. Schools applied normal assessment procedures and so there could have been some special needs students included. But there wasn’t special consideration, this should happen before or while they are doing the task not in the grading after ... unless of course a student was sick [on the day]. ”

So although there was some feeling amongst the teachers and markers that special needs and students requiring special consideration were given help:  
“... some answers may have been scribed. ”

The QAT developers indicated that they were unable to implement their plans due the lack of finance.

In addition it was reported in the *QAT Report on the Computer Based Task* (Owens, 2005) that:

“Unfortunately a decision was taken to drop any support for special needs before the first trials took place. No further testing was done and much of the instruction and help system were implemented in a manner that did not make it able to use the TextHelp software without modification.” (Owens, 2005:15)

## **Conclusion**

To illustrate the extent to which the QAT met the design brief the student results have been examined in term of validity and reliability aspects of the achievement data that have been extracted from the students’ results.

Overall the QAT appears to have provided intellectual challenge and made connections to the wide world from the perspectives of the teachers, markers, marker advisors and the QAT team.

It has been possible to assess the student’s achievements in transforming ideas and/or information and in the underlying generic skills and dispositions, this is evident from the students overall results and the way in which these have been reported.

The QAT has provided assessment data on processing (transforming ideas and/or information) and in so doing was contextualized in SOSE and the Arts which required students to demonstrate understanding of knowledges (facts, concepts, procedures).

As reported (EQ, 2005: 65) the QAT did not provide measures of performance in literacy and numeracy as specified in the brief because:

“... literacy and numeracy skills were fundamental to, and interwoven in, the assessment tasks. Literacy and numeracy skills were required for students to access the interactive computer-based tasks and the constructed response tasks that made up the 2005 QAT. That is,

literacy and numeracy were, in fact, assumed knowledge for the 2005 QAT. Measures of literacy and numeracy were subsumed in the overarching measure of Transforming ideas and information. ... Grade C could arguably be said to identify a student who is functionally literate and numerate.”

## STUDENTS' PERSPECTIVES

This chapter presents the students' perspectives of both the constructed response (the paper-based task) and the computer-based task. It outlines the sample of students surveyed and interviewed, the approach adopted in analysing and triangulating the data sets and summarises the key findings associated with the students' perspectives of the 2005 Queensland Assessment Task (QAT).

### Student Sample

Immediately after the students had completed the 2005 Queensland Assessment Task (QAT), interviews were conducted with four focus groups of students (30 interviewees) at two schools (FG1 and FG2), one an inner city, independent girls school (ICGS) the other a country, state school (CS). In the former school the evaluator was only present for the computer-based task while in the country, state school the evaluator observed students completing both tasks. The focus group interviews included questions on both tasks.

In addition, in two different schools, one an inner city, independent boys school (ICBS) and the other a remote, state school (RS), students completed an open-ended survey (OS) which sought their views of the experience of both the constructed response (44 responses) and the computer-based task (183 responses). This occurred on the day of the administration of the tasks.

The survey was paper-based, to maximise students' access to it in all schools, including those with limited computer resources. It sought students' perspectives on their experience of the QAT. The survey was constructed using a five point likert scale (strongly agree, agree, undecided, disagree and strongly disagree) and required students to indicate their response by shading in the appropriate 'oval' on the form. There were also some short answer questions included. Over 4000 students in 56 schools undertook the 2005 QAT. Each of these schools was requested to have 20 students who undertook the QAT to complete a survey form; 794 (out of a notional 1120) completed forms were returned. This constitutes a response rate of 71%. The written components of the student responses amounted to over 48,000 words. Due to the employment of the evaluator at the time of the administration of the QAT seven weeks had lapsed by the time the survey was constructed, piloted, printed and ready for distribution. Many of the students' responses refer to this time difference, with implications that are discussed below.

These three data sets have collectively provided sufficient data that has been triangulated in the analysis to enable a rich and valid representation of students' opinions and experiences of the 2005 QAT.

### Data Sets

The student focus group interviews focused on:

- Opinions of both tasks
- Evaluation of the support available to complete the tasks

- The nature of the learning that was being assessed
- What was learnt from the tasks
- Suggestions for change
- Comparisons with teachers' assessment tasks
- Advantages and disadvantages of using the computer

The open-ended survey simply sought students' views on their experience of the 2005 QAT and encouraged them to express their opinions on both tasks.

The student survey consisted of the following parts:

- Computer-based task
  - Opinions of the task
  - Evaluation of the support available for the task
  - The skills used in completing the task
  - The most difficult part
  - The most interesting part
  - What was learnt from the task
  - Suggestions for improvement
- Paper-based task
  - Opinions of the task
  - The most difficult part
  - The most interesting part
  - What was learnt from the task
  - Suggestions for improvement
- Additional comments.

## **Analysis**

A mixed methods approach of quantitative and qualitative analyses has been used. The data from the focus group interviews, the open-ended survey and the written responses of the survey has been analysed according to the constant comparison method introduced by Glaser and Strauss (1967), expanded by Lincoln and Guba (1985) and Silverman (1993). This has involved the collection, organization and collation of the transcribed interview data and the other two sets (open-ended survey data and written responses from the survey) into three distinct data bases. Analysis involved several phases. First the evaluator read the entire data bank to identify key themes. The second phase required the evaluator to re-read the data bank and begin coding the data according to the identified similarities, differences, patterns and consistencies of meaning. The emergent themes from the first phase of analysis were coded during the second phase. The third phase involved the validation of these themes by triangulation

and further comparison with the findings from the survey data that had been analysed separately using percentage values for the five categories of response.

The findings of the analysed data sets are presented according to the two tasks and identified themes.

### **Computer-based task**

The survey questions asked the students to indicate the extent of their agreement with positively worded statements, in relation to their opinion of the computer-based task (Questions 1–9) and the support available for the task (Questions 10–18).

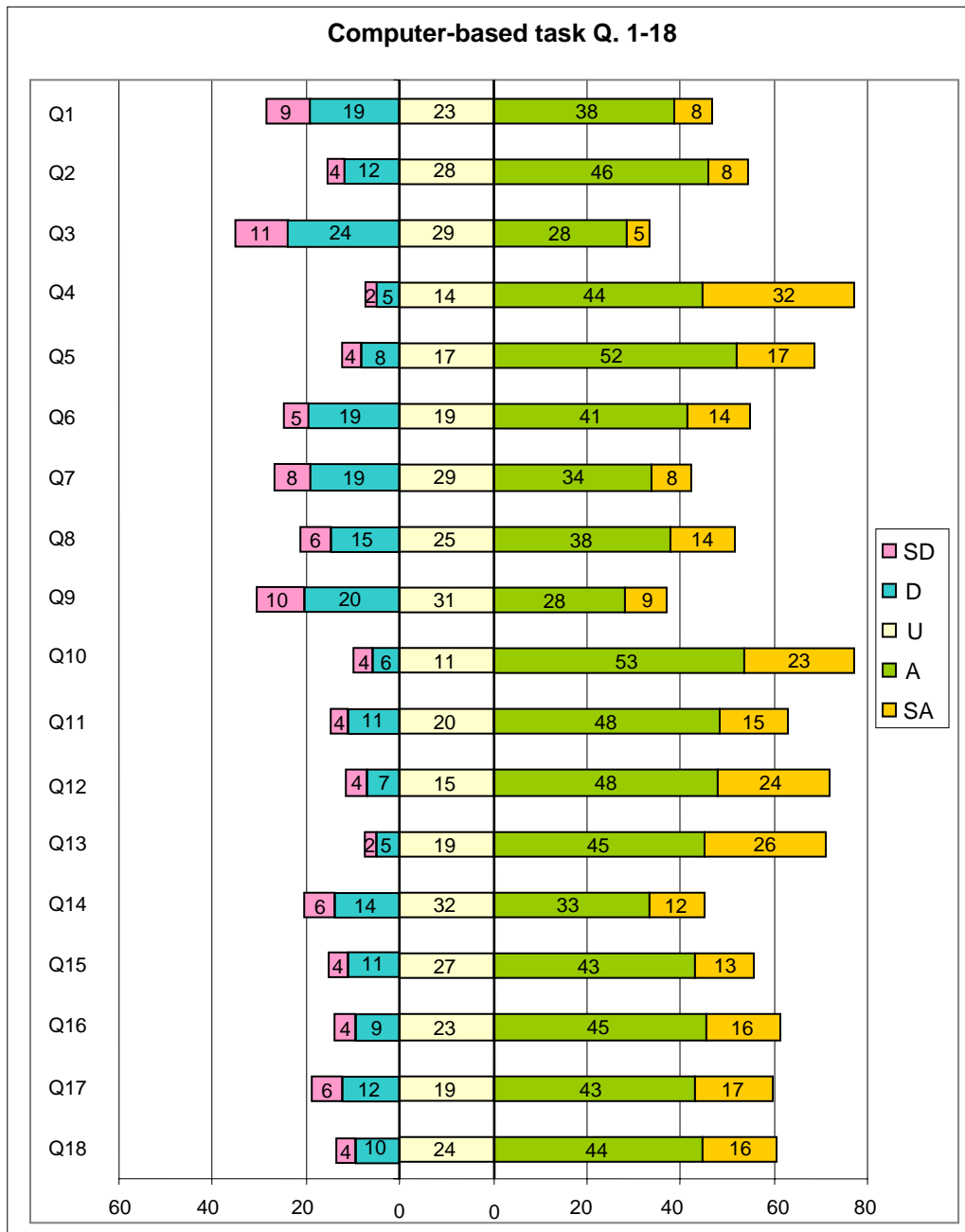
#### *Opinion*

1. The computer-based task was fun.
2. The questions ranged from easy to difficult.
3. The content material was interesting.
4. The computer-based task was very different from school assessment tasks.
5. The computer-based task required me to think about my answers.
6. The computer-based task required me to use my computer skills.
7. I was adequately prepared for the computer-based task.
8. The computer-based task was very different to what I expected.
9. The questions allowed me to show what I could do.

#### *Support*

10. The login instructions were clear.
11. The task instructions were clear.
12. There was adequate time to complete the tasks.
13. My computer skills were adequate to complete the tasks.
14. The lead-in activity on contours was helpful.
15. There was adequate computer-based information and help provided.
16. I found it easy to navigate around the question.
17. I made use of the help sections provided.
18. I had adequate content knowledge to complete the task.

Students' responses are presented in Figure 3. In this graph, responses of Undecided are represented by a bar of constant width, in order to emphasise the differences in actual agreement and disagreement. Percentage values for all categories of response (including Undecided) are given.



**Figure 3. Percentage student responses to Questions 1-18, Computer-based task**

### Differences from School Based Assessment

In the Opinion section of the survey (Questions 1–9), students tended to agree or strongly agree with the positively worded statements, indicating a general positive opinion of the task. This agreement is most marked for Question 4 (“The computer-based task was very different from school assessment tasks”), where the proportion of “Strongly agree” responses is the highest for any question in this section. Interestingly, there was greater agreement with this statement than with the statement in Question 8 (“The computer-based task was very different to what I expected”), perhaps suggesting that students can conceive of a greater range of assessment types than they are commonly presented with in class.

These results align with the analysis of the qualitative data sets. Students indicated not only that they agreed that the computer-based task differed to school assessment but suggested how it differed:

- Better level of explanation than other teacher directed tasks (CS FG1)
- Not much typing, that is written work (CS FG2)
- Much more variety, wasn't just black and white (CS FG1)
- It was the first time I was asked to text. (When this student was asked why she thought this question had been asked she responded with the following.) To see if we could explain (the view) using text messaging. (CS FG2)
- It was different because they guided you through the questions more than usual (ICBS OS)
- The nature of the questions was interesting because usually there is no activity to make it entertaining (ICBS OS)
- The test was a good change because it only has a few questions. They are not easy though but because there are on a few questions it gives us more time to work on the question and doesn't (sic) really cause people to freak out like a test with 100 questions do (sic). (ICBS OS)

### **Level of Challenge**

Agreement was also very strong for Question 5 ("The computer-based task required me to think about my answers"). An assessment task that is different from the usual and that requires students to think about their answers would seem to be in keeping with the vision of the QAT. However, a cautionary note is sounded by the responses to Questions 3 and 9, in which agreement and disagreement are fairly even; that is, students were evenly divided on the issues of the interest of the content material and the task allowing them to show what they can do. They agreed more that the task was fun (Question 1) than that the content material was interesting — the fun presumably coming more from the medium than from the content. The evenly divided responses on the extent to which the task allowed them to show what they can do can be taken as an indication that different mediums of assessment allow different students to show what they can do. If it was proposed that there should be a completely computer-based assessment regime, the responses to Question 9 would not support the position; they might, however, support the position that assessment should be sufficiently varied to allow the students who responded positively to Question 9, as well as those who responded negatively, the opportunity to show what they can do in an appropriate medium.

In general the survey data of students' opinions of the computer-based task can be characterised as positive to very positive; in no instances did they decisively *disagree* with a positive statement about the task.

Again the findings from the analysis of the qualitative data sets correspond to the findings of the student survey in that many students indicated that the task did require them to think about their answers and indicated that the task was fun, enjoyable and interesting. Question 5 of the computer-based task that required students to create a diagram of the cross-section using only 10

markers, was mentioned frequently as one that provided a great deal of challenge.

- I found the cross-sections challenging because it was bit hard to pick out where the markers went (ICBS OS)
- The test was harder then (sic) I thought. This test made me think. I enjoyed it. (ICBS OS)
- I found that most of the questions made you think a lot and weren't like normal test questions. (ICBS OS)
- It was a good test because it challaged (sic) my testing skills. I discovered that you can't study for this type of test. (ICBS OS)
- The questions were challenging and made me think laterally. (ICBS OS)

As the survey data suggested not all students found the task interesting. However there were very few comments such as the following that emerged from the qualitative data analysis.

- It was challenging but made me think about the question also a bit boring too much work on contour lines. (ICBS OS)
- It was hard (to be honest) and a bit boring. (ICBS OS)

There was also agreement in the qualitative data analysis that students believe that there should be a variety of different mediums of assessment to allow them to show what they can do. Suggestions included:

- Maybe more subjects such as maths (ICBS OS)
- Multiple choice questions needed (ICBS OS)
- I think the whole concept of computer-based tests is great and there should be more of them – education needs to adjust to the fact that technology is becoming more integrated into everyday life. (ICBS OS)
- The computer test didn't make me as stressed as paper tests. (ICBS OS)
- Pleased that tests aren't always on paper (ICBS OS)

### **Level of Support**

The analysis of the survey responses relating to the support available for the computer-based task (Questions 10–18) indicated that the students were even more inclined to be positive, with no more than 20 percent of students indicating any disagreement with any of the positive statements. The task's instructions, interface and provisions for online help seem to have been responded to favourably by the students. Noteworthy are the responses to Question 13: only 7 percent of students disagreed that their computer skills were adequate to complete the tasks. If students found the task difficult (as the written responses indicate that many of them did), this would not seem to be a reflection on the specific *computer skills* demanded by the task.

Similarly the analysis of the qualitative data suggests that students were very positive about the instructions and the level of support provided for the computer-based task, although there was infrequently a dissenting voice.



- The tests (sic) instructions were some times (sic) difficult to understand (ICBS OS)
- The instructions were helpful (CS FG2)
- There was plenty of explanation which made it easy (CS FG1)
- There was plenty of explanation and information and time to complete the task (ICGS FG2)
- The questions were layed (sic) out well (ICBS OS)

The medium of the computer appeared to be appreciated by the students, which is suggested from the analysis of the qualitative data.

- I found the 3D images useful and could see the benefits of the computer (ICGS FG2)
- Not much typing or written work (CS FG1)
- I liked the way with the contour lines that there were three ways that it could be communicated via the screen, by the shaded areas, the colour and 3D. (CS FG2)
- The reason why I liked it was because it was on a computer (RS OS)
- I was a convenient way to do a test because every question is at the click of a button. It is good doing the test on a computer because you don't have to worry about writing really. (RS OS)

### Perceived Skills Required

Survey questions 19 and 20 asked students about the skills they used in completing the computer-based task. Figure 4 presents their responses in relation to the nominated skills of deciding, designing and mapping.

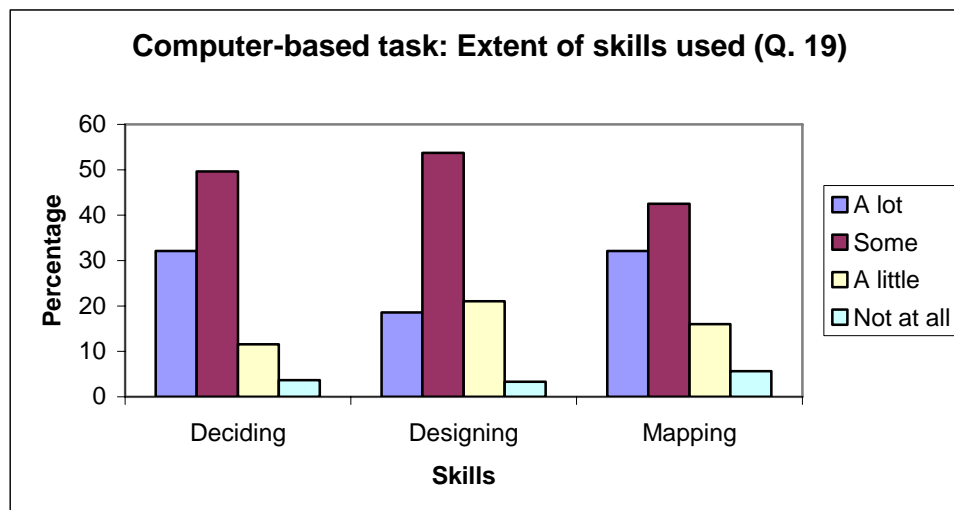


Figure 4. Extent of skills used in completing computer-based task

Most students indicated that they used each of the skills more than a little. What is perhaps most interesting about these responses is that mapping — the most superficially obvious skill required by a task dealing so strongly with contours — is not seen as dominating other, perhaps less obvious, certainly more *generic* skills (deciding and designing).

Survey question 20 asked students to nominate other skills they used in completing the computer-based task, and to indicate the extent to which they

used these. Nominated skills tended to be either very general (thinking, imagination, creativity, perception, understanding), or very specific (typing, keyboard skills, clicking and dragging). A quite frequent response, which perhaps conveys better than these responses how a proportion of students experienced the task, was “guessing” (only one student referred to “educated guessing”).

Noteworthy in responses to this question, as to other questions requiring written responses, is the rarity of what *may* be intentionally dismissive responses (sitting, watching the clock). Such responses may of course be genuinely descriptive of an experience and should not be overlooked. Nevertheless, their rarity might suggest that students did not, in general, approach the activity of completing the survey in a flippant frame of mind — thus providing more weight to the findings.

Likewise with the qualitative data analysis students indicated that the skills they used in the completion of the computer-based task included problem solving and computer skills. These quotations illustrate the range of suggested skills:

- I think the task was important for assessing our ability, competence with using the computer, and some generic skills. (ICGS FG1)
- They got ous (sic) to draw making it look at out (sic) artistic talents (ICBS OS)
- It was good how you use drawing skills as well. (ICBS OS)
- The test had a lot in it, like reading a map to drawing a sign in paint. (ICBS OS)
- Not having exact answers allows people to use their imagination and express their opinion. Gives us freedom. (ICBS OS)

Students recognised that their creative skills were required in the completion of question four (constructing a road sign) and question seven (writing a text message).

### **Level of Difficulty**

Question 21 asked students to nominate the most difficult part of the computer-based task. Every part of the task was nominated by at least a few students, but by far the most frequent response was Mapping or Contours, with quite a few students also nominating text messaging. Some students referred to specific contour-related questions (for example, “Placing icons on a selection of the road, using the slope indicator”). Some students explained *why* they found a particular part difficult, in ways that reinforce the idea that difficulty is not always an inherent, given feature of an item but rather is often something that lies in the relationship between the item and what the student brings to it:

- The contour lines as we had not been taught that at school.
- The contour lines, because I think that use of contour lines at school is stupid and tiring. I don't think the majority of us using contour lines in the future is very high.
- The drawing because I can't draw properly.

- The text messaging was the hardest because I DO NOT own a mobile phone. I felt the question was expecting people to have a mobile phone.
- The timeline was challenging as I had to visualise the times in my mind to decide how much distance to leave between them.
- The mapping as I don't really like seeing places on paper. I like to see them in a picture or in person.

If all parts of the task were mentioned in Question 21 as being the most difficult, so also were they mentioned in Question 22 as being the most interesting. But whereas the mapping dominated the responses in Question 21, designing signs and labels and writing text messages dominated the responses in Question 22. There are dissenting voices (nominating messaging as most difficult and contours as most interesting), but the most common view was that mapping was hard and designing and messaging were interesting.

From the analysis of the qualitative data similar conclusions are drawn. Again students indicated that mapping and contours proved quite difficult for similar reasons with some students indicating that text messaging was more difficult for them:

- I hadn't learnt about contour lines before (CS FG2)
- I found it difficult to transfer the picture to words. I only sent a text message. (CS FG1)
- The mapping skills were hard. We don't learn much about that stuff at school any more. (RS OS)
- Question 2 (placing icons on the matching feature) was difficult because it was completely different from any I've ever done. The map question also let me think hard on how the place will look if I was there. I was surprised it took so long to do seven questions. (ICBS OS)

## Learning

For Question 23 ("Did you learn anything in completing the computer-based task? What did you learn?"), the most frequent response was No or Nothing, but among the responses that did indicate that something had been learned, by far the most frequent response was Contour lines. Very few students indicated that they had learned anything from those parts that were most commonly considered to be the most interesting — designing and messaging. Some students indicated that it was specifically the computer medium that helped them to learn about contours, while a couple of students said that they learned that computers did not make any difference (for example, "Yes. Contour lines will always be boring, whether computer or on plain A4 paper.")

The learning that was suggested from the analysis of the qualitative data again corresponds with that of the findings of the survey data. Most students suggested that they learnt about contour lines. To illustrate:

- I found the closer together the contour lines were, the steeper it is. The further apart contour lines are the flatter it is. (ICBS OS)

There were a few others who made some differing suggestions:

- I learnt how different signs could make use of pictures to explain. (CS FG2)

### **Recommended Changes**

In response to being asked how the computer-based task could be improved (Survey Question 24), many students replied that the instructions should be clearer, or simply that it should be made more fun and more interesting. Many students, however, went further and suggested how this could be done, and the key feature of most of these suggestions was *variety*: variety in content (“not just contours”), and variety in computer tasks (“Powerpoint and Publisher”). Some wanted it easier but some wanted it more difficult. In particular, some students seemed to see the QAT as an opportunity for some truly challenging assessment: “Put something more interesting in it, like MATHS. Also, these tasks don’t test your computer skills, unless computer skills are clicking a mouse & typing a few letters”; “More challenging problems. Maths equations. General knowledge questions.” Together with a number of suggestions simply not to do it at all, were some that expressed satisfaction with it the way that it was: “Nothing it’s brilliant”, “Nothing. I believe that it was well structured & presented. It wasn’t too easy, wasn’t too difficult but the questions were challenging and made you think.”

There was strong agreement between these conclusions drawn from the survey data and those derived from the qualitative data analysis. Students not only thought that the instructions could be clearer but wanted greater emphasis on reading the instructions carefully, for example: “Make sure that we are told to read the instructions more carefully” (CS FG2). This was borne out in others’ responses such as: “My first reaction during the sign task was to use words. Then I realised it had to be pictorial” (ICBS OS). Another similar suggestion:

- What you could improve is how well the questions are explained. I got stuck on one of them because I didn’t know what to do. (ICBS OS)
- Question 7 needed some fine tuning, it was very vague in what to do. (ICBS OS)

Students also wanted greater variety and more questions:

- Less landscape questions and more on tourism etc. (ICBS OS).
- More variety please (ICBS OS)

What is overwhelmingly apparent from the qualitative data analyses is that students were uncertain about what was being assessed. This is an important issue that in future must be addressed. Students need to understand the purpose of the assessment which some saw as an assessment of SOSE, Geography, general knowledge, level of Year 9 students or ‘smartness’. To further illustrate the range of comments students made about this aspect of the QAT a selection of responses are offered:

- How to use symbols to convey meaning without words. (CS FG1)
- Content of SOSE and in particular mapping. (CS FG2)
- Couldn’t see the point of the last question about texting. (ICGS FG1)

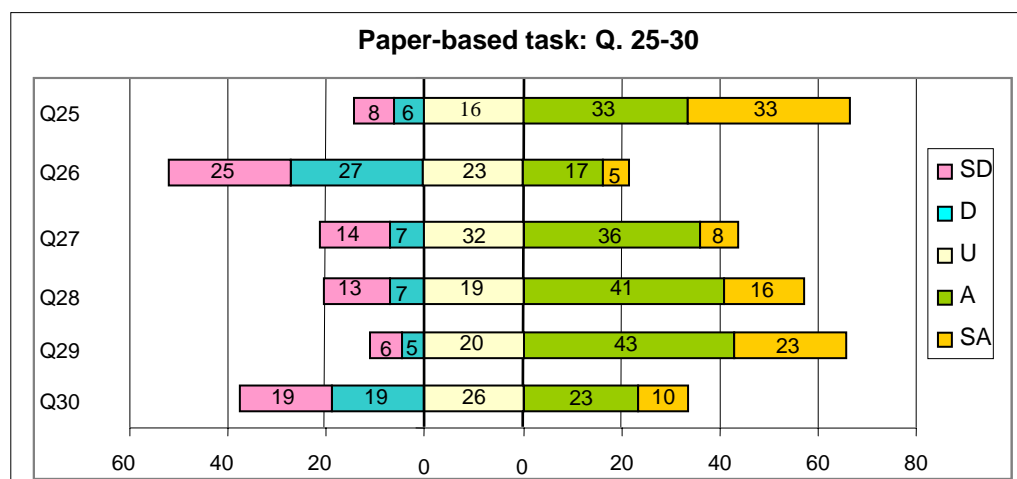
- What will be gained from this? What this shows us about us? (sic) (ICBS OS)
- Questions requiring common sense and logic not memory. (ICBS OS)
- One question I thought was really pointless was the question on painting/drawing a picture on a road sign.

The implication of the range and variety of responses suggests that students need to be further informed of what is being assessed and why. This aspect is important to consider because if it is neglected then it becomes a threat to the validity of the assessment task.

### Paper-based task

The survey questions relating to the paper-based task broadly mirror those asked about the computer-based task, except that there was no section on support. Students were asked to express their opinions of the task by indicating the extent to which they agreed with the following statements.

- 25. The paper-based task was more difficult to complete than the computer-based task.
- 26. I enjoyed completing the paper-based task.
- 27. There was enough information given to complete the paper-based task.
- 28. There was enough time to complete all the questions.
- 29. The paper-based task caused me to think carefully about my answers.
- 30. The content of the paper-based task was interesting.



**Figure 5. Percentage student responses to Questions 25-30, Paper-based task**

The students were also asked about the paper-based task in the focus group interviews and the open-ended survey (44 responses). The analysis of this data has been compared with the survey results and findings have again been triangulated to increase the validity of the findings.

## **Level of Difficulty**

The analysis of the survey data indicates that although most students believed there was enough time to complete all the questions (question 28) and that enough information was given to complete the task, 66 percent of them agreed that the paper-based task was more difficult than the computer-based task, only 21 percent indicated that they enjoyed doing it, and only 33 percent believed the content was interesting. Most students agreed that they had had to think carefully about their answers. The clear sense that emerges from these responses is that of a more demanding, less enjoyable task than the computer-based task. The findings from the qualitative data align closely with these conclusions and also suggest that the students found some of the language used in the questions beyond their comprehension. For example words like 'parody' and 'driving force' proved to be terms that were not entirely understood.

The most difficult part of the paper-based task, according to most students (survey question 31), was the section "American Gothic — Not!". However, few students explained what it was that was particularly difficult about it, except to say that the instructions were not clear. The other sections of the task were also nominated as the most difficult by some students, but again with little indication of where the difficulty lay. A couple of students highlighted that difficulty did not preclude interest: "The American Gothic section was challenging, but also fun and interesting"; "The American Gothic part was difficult but was still interesting." Some insight into why the students found this section particularly difficult emerged from the qualitative data analysis. Students had difficulty understanding some questions in this section and also found the questions ambiguous in the way they were framed. Some didn't understand what was being asked of them and indicated that not enough information had been provided.

## **Level of Interest**

In response to survey question 32 ("What did you find the most interesting part of the paper-based task?"), it was "American Gothic — Not!" that also featured most prominently, and there were some specific indications of what it was that was interesting. Many students enjoyed locating the pitchfork shapes in the painting, but some also mentioned the probably more challenging questions about the nature of parody. The calculations across currencies and numerical systems in "An Australian Traveller in Egypt" were also referred to quite frequently. The findings from the qualitative data aligned with these from the survey. Students indicated that the section "American Gothic – Not!" was most interesting and some students indicated "An Australian Traveller in Egypt" was also interesting.

## **Learning**

Most students indicated (in response to survey question 33) that they did not learn anything from the paper-based task. Those that indicated that they did learn something generally referred to the factual content of the questions, sometimes seriously, sometimes flippantly ("Yes, that Tasmanian tigers scientific name is Thylacines! Wow! such knowledge!"), sometimes erroneously ("Just that Hitler was connected to the death of the Tasmanian tiger. I also learned that there is a place called QAT National Park"). One student's comment — "your (sic) not supposed to learn anything in a test, your

(sic) supposed to know it already” — does not seem inconsistent with most of the responses to this question. From the qualitative data analysis it was also apparent that students did not think that they had learnt anything from the paper-based task. There were some comments that suggested the student had a better understanding of parody and that in Egypt “their money is different”. Again some students demonstrated their lack of understanding, for example: “Who’s Hitler?”

### **Required Skills**

Students’ responses to survey question 43 (“What skills did you have to use to complete the paper-based task?”) also seemed to display less insight than their responses to the corresponding question for the computer-based task. Reading, writing and knowledge account for the majority of responses. The findings from the qualitative data do not expand on this, for example: reading, writing, common sense and general knowledge.

### **Suggestions for Improvement**

Similarly, the most common suggestions for improving the paper-based task (survey question 44) resemble those for the computer-based task (more fun, more interesting), but without providing the additional information that was present in some of those earlier responses. Several of the more articulate responses called for more multiple-choice questions or for less open-ended ones: “Some of the tasks were a bit abstract and irrelevant. A lot of the questions were very opinion based and had no clear wrong or write answer making the question somewhat difficult to understand. Needed a few more practical and challenging tasks.” These findings aligned with those from the qualitative data where there were suggestions for “more explanation”, “more information about the questions” and generally “more help”. Students indicated that they had adequate time to complete the tasks.

### **Timing for the Collection of Student Responses**

A recurrent theme throughout all the written responses was that the survey should have been administered at the conclusion of the task, not at a considerable time afterwards. Many students said they had forgotten it. However, given this view, it is interesting that so many students still had detailed and forceful things to say about the QAT. Although many of them claimed not to have learned anything from it, many nevertheless retained clear impressions of the experience and opinions about its quality, and even suggestions for improving it.

While the open-ended survey and the interviews were conducted immediately after the administration of the 2005 QAT the findings and the emergent themes correlate with those from the survey. The students’ enthusiasm and interest in the computer-based task were expressed particularly emphatically in both the interviews and the open-ended surveys. One example:

- I really enjoyed completing this test because it was an exciting, challenging new experience. I liked the way they gave you the opportunity to interact and do everything yourself in a new style. (ICBS OS)

### **Overall Opinions**

The final section of the survey called for “additional comments” which provided some expressions of opinion on which of the tasks was better (with

more students expressing a preference for the computer-based one), and also overall verdicts, both positive and negative:

- It's good to have a different type of assessment to the normal!
- It was very boring. You have to completely change it!!
- I think it would be good for all year nines to do the task as it is different and interesting and has something for everyone.
- The computer-based task seemed pointless. Maybe if you added something telling us what we were being marked on and why it was important for each question.
- Nice, different and unusual but I like it more than other exams.
- This was a pointless waste of time. Pick questions that all Yr 9s are learning about. Spare the paper and save the trees.
- I found this interesting compared to other state/national tests because of the computer section, but also the written section has the information provided and it was new, and different to things we'd learnt in school.

Similarly an analysis of the interview and open-ended survey data collected immediately after the QAT correlated with those views expressed above. A selection is given below:

- I feel lucky to be able to take this test. The test did not leave me exasperated. It gave me good practice for problem solving. It was good that I got the opportunity to learn other non-school related things. (ICBS OS)
- The test was interesting because I wasn't expecting to have make (sic) a sign but it was fun and made it more fun. Though I found the contour drawing a bit challenging. But it was good because it was different. Though I do not see the point of this and how it relates to any thing. (ICBS OS)
- Overall the written test was bit poor but the computer one was GREAT!!! (RS OS)
- If I had taken the task more seriously I would have understood the task a bit better. (CS FG2)
- I found the computer-based task easy, enjoyable, fun and felt that it related to our own experiences. (ICGS FG1)

## **Conclusion**

In summary, the students' quantitative and qualitative responses give a clear sense of acceptance of the computer-based task, with the paper-based task being generally acknowledged as the harder and less interesting of the tasks. The qualitative responses give a more mixed sense of the students' responses, with blunt negative comments like "boring" more prominent than the quantitative data might have led one to expect. However, there are strong positive threads as well as these negative ones running throughout the entire qualitative data bank. In particular, there is a strong sense that the QAT is different from other forms of assessment, both in terms of its use of computers and its challenging nature. This student's listed response tends to sum up the range of comments made by others:



“Very creative!  
Very modern  
Great use of graphics  
It was very user friendly  
Very fun  
Didn't feel like a test  
Good technology bonds  
Great help functions  
Hard to go wrong  
Pretty useful learning experience  
Lives through the student  
This will revolutionise the word “TEST”  
Bring on MORE!” (ICBS OP)

## TEACHERS' PERSPECTIVES

Following the implementation of the Queensland Assessment Task (QAT) in 2005, 28 teachers and principals from 27 schools (26 state and one non-state) that had been involved in the QAT were interviewed. The schools were geographically and socioeconomically diverse. The interviews were transcribed (amounting to almost 50,000 words).

The interviews were structured, with the questions relating to the following main issues:

- the interviewee's role
- relevance of the QAT
- the intellectual challenge and motivation provided by the QAT
- equity
- alignment of required performance and targeted intentions
- the preparedness of the school
- the usefulness of the QAT (in terms of teaching, student learning, planning and school publicity)
- the implications of the QAT for the development of a common assessment task
- suggestions for improvement.

### Interviewee's Role

The role of the interviewee in the QAT varied from being a principal with a close interest in the QAT but little direct involvement in its implementation, to teachers and deputies who were almost solely responsible for the implementation of the QAT in the school, to a teacher who was involved only in the marking of one component of the QAT. Most of the interviewees, however, had been involved in the practical preparations for the QAT, including dealing with the QAT team and enabling the implementation of the computer-based task. This may partly account for an important common feature of the responses: a tendency to think of the QAT primarily as a computer-based task. There are probably other reasons for this perception as well (discussed below), but it does seem that the involvement of many of the interviewees in the logistics of this component (buying USB drives, negotiating use of computer labs, preparing computers) flavoured their perception of the nature of the whole enterprise. In some instances, teachers who responded quite perfunctorily to questions about the relevance of the QAT would express themselves more passionately about the logistics of it. This side of the QAT, for those teachers responsible for it, was clearly a major undertaking, which may have contributed to the computer-based task looming comparatively large in their recollections.

For some teachers, however, the situation was reversed:

I only saw the written test. I didn't see the computer one because that was on the other side of the school.

## Relevance

Teachers were asked one question about the relevance of the QAT to teaching and another about its relevance to student learning, but their responses did not generally make a clear distinction between the two areas and can probably best be considered together.

Responses about the relevance of the QAT were quite divided. The teachers who saw little relevance in it often had a sense (or reported that their students had a sense) that it was a one-off, out-of-the-blue experience which did not fit into the context of what teachers and students were already doing:

Not particularly relevant to what we were teaching at the time, but you know we will never change what we were necessarily teaching.

Well, we haven't, the way that we've actually done it here we've just conducted the test we really didn't build it in anything at all towards our teaching.

It didn't really affect them learning different things cause they really weren't aware of what was going on, what with it being a trial.

They didn't see it as relevant because they thought, "Oh it's not for a final mark, it's not going anywhere" but we tried to explain that's not necessarily the point of testing, it's good experience for you etc.

It's so limited, because it was done over a short period of time and very specific, had a very specific focus.

Teaching at this point not that relevant because it was just a one-off test, no results back.

Well, it was an add-on to what the students were doing. It wasn't integrated into what the students were doing as such.

The strength of this theme in the more negative responses seems to point not towards specific features of the 2005 QAT but, by implication, towards any manifestation of external, standardised testing, and may reflect the strength of the tradition of school-based assessment in Queensland. One response — "See we had a bit of a problem. We didn't know the kids were supposed to learn, I suppose" — would seem to express the extreme if this tendency to dissociate valued learning from external assessment. That the QAT was not only external and standardised but was also "just a trial" brought its relevance, for these teachers, even more severely into question.

On the other hand, teachers who saw relevance in the QAT tended to find it precisely where the other teachers located its irrelevance. Its external nature was seen by some teachers as a valuable experience for their students in terms of test-wiseness, in particular with a view towards the QCS Test in Year 12:

Pretty relevant in the fact that we actually looked through the written response task and felt that it was really like a junior QCS Test which is one of our focuses here, to get the kids better prepared.

I think it's provided a good opportunity for them to do formal testing because of the way that the curriculum is at the moment we do a lot of problem-based work and group activities and so it provided a really real

situation that would prepare them for senior, the assessment process in senior. So the process was relevant because it filled the gap that we needed to fill.

I think it introduced the idea of the type of testing that is done in Year 12, the QCS testing, like it's starting to get them to think along the lines to think like that.

I think doing the test was a good learning experience certainly in terms of the fact that they all know they have to do QCS at the end of Year 12. So just learning the structure of that test. In terms of the IT it was quite easy, the IT section the skills that they had to use in IT were not difficult at all. So as a test of their IT skills it was very simple, I would say too simple if you were trying to test IT. But in terms of learning about that type of test it was good, it was worthwhile.

It seems likely that these opposed positions on the relevance of the QAT represent previously established positions on the value of external testing. While they may not relate directly to the quality of the 2005 QAT, they certainly point to an important aspect of the Queensland educational climate.

One teacher made a comprehensive effort to place the relevance of the QAT within educational developments in Queensland:

That's difficult. I found that it was relevant in terms of where we're at in our outcomes roll out and working out how to assess and developing good assessment. It was relevant in that it gave us a really good example of good assessment technique I thought. It was also relevant in terms of as a school understanding where we sit within things like the standards map and how our practices at school align with those things being developed elsewhere in the department. So in terms of relevancy they're the key things I'd say.

Comments about the relevance of the QAT did not usually mention specific parts of the QAT or the specific content or skills or domains found in the QAT. (This is true of the interviews as a whole, and in this respect they stand in marked contrast to responses to the student survey. The students were much more ready — perhaps more able — to differentiate one part of the QAT from another, and express opinions about them.) The extent of the distinctions made in the interview tended to be computer-based versus paper-based, with more attention being given to the computer-based task. Where teachers did make this distinction, they tended to be more positive about the computer-based task; in fact, while some teachers had positive things to say about the paper-based task, none of them actually expressed a preference for it over the computer-based one. Positive estimates of the relevance of the computer-based task were made by teachers who were generally positive about the relevance of the QAT, as well by some who were generally negative:

The fact that the students said they would prefer that type of exam rather than pen and paper means to me that it is relevant because their world is more online.

The actual computer-based task I thought was relevant. But in terms of the content of the task we kind of had to teach things like contours out of

context. So relevancy I'd say is an issue there, but in terms of the actual skills required in the computer aspect, really relevant, yeah.

They got right into the computer task. I don't know how well they did but they certainly seemed au fait with the IT part of it all and so I'd say better than average in terms of learning.

In summary, then, the main features of the QAT that were referred to when discussing its relevance to teaching and learning were its external nature (on which opinions were divided) and its use of computers (on which opinion was positive). Positive opinions on either of these issues are presumably based on positive reactions to other issues (that is, teachers who value external assessment and computer-based assessment presumably would not be positive about a *bad* external test or a *bad* computer-based task), but it is worth noting that these probably deeper issues of relevance and quality were passed over in favour of dealing with the more general issues of external and computer-based assessment.

### **Intellectual challenge and Motivation**

Teachers were asked if the QAT was intellectually demanding and if it motivated the students.

All teachers agreed the QAT was intellectually challenging. However, in considering these responses, it is important to bear in mind two related but not identical ways in which this question seems to have been understood. On the one hand, intellectual challenge was seen as a good thing, offering students the opportunity to extend themselves beyond what they are normally expected to do.

There were things that stretched the students and pushed some of the better students along the way, yeah. It was interesting that some students took a lot of time over some item and tried to get absolutely perfectly right.

I think the tasks were well graduated in that the first part the girls said was not all that difficult, and then by the time that they got to the American Gothic, you know, the better girls said "You know I could do that" and the girls who are not as good said that they found that challenging.

I found that the written component of the QAT was really quite demanding for our students. I can see that there were easy questions on it but the questions that required higher order thinking, I think our kids would really find that challenging, yep.

On the other hand, intellectual challenge could be treated as a euphemism for "far too difficult for my students".

Yep, definitely, cause students didn't really get what they were doing.

The overall task itself I think was demanding of our kids and I take it that's what you mean for me to answer for my kids, found it very difficult when we couldn't answer any questions.

A thread throughout the general opinion that the QAT was challenging was the view that the questions ranged in difficulty:

Yeah I thought there were some good tasks in there and the students enjoyed them. They enjoyed doing something that was different. You know some were. Some of the tasks were challenging, some were obviously very easy but I think there was a good spread there to keep most of the students involved and interested.

Yes I do especially the written task. My concern with the written task would be the length for a year 9 student that they were at a fairly high level and to ask a year 9 student to sit for 90 minutes and expect that the last set of questions are going to be as well done as the first set of question is a very high expectation. The computing situation may be they didn't feel ... I think it was challenging without them realising it was challenging. Because of the difference, it wasn't all written and reading response.

I thought in making up a range of tasks, because I've certainly had plenty of experience in this area, you'd need to be able to set tasks obviously which kids feel as though they can achieve in, because if you just blow them away with stuff then the kid will get to the third question and just look. They'll have a go at it but there won't be any commitment. Good tests need to spread kids in my book and need to challenge kids and there's plenty of challenge in that. No doubt there will be plenty of people who will say 'way too hard, over the top' etc etc but there will be a kid sitting back there who will have done that test somewhere who will come away and say that was a piece of cake. He wasn't challenged. And maybe 99% of them were, so I say good test.

Again it is evident that there were different responses to the computer-based and paper-based tasks, with the paper-based one being seen as the more challenging. Clearly this does not mean that the teachers believe that paper-based assessment is necessarily more challenging than computer-based, but it is noteworthy that in this particular case the teachers believed that the students had more difficulty with what was the more traditional medium of assessment.

In response to being asked if the students were motivated by the QAT, almost all teachers said that the students found at least parts of the QAT motivating. The wholly negative responses were few:

I would say no.

Probably not, no. The whole thing just seemed to come out of nowhere. We got very little information as to what the test would be like. All they knew was that they were going down to the hall to sit this test. I don't think the students give it much importance really.

Nup, not at all.

There were more wholly positive responses, which tended to emphasise fun, novelty and stimulation:

Yeah they enjoyed it. They thought it was fun.

You are always going to have those kids who are intrinsically motivated and always reach to strive to achieve at that level but on the opposite hand to that you are always going to get those kids that aren't. But we found our kids very keen to do it, especially the computer aspect of it.

By that question I sort of go to did they find them interesting or fun, which is what middle schooling is about as well, as you know, engaging them in their learning through that sort of fun approach. I think there are lots of different stimulus (sic) there to motivate them. At least it wasn't just all lots of text, written text so in that sense there was a variety I suppose. It didn't motivate them to go on and pursue it afterwards. If they did I wouldn't know about it but the nature of the exam itself, the written one, I think there was enough variety there, yes.

Yeah I do actually. I read quite a bit of the student feedback and I think they were quite engaged in the task, yep.

I think the kids found it a novel experience. I think it was good for year nines to actually go into a situation like that. Especially since we did it straight after our QCS Test with the Year 12s and so they sort of felt quite grown up about the whole process and having to do something to time and something that was new to them. Yeah, it was pretty good I thought for them.

Most of them were interested in the task and carried it out seriously and some of them actually said they quite enjoyed it.

Our students seemed in the majority to be fairly focused on it. I mean, as I said, the computer task I mean nobody stared out the window at any stage. They were all focused. And I think in the written task, I wasn't in every room, there were three or four classes that did that test. I think in the main most of them were quite focused, quite stimulated, yep.

Some teachers expressed the view that some parts were motivating and others weren't, or that some students found different parts more motivating. The computer-based task tended to be seen as more motivating than the paper-based one:

Initially when they did the written task they were motivated when they went in to it because it was a different form, seeing how the level was pitched at, was way above them, in a lot of areas, and it required so much prior knowledge that they didn't have. Half way through, it was actually demoralising. They were very negative. They felt as though, that if they were getting assessed at a Year 9 standard that they weren't meeting the mark. Therefore it created a lot of agitation. It was not a very good atmosphere to be in about three quarters of an hour in. However, the one that was done on the computer was met. That was great. It was fine, loved it, motivation was fine. No grumbling whatsoever for that one.

They enjoyed the computer task but the written task they probably weren't as motivated by it.

Students are all different, so some parts motivated students more than others. From the student surveys that were done what appealed to some students, and they really enjoyed doing, didn't appeal to others. It will

vary. There was no particular (task). The ones that kind of stood out like the Gothic-not, some just absolutely hated it where other students loved doing it, the same with the computing task. Some students love the idea of the visual and the diagrammatic ways of expressing themselves and others didn't. I think that shows the diversity of interest. Depending on how interested in some of those situations and topics as to how well they went.

Definitely for the computer-based one. ... at our school anything to do with computers they are ... switched on, full attention and they'll do it. And the mobile phone question was particularly good, cause that's quite up with technology. ... the kids loved the one with the gradients and compasses ... Where, you know, the opportunities for them to get it right or wrong were more so. Heading towards letting them pass or a correct thing because they had interactive, and they could watch the arrow move. ... so they would probably get a closer answer than if they just had to guess without all the other interacting pieces. With the written side they definitely, they loved the maths part of it, they got into that quite easily. But, yeah, the one about the Gothic was a bit more challenging and a lot of the kids aren't into art at our school so they weren't as motivated towards those questions and I found that lots of them would just skip those questions.

An interesting aspect of the responses to this question is that they tend to be more elaborated than responses to most other questions. Teachers seemed particularly interested in discussing the issue of students' motivation to do the task, and in discriminating between *degrees* of motivation. The detail of many of the positive responses here is as telling as the explicit, at least partly positive view being presented. The QAT appears to have engaged the students to an unusual extent, and this fact itself seems in turn to have engaged the teachers. (The teachers who saw no motivation in their students were not correspondingly interested in the source of this lack of motivation.)

Taking the responses to the two questions about intellectual challenge and motivation together, there is frequent agreement that the QAT was both challenging and motivating.

### **Equity Issues**

An assessment task that is seen as being intellectually challenging and heavily demanding in terms of technology may perhaps be expected to prompt questions about equity. Responses to the question, "Do you think the task supported all students in the production of a performance of high quality?" certainly did include a range of negative views:

Special needs students, with some of our ascertained students there's no way they could have coped with that sort of, not just the test but the change... You know it can send kids off so I don't think ascertained kids... It depends on the nature of the ascertained kids. So I can't give a general question there. We did evaluate it in terms of each individual kid and as I said only one ascertained kid was put in there because we knew he could cope with the circumstances.



No it didn't cater for them at all because obviously the students that didn't work well with having an exam were stuck in the situation where they had to do it so they couldn't cope with it.

We have ... 42, probably a 40 odd per cent if not more indigenous population and remote. So I think ... I don't think it sort of really helped those kids out that much, no.

No I don't. In some ways it was too culturally specific.

No, way off the mark in both cases. For your average student it was way off the mark but in terms of intellectually impaired or Indigenous no way did it meet anywhere near what they needed.

The technology could have been a bit of a problem with students with special needs and particularly indigenous students. They would have difficulty in accessing as easily. They're just not as familiar with it. Not as quick to pick up what was required.

Again, however — despite the final comment just quoted — there was a tendency for the computer-based task to be regarded more positively than the paper-based one:

Some of our students did have, saying that they enjoyed the computer pilot, which they did but some students had difficulty in completing those tasks successfully. And I guess that comes from not necessarily being familiar with the programs that we used. But, you know, they all had a go even if they didn't necessarily complete the tasks successfully. So yeah you know a lot of our LD kids did quite well within that.

Probably, on the computer one I would say yes because there was lots of interaction and things and the kids could look up things if they needed to or go back and check. With the written side, I felt that myself struggling with reading some of the questions and thinking I had it down pat the first time and then re reading, kids don't re read so and the ones that have reading and literacy problems they would have found and struggled quite a lot with those ones.

I'd say so yes. The computer task was good for that.

I think the computer task used that more effectively than the written task. I think you always get, while in the computer task there was there was a lot of information to read which could have disadvantaged some students who aren't great in their English ability, there were lots of visual clues to support the students and the practice was supportive of that. I think once again in the written task, obviously, there would be disadvantaged students from ESL or others from Torres Strait Islander backgrounds and some of the language may not have reflected cultural groups within Australia.

Comments that were wholly positive on this issue tended also to be perfunctory:

Yeah no problem there. I didn't see that as being anything major. Yep pretty good.

It seems, then, that there were quite a few concerns about equity issues (in relation to language level, cultural references and the actual situation of a novel form of assessment), but that, in some cases, these were assuaged to some extent by the computer-based task. Again, this clearly does not mean that computer-based assessment is regarded by any of these teachers as *necessarily* more equitable, but in this particular case the approach to computer-based assessment manifested in the 2005 QAT was seen by some teachers as an enabling one for some groups of students.

### **Alignment of Performance with Targeted Intentions**

Teachers were asked, “Do the tasks require performances that are relevant and adequate for the targeted intentions?” This question elicited mostly positive but perfunctory responses. The following are representative:

Yeah, I think they did that, yep.

Yep, yes definitely.

Yep, I think they covered that quite well.

Yeah I think they did.

The lack of elaboration here may suggest that the question did not clearly communicate its meaning, and that teachers were content to agree with something that sounded as if it should be a good thing. In particular, the meaning of the “targeted intentions” may not have been understood. If this is so, the generally positive responses may suggest enough of a feeling of good will towards the undertaking to choose perfunctorily positive over perfunctorily negative, but cannot be used to obtain any information about teachers’ views of the alignment of the QAT with curriculum aims.

There were, however, a few responses that were both positive and reasonably elaborated, expressing the view that the QAT did in fact assess important elements of the curriculum:

I think if you answered those questions, they certainly had to analyse, they had to construct, they had to deconstruct they had to do all of that.

I think in general ... yes, things like transforming from one text to another the processing that students have to go to is a very important thing that we in the classroom are doing and requiring our students to move towards. And in terms of the test reflecting of the task, reflecting that, that was very good.

Two more negative responses were also reasonably elaborated, and both suggest a sense that the QAT represented a narrow view of a potentially much richer field of curriculum and assessment:

I think that there are better ways of assessing those skills. As I said we are a New Basics school. We’ve been doing the New Basics curriculum now for four years and I believe that the Rich Tasks are a much better way of assessing those types of skills than the QAT test.

I think the computer one was too narrow ... a focus. It was high on those topographical maps, so it wasn't broad enough. It also tested stuff that could have been done on pen and paper, it didn't gain anything greatly.

## Preparedness of the School

In relation to the school's readiness for the implementation of the QAT, teachers were asked if they were adequately consulted, informed, resourced and prepared. Generally teachers said they were prepared; in a few instances where they were not, it was because of the situation within the school (staff changes, or poor communication). They tended, however, to point out various difficulties that they had had to overcome on the way to being adequately prepared. While personal comments about the helpfulness and expertise of the QAT team were always positive, some comments about processes were not. In particular, the use of the website to communicate with teachers was questioned:

Consulted yes, but not, teachers through websites, you know they just do not have time to get on and keep checking the website everyday. That was absolutely ridiculous, even HODs, myself, you know we are so busy that to just get an email saying 'Oh just check the website' was totally inappropriate. And so in that sense we were not as ready, as ready as we would like to have been. That's the main feedback there.

Yes, except their web sitey thing wasn't a huge success. With the timelines and stuff it was another thing to learn. Mate, there are some technos in there that got carried away.

Yes, the information process was interesting. I don't know that the QAT's website was a very effective way of doing it. It was rather complicated. The principal, he couldn't understand how to do it. It took a fair bit of fathoming to work out what was going on but yeah besides the QAT website, the information was there it was just a little bit hard to access sometimes.

Other sources of frustration include difficult timelines, not knowing exactly what the QAT would look like, and negotiating the use of resources within schools:

We kept getting lots of information that it was coming that there would be a written test and an online test and I really don't think til we saw the paper we had any idea what the questions would be like. It was very difficult to try and explain to the students what the test would be about because we just didn't know.

I'd say yes but ... I'd the timelines for consultation. We'd get the information really quickly, and have to apply it, and so in terms of me then being able to consult my staff, they felt really pressured and not having a lot of time to digest or to get ready.

It was frustrating at times. In terms of, we plan a year ahead, it was frustrating at times when people were asking us to suddenly find these rooms, these computer rooms, and ask us to set them up a certain way which meant that nobody was allowed in the rooms before because of the set-up procedures. A lot of people book computer rooms a term ahead and we were having to deal with this with maybe 4 weeks ahead. I understand this because it was a trial. That was quite difficult at times. The immediacy of things with the workload we already have. I really did find hard the getting the names downloading the names and handing

them over to QSA and having to have it done by, one minute it was a Friday and then they changed the date and with everything else going on in the school some of the deadlines were a little difficult to meet and I felt bad about that but I was trying as best I could.

In light of the considerable technological requirements of the computer-based task and the inherent difficulties of doing something out of the ordinary with numbers of students within the often strictly defined structure of a school, it is interesting that while teachers often mentioned the workload involved in preparing for the QAT, they generally said that they were in fact ready for it. Several teachers warned, however, that expanding the QAT to a larger base of schools would prove difficult, especially if it entailed the use (especially the purchase) of USB drives.

## **Usefulness**

Teachers were asked how useful they thought the QAT would be in terms of informing teaching, improving student learning through feedback, a resource for planning, and school publicity.

There was a strong sense of potential usefulness in the responses, often tempered by the observation that until they received the results this usefulness would remain only potential. The usefulness tended to relate to building on information about how their students are performing in relation to others, or to identifying issues in curriculum planning:

Well it will tell us whether or not we're on the right track, whether we're using online material properly and in an intelligent manner so that it is part and parcel of our curriculum not just, not just something that is quite divorced from it. I think in today's world, in today's society if you can't find your way around a computer screen you are illiterate. It will tell us whether or not our students are literate.

I think it will be very useful and I actually plan on sitting down with the teacher, whose class we used, myself, the Deputy and the Principal and talk about how that will actually inform teaching, learning.

We're always comparing ourselves to other schools. We're always looking at data. I think it will be a good starting point. I suppose we'll get two different lots of data, one for the written and one for the technology. It would be good if when we received that data if we could have a copy of the exam again so that when we discussed the data with the staff we have a reference point. We could talk about where they've done well and where they need better skills, where they need to develop, yep.

It will be useful for next year. Depending on how the information is presented to us whether it's just an overall grade or it's broken down into sections. It could allow us to see where individual students strengths or weaknesses lie and also where schools strengths or weaknesses lie.

We're always keen at our school to actually align our assessment with overall assessment. So we want our assessment to complement what's going into the standards map. So we want the constructs to mirror what we've got in our assessments and those sorts of things. We'll definitely be using that information to keep that alignment and probably tweak that, yeah.

I think it will be very useful because we have been searching for baseline data for a while now, in terms of just where our kids are sitting. It's also useful to us because we have such a large group of non-English speaking background students that it will be interesting to see how they perform across the state because there are lots of issues in terms of language with those kids and being able to do a test like QAT which is really very heavily language based. For them it will be interesting to see how they perform.

While these responses suggest a sense of enough congruence between what was done in the QAT and what is generally done in the school, this view was not universal:

I don't think it's going to be of any use. Mainly for the reason that it hasn't aligned with what we've been teaching so it's just assessing something so far from what we're doing. So at this stage I don't think it's going to be of a great deal of use. Also because of the motivation factor of the students cos I was there when they were sitting it I know that it's not going to be a true indication of what they're capable of.

On the issue of using the QAT for school publicity, this did not seem to be something the teachers had considered. One response, however, sums up a common immediate response to the possibility:

Yes, if we do well. No, if we don't. I didn't say that.

### **Implications for A Common Assessment Task**

The Queensland culture of school-based assessment was evident again in responses to questions about the usefulness of the QAT for the development of a common assessment task. Some teachers called for teacher involvement in the creation of future tasks, particularly to try to ensure an appropriate level of difficulty for Year 9 students. There was frequent support for the idea of a computer-based component, but concern about the practical problems inherent in that approach.

Several teachers responded to the question with reference to the place of common assessment tasks within QCAR, how they would relate to the essential learnings, and the implications of this for possibly increasing the uniformity of Queensland education.

### **Conclusion**

The stated opinions of these 28 teachers, while not intended to lead to a definitive verdict on the nature and success of the QAT, can provide a useful insight into the educational climate in which the QAT was piloted, and in which any future common assessment task will be implemented. On the basis of these teachers' opinions, the QAT encountered a climate of conflicting views about the very thing it was intended to be — an external form of assessment, in addition to the established school-based assessment of the Queensland system. Where some teachers saw this as potentially providing a firm point of reference, or at least useful preparation for high-stakes tests in later years, others saw mainly a lack of congruity between the external assessment and internal curriculum.

The positive response to the computer-based task would seem to indicate a climate that is conducive to innovations that enhance the students' motivation, while the response to the paper-based task was more mixed. There is obviously no agreement among teachers from a wide range of schools on just how challenging a Year 9 task should be, however, all teachers felt that the QAT was in fact intellectually challenging (in some cases, too challenging). In addition, the teachers believed that in general the students were strongly motivated by the task. Perhaps the most powerful message to come from what the teachers said is that, despite a perhaps common assumption that motivating the range of students in the middle years necessitates a lowering of standards, these students can in fact be strongly motivated by work of genuine intellectual challenge.

## IMPLICATIONS FOR POLICY

This evaluation of the pilot of the 2005 Queensland Assessment Task (QAT) was conducted in the context of the development of the Queensland Curriculum, Assessment and Reporting (QCAR) framework. This new policy aims to identify the essential learnings and the standards that describe student achievement (P-10).

Other objectives of this framework include the provision of a bank of assessment tools that require a range of responses such as short answers, extended projects, practical demonstrations, performances and oral presentations. Teachers' professional learning related to assessment will continue to be promoted. A common reporting framework to describe student achievement using a common five point scale (A-E) will be introduced and comparable statewide assessment of student learning, in Years 4, 6 and 9 in the essential learnings in English, Mathematics, Science and one other subject, is intended.

The final dimension of this evaluation focused on the policy context. The prime purpose of this was to identify the key issues, challenges, opportunities and the significance of the pilot of the 2005 QAT for the QCAR framework policy-making arena. The implications for action and the possible areas for collaborative research that are derived from the experience and learning associated with the QATs are outlined and discussed. The utility of the experience for the future development of common assessment tasks was of particular interest.

### KEY ISSUES

Given that a major purpose of the QCAR framework is to align the intended learning outcomes of the school curriculum with the assessment and reporting of student achievement there is a need to consider the factors which affect the implementation of such change. In this context the major change for teachers, students, parents, carers and the system is the use of standards-referenced assessment. The standards will provide the common linchpin for connecting curriculum, pedagogy, assessment and reporting.

#### ***The Change: Standards-Referenced Assessment***

Communication of the centrality of the use of standards to parents, carers, students, teachers and the system will be fundamental. To begin it will be important to develop a common understanding of the concept and the purpose of the standards in the QCAR framework. To illustrate, standards can be defined in several ways such as follows:

- moral or ethical imperatives (i.e. should do)
- legal or regulatory requirements (i.e. must do)
- quality benchmarks (i.e. expected)

- arbiters of performance quality (i.e. defining success or merit) and/or
- learning milestones (i.e. progressive targets) (Maxwell, 2002).

In the Queensland context of KLA syllabuses, to date, standards have been interpreted in curriculum outcome terms as quality benchmarks (learning outcomes that illustrate the standard attained relative to that expected at a particular Level) or learning milestones (learning outcomes that represent the progress that has been achieved relative to the identified targets at a particular Level). That is, the progress a student has achieved towards the expected quality benchmarks or learning milestones reported as learning outcomes in relation to the KLA syllabus Levels.

What is to be introduced that will need to be understood by **all** is the interpretation of standards as 'arbiters of performance quality' or standards as 'defining success or merit'. This has not been the focus of standards in the Queensland KLA syllabus context and will need to be emphasised.

In terms of the QCAR framework teachers will use the standards to describe both the quality **and** progress of student learning. The standards will provide a common frame of reference for making judgments about the student quality **and** progress of student learning, and a common language for reporting. With the emphasis on standards as statements that indicate different levels of quality of performance, teachers will need to meet regularly to discuss student work for moderation purposes. Such professional collaboration will help prevent the repeated failure of attempts to implement standards that are based on the 'mechanical application' of explicit criteria (Cresswell cited in Klenowski, 2002: 64).

It is important to make the distinction, here, between criteria and standards. Criteria are the characteristics or dimensions on which the quality of student performance is judged. Standards are the levels of quality or performance expressed as reference points along a developmental scale or continuum (Sadler, 1987). A five point scale (A-E) will be trialed to describe the standards. The A grade will be the aspirational, C as acceptable and E as a positive description of what students know and can do (QSA, 2005: 21). Standards can be specified as descriptive statements to specify the points on the quality continuum and can also be conveyed in part by means of a set of exemplars or key examples to illustrate what distinguishes high quality from low (Sadler, 1989: 128).

The elements of the QCAR framework to be developed by QSA and the three school sectors include the essential learnings, to assist teachers understand what students need to know and be able to do, and the standards that will help them determine what students have learnt and how well they have achieved. In implementing this change the following issues will need to be addressed. Some of these emerged in the pilot of the 2005 QAT and are considered significant in the evaluation.



### ***Understanding the Change Process***

In the implementation of the QCAR framework it will be important for policy officers, teachers and principals to be aware of factors that impact on change. Some of these issues will now be illustrated and discussed to highlight their importance for future consideration.

#### *Purpose*

One of the difficulties that emerged from the pilot was the lack of understanding by students and teachers of the **purpose** of the 2005 QAT. Students were uncertain about what was being assessed, some saw the purpose of the QAT as an assessment of SOSE, Geography, general knowledge, level of Year 9 students or 'smartness'. Teachers too were not always clear of the purpose and those who saw it as a one-off experience did not believe it aligned with their teaching and learning context.

The tradition of school-based assessment in Queensland will need to be harnessed and teachers will need to understand how valued learning can be assessed both by internal and external means. The fact that the QAT was external, standardised and, for some teachers seen as 'just a trial', diminished the relevance of the experience.

This particular teacher's response tends to sum up the lack of understanding about the purpose of the 2005 QAT:

"I'm not sure what their goal or purpose was for this. Was it literacy and numeracy measures for QSA? It hasn't been made really clear. ... whether this means that there will be a Year 9 exam or whether they are just trialing this to see if there should be an exam? ... you wonder if it's going to be used to compare schools, or if it's going to support ... more of the Smart State stuff that comes out in the media and to the government."

A poorly conceptualised change or one that cannot be demonstrated will be difficult to implement. Teachers, students, parents, carers and community members need to know who will benefit from the QCAR framework and how. What will be achieved for students needs to be made explicit. As a member of the QAT team indicated:

"Some people say that teachers won't accept state-wide testing or assessment in other forms ... but our experience has been that in fact a lot ... will, provided that they believe it has been well thought out, that it's doing high level stuff, the kids are enjoying it, that marking is valid ... They don't throw out something just on the basis of its form. They are more likely to worry about the quality of what is presented. ... do I think this is worthwhile? ... teachers are looking for experience and direction."

#### *Resources*

When asked about the lessons learnt from the 2005 QAT, and the implications for the QCAR framework, interviewees identified the following

deficiencies. The timeline for completion of the project was shortened making the intended outcomes more difficult to achieve. For example, the planned trial of the constructed-response could not go ahead. The team also suffered from having to rely on part-time staff or contracted staff whose roles and responsibilities were not always clearly defined. This lack of clarification of role relationships resulted in misunderstandings among team members. Support and direction were also missing and this too contributed to demoralising and disheartening the QAT team. To illustrate:

“We literally dragged people out of homes with new born babies to say we need your expertise, come in now, and we will pay you and yes we can work around three days a week ... that is not really the way this should work.”

“...I think you need to have a long term plan, you have to have it established, you have to have people trained, you have to have the right people to do the computer programming and so on. You have to have stability, professionalism and you have to know what it is you are trying to do. You have to have the capacity to do it. ... You have to have will – which means you know what you want and what you are trying to achieve. And you have to have wherewithal – which is given that you know what you want and given that you know how to do it, you have to make sure you actually have the capacity to deliver. ... (in terms) of the history of the QATs – I would say lots of will, not enough wherewithal. I would choose wherewithal and make sure I was a lot clearer on the long-term will.”

If a change is poorly resourced or resources are withdrawn after the first wave of innovation is over then this will be a barrier to change. Insufficient funds for materials or time for teachers to plan can mean that the change is built on the good will of staff who will not persist for too long without additional support. From the many references to the under-resourced aspects of the 2005 QAT it would appear that these were factors relevant to the problems experienced.

#### *Communication and Commitment*

The following reference to the ‘Report of the Assessment and Reporting Taskforce’ identifies the importance of communication.

“For the enactment and implementation of a coherent curriculum, pedagogy, assessment and reporting policy across Years 1-12 in Education Queensland schools the following elements are needed: ....

- Clarification of the roles, responsibilities and communication links for all parties and better communication within Education Queensland, between elements of the new statutory authority and Education Queensland, and between Education Queensland and existing authorities until they are replaced;
- A spirit of collaboration among policy-making groups: the current inconsistency in policy is confusing for external and internal

audiences alike: it is inefficient and runs the danger of hindering strategic development. ....

The nettle of strategic leadership needs to be grasped if Education Queensland is to move forward: the taskforce sensed the feelings of frustration from teachers, school principals, curriculum developers and Education Queensland officials at the current situation.” (Education Queensland, 2002: 3).

The following quotations demonstrate the frustration QAT team members experienced due to what they perceived as a lack of collaboration and communication.

“I suspect ... that as one project ended (New Basics) and as QSA started getting responsibility for things and having to absorb things into their corporate structure a lot of messiness and confusion ensued ... At that level above ... operations level I think things weren't clearly defined quickly enough so that everything could run ... smoothly and it's to everyone's credit that in spite of all of that turmoil above us, people just got on with the job and tried to get it done.”

“Make sure that your people resources are managed correctly. It's very hard to deal with the situation. One of the problems we have had this year is the politics between QSA and EQ where we were left in limbo half way through the project ...”

“The management of the QAT project was changed half-way through which meant for most there was no job security, this was damaging to the people involved ...”

“It probably would have made a difference too if we hadn't 've changed half-way through. It would have made a big difference because a lot of people left. There was a lot of time there, where job security was very, and it still is, very up in the air. And so it's difficult ... “

The above quotes and the following one illustrate the need for commitment throughout the reform effort.

“Where I do think that they fall down is like this, a number of schools have rung me up to say that this was great and have asked what is happening next year. If we turn around and say we don't know what is happening next year, which is the absolute truth, then they will ask, 'What was the point of this year?' I think they have a cynical view too, the Department starting out big projects and nothing comes of them and schools put a lot of time and effort in, like everyone else. There's no follow on... The kids love it they want to know when there will be a next one.”

In 2005 there appeared to be no long-term commitment for the QAT to carry people through the anxiety and frustration of early experimentation or

unavoidable setbacks. Related to this factor is the lack of commitment by key staff who could contribute to the change or who might be affected by it. At the same time it is possible for key staff to become over-involved, as an administrative or innovative elite, from which teachers feel excluded resulting in resistance and resentment. These are further issues for consideration.

#### *Co-ordination of the Change*

If a change is pursued in isolation it can be undermined by other unchanged structures such as when, for example, cross curricular learning standards are juxtaposed with subject-based report cards or standardised tests. Conversely if the change is poorly co-ordinated with, and overwhelmed by, many parallel changes teachers will find it hard to focus their efforts. Such factors have considerable implications for the QCAR framework as there will need to be parallel changes to curriculum, pedagogy, assessment and reporting and these will need to be carefully co-ordinated and supported.

With regards the 2005 QAT what has been highlighted is the importance of the co-ordination of the intended change and the need to understand how the parallel changes intersect. To illustrate, because of the existing curriculum context in Queensland a generic assessment approach that focused on the construct of processing was chosen rather than a curriculum content approach. This decision was based on the 2003 QAT findings that curriculum coherence was deficient and there was a possible “dumbing down” of the curriculum. The development of an innovative computer-based task that would enable the assessment of generic skills became the priority. What the QCAR framework will build on in terms of the assessment process is the assessment of essential learnings which include:

- Cross-curriculum aims (purposes of schooling)
- Generic skills and attributes (lifelong learning attributes, cross-curricula priorities, employability skills) **and**
- Domain-specific knowledge, skills, understandings, dispositions and values (in and across KLAs, subjects, courses of study, training packages) (QSA, 2005b).

#### *Pace and Scope of the Change*

If change occurs too quickly then teachers, students, parents and carers may not be able to cope. On the other hand if the change is implemented too slowly then teacher and students can become impatient or bored and move on to something else. A teacher, involved in the marking of the 2005 QAT commented in this light that reporting back to the schools on the students' results was necessary but should also be timely. For example:

“I think to have schools participate in something and not give them some feedback would be inappropriate. And that is, it is not just done, but done quickly - within a month – otherwise it runs out of the mindset of people. ... these schools have been involved in the pilot. They volunteered and they have an expectation of something coming back, be it minimal or whatever. Otherwise it doesn't help to get other people involved the next time.”

As was also clear in the survey data the timing of the administration of questionnaires for evaluative or monitoring purposes should occur in a timely fashion and be fed back to teachers, students and policy developers regularly throughout the implementation process. In this case many students said that they had forgotten the 2005 QAT when they were surveyed nevertheless they did provide detailed and forceful comments about the QAT. The key principle is the importance of timely and regular feedback to the implementers and developers of policy.

If the scope of the intended change is too broad and ambitious and teachers have to work on too many initiatives this can impact negatively on the change too. For example in the United Kingdom, researchers Earl, Watson, Levin, Leithwood, Fullan & Torrance (2003) in their evaluation of England's National Literacy and Numeracy Strategies found that teachers were struggling with 'initiative overload' because of the scope and number of changes introduced simultaneously. Alternatively, the change can be too limited and specific which will not result in much change at the classroom level. If the findings from the experience of the 2005 QAT pilot are not used to inform the implementation of the QCAR framework then a lot of important learning and intellectual capital will be lost which is not at all cost effective, given the intent of a pilot.

#### *Student and Parent Involvement in the Change*

Students need to be involved in the change or have it explained to them. This did not happen to a satisfactory extent in the 2005 QAT pilot as was clear from students' and teachers' varied explanations of what they thought was being assessed and why. Students want and cling to ways of learning that are familiar and can become the school's most powerful resisters to change (Hargreaves, Lieberman, Fullan, and Hopkins, (Eds.) 2002). Similarly parents can oppose the change because they have not been kept informed and feel distanced from it. The changes that accompany the implementation of the QCAR framework need to be explained in student-friendly terms and parents need to be kept informed.

#### *Leadership of the Change*

The leadership of the change needs to be consistent, organised and managed effectively. Sometimes the leader can be controlling, ineffectual, or decides to take advantage of early success and moves on to higher things thereby jeopardising the change (ibid). To a considerable extent the management and policy leadership of the 2005 QAT suffered from a lack of consistency and support.

As has been reported elsewhere "transforming teaching and assessment approaches across education systems requires strong policy leadership, serious investment in training and professional development and innovative programmes and incentives for change" (Centre for Educational Research and Innovation, cited in Elwood, in press). The implementation of the QCAR Framework will require both QSA and EQ to commit to the change process. Policy groups will need to plan for the implementation and as highlighted here be aware of:

- The situational constraints;
- The values, ideas and experiences of those responsible for implementing the change and
- Knowledge and theory about the processes of change.

## **SIGNIFICANCE OF THE PILOT**

The significance of the pilot of the 2005 QAT has been analysed in terms of the implications for the policy context of the QCAR framework. The lessons learnt have emerged from carefully considering what has been valued and what has proven challenging in the development, implementation and evaluation of the 2005 QAT.

### **Changing Policy Context**

What becomes clear from this evaluation of the 2005 QAT pilot is the rapidity of change in the policy-making arena and the consequent impact at the level of design and development of intended assessment strategies. It is important however that from the outset teachers, students, parents, carers and administrators are aware of the intended purposes of the innovations or changes and why they are being introduced. To illustrate how the changing policy context impacted on the 2005 QAT some modifications and changes in emphases to the purposes of the QAT have been analysed.

One important original purpose of the QAT was to overcome the lack of state-wide data for the system. With better state-wide data about student achievement more informed decisions can be made about school development priorities and what measures need to be taken to improve student learning for all. As emphasized by one of the QAT team members:

“One of the overwhelming aims of the 2005 QAT was to provide the system with data. ... We tend to talk about the QAT in its original conception of being something that would be a statewide assessment task, which means that everyone in a year might do it.”

And as described by another QAT team member a further role of the 2005 QAT was to:

“... investigate whether we could go beyond school-based assessment to discover if we could administer a common task and as the foundations strengthen then the results will become more meaningful ... constituted a learning experience to build on.”

In the 2003 QAT design specifications the following was specified:

“... 6. A QAT is suitable for administration annually, in September, to students in at least one of the *junction years* ...

7. A QAT is suitable for providing systemic data on the achievement of a statewide sample of students from the cohort(s) described ...” (EQ, 2004:1)

The changing policy demands impacted on the 2005 QAT in that it was designed to provide assessment data related to achievement in the KLA syllabuses as well as fulfil other emerging purposes. The design incorporated the usual purposes of defining the task’s content and constructs to be assessed but was modified to include the use of ICTs, the assessment of generic skills and multi-literacies as well as literacy and numeracy at Year 9 level. The 2005 QAT was intended to serve these multiple purposes (EQ, 2005:65). During the period of design, administration and marking of the 2005 QAT the federal government announced that there would be national testing of literacy and numeracy at Year 9. As stated in the *QAT 2005 Reflections* report “The potential of a transdisciplinary QAT to provide an assessment of literacy and numeracy at the state level no longer required investigation” (ibid).

The 2005 QAT was able to provide assessment data on the construct of processing, in particular transforming ideas and/or information drawn from SOSE and the Arts. It also incorporated a computer-based task. Scores in literacy and numeracy were not extracted as these skills were interwoven in the interactive computer-based task and the constructed response task of the 2005 QAT.

A major difference in the context of the QCAR framework will be the identification of essential learnings and the standards. The options for statewide assessment that have been suggested include:

- QSA developed common tasks in accordance with the requirements of a design brief OR
- School-based assessment tasks, developed by or prepared for a school and approved by the QSA in accordance with the requirements of a design brief OR
- A QSA approved school assessment plan that indicates how schools assess student achievement and meet the requirements of a design brief provided by QSA.

These options for statewide assessment will be trialed to inform the development of future sustainable models of assessment. From 2008 at Years 4, 6 and 9 a common assessment process is planned to provide comparable statements of student achievement in the essential learnings in English, Maths and Science incorporating associated capabilities. It is intended that teacher capacity be developed to use the essential learnings and standards, develop assessment, and report student achievement that will contribute to improved student learning, school improvement and comparability of reported student results. (QSA, 2005b)

As QAT team members were aware:

“In devising a common assessment task you would be looking at what is the essence of (for instance) maths, ... That was the obstacle with

us; different people did different pieces so we avoided the specificities because everyone was doing different things ... in doing that we failed to significantly address the disciplinary areas. ... these need to be identified, the essential learnings, and they *will* be identified so there (will be) clear guidelines to teachers of what they are to teach ...”

“A major limitation of the *context* of the QAT was the lack of identification of the essential learnings.”

“The identification of the essential learnings and essential capabilities will help to sharpen the common assessment task. That is, make it a more powerful device if maintained as a complex of three (computer-based, constructed-response and performance-based).”

Teachers too indicated that:

“... common assessment tasks, as that concept grows there will be a real need to re-evaluate how we structure the curriculum and what informs that structure and move away from KLAs as we have known them. ... If the common assessment task can be of the same quality (as the 2005 QAT) because I think there is a lot of value in a test that is similar to what we are marking. I think that it’s on the right track and the sorts of skills that it’s identifying are essential, they are essential learnings for kids of today.”

To respond to such a dynamic policy context requires the team responsible for the design and development of the tasks to be supported and directed with responsible leadership which includes leadership in policy.

### **Co-ordination of Roles, Responsibilities and Relationships**

A key message from the 2005 QAT pilot was the uncertainty associated with the policy context and the role relationships between QSA and EQ. This meant that some hindrances to communication and inefficiencies occurred due the lack of articulation of professional boundaries. It is important that the roles and role relationships are made explicit to all and that channels of communication are kept open throughout the trial and implementation phases of the QCAR framework. Each organization needs to effectively support the other in harnessing and building the capacity of schools to use the essential learnings and standards, develop assessment and report student achievement for the purposes of improving student learning, school improvement and comparability of reported student results.

The responsibilities of key organizations need to align so that important learning and support can occur concurrently. This will help to provide direction and useful information in responding to the changes introduced by this major policy reform. Conceptually, the important relationships between curriculum, pedagogy and assessment need to align and this requires careful co-ordination across agencies, institutions and sectors.



One way of achieving such co-ordination and useful symbiosis has been demonstrated throughout the trial of the QATs by the work of ARFIC. The work of this committee should be continued because to date important conceptual work and intellectual development has already been achieved by this group and will be lost if not managed and disseminated widely throughout the state.

## **Standards**

Important learning about the use of standards for assessment of student achievement has been gained. For teachers this was an important opportunity to engage in standards-referenced assessment. The benefits derived from this experience were many:

“... it has been beneficial to be engaged in marking because it ... provided an overview of the students’ performance on different skill areas which could then inform curriculum development. ... the experience could also encourage teachers to rethink and redevelop their assessment and teaching tasks.”

“Glaring gaps of what should be taught/learnt at Year 9 apparent from the QAT. Middle schooling is the phase where teachers are responsible for teaching two or more subjects”

“... useful for teachers so that we can actually get feedback on how the students are doing and what we need to improve on because there is always going to be something that we need to improve ...”

“... as a teacher I am going to be prepared ... and ...if there are any holes in what we are doing - making sure that the kids are prepared not necessarily as a focused test preparation, like there is in lots of schools for QCS – but in general. It should help improve higher order thinking responses in the general classroom and I just feel that being here I am able to help my school and help lead the other teachers. I believe I am able to report back to the school about the types of questions.”

The implications for curriculum and school development priorities become apparent. Teachers appear to be able to determine from the assessment data provided how they might change their emphases in terms of curriculum priorities and how they might develop their pedagogy to address the gaps identified in student learning.

Teachers were also able to see that some students had achieved particularly well in some of the tasks that were considered high-level and quite sophisticated. As a QAT team member observed;

“... the most important principle I think is that it allows students and schools to get a feeling for state-wide standards ... I mean ...(in) a quality sense. What is the quality of assessment coming through? Not relative to how well the state did but relative to the sorts of standards

that people believe the state is capable of ... Relative to the aspiration ... where the aspiration ... has actually been met by some students, so it's not an unreasonable aspiration. It is about people knowing that here are students ... who can do stuff of this quality and seeing where they are relative to that."

Another QAT team member reflected on the learning for teachers from their experience of being involved in the marking.

"I think ... it's a good opportunity to get teachers engaged with standards-referenced assessment and I think there is a legitimate professional development mode there... As the QCAR framework develops one of the commitments is towards the building of teachers' skills in assessment and I think particularly in the P-10 (school sector) there isn't necessarily a sense that there needs to be the qualitative differences ... specifically sought in relation to student work. It is not good enough to say ... 'basically they are working at level...'  
(Teachers) are writing (reports) of this generic form. I think that there is a very important role in getting teachers used to and educating them in standards-referenced assessment."

When asked specifically about whether teachers were involved in determining the standards it was made clear that this was not the intention. Rather teachers were gaining an understanding of the standards from their marking experience and their assessment of student work. As one of the teacher markers commented:

"Standards come from students' work ... what students can actually do not what you think they can do. The first one is what we are thinking they can do and that's been verified from this whole range of students' work. That's where I say the exemplar material comes in. What you need is actual students' work which illustrates the standard or an application of a marking scheme."

Similarly, other teachers indicated how their understanding of standards developed from their experience of marking the constructed response and this helped them realize the implications for curriculum, teaching and learning at the school level:

"For me as a marker ... it keeps me involved. It gives me great encouragement to see the developments that are taking place. Year 9 has always been a lost year ... some form of monitoring situation of what is actually taking place in schools ... is necessary."

"I think it gives us an indication of how we should change things. ... an opportunity to re-evaluate our curriculum at that middle school level because to me the expectations of the test are vastly different to the outcomes that the curriculum is operating (at) in the schools today... It showed me that there is a bit of a gap between the perceptions of what outcomes we should be providing for our middle school students."

“... as a marker I guess it has shown me that our children are about average. They’re a long way from the bottom and they’re pretty well a long way from the top but they are scattered around the mid-point of the continuum. It will be interesting to see what the final results show.”

“The model response (exemplar) and how you use that ... it became really clear to me ... the use of those model responses (exemplars) for the marking criteria. That was really useful and in terms of how we did it, the training the other day, is how I would conduct some training for teachers to start thinking about how we make a judgement on those responses and what is it going to be?...”

The implication for the QCAR framework is the importance of pitching the tasks at the correct levels for Years 4, 6 & 9. As has been evident from the pilot the student work has helped to inform the development of the standards.

### **Computer-Based Task**

Undoubtedly the most significant development has been the computer-based task. When asked about the key lessons learnt from the trial the majority of responses referred to the computer-based assessment task or some aspect of it. The following was a typical response:

“Most exciting development has been the computer-based task ... this needs to continue.”

The *QAT 2005 Reflections* report states:

“The Queensland Assessment Task is an exploration which combines the more traditional paper-based mode with a foray into the electronic medium requiring a student to ‘take the test’ on the computer reacting to novel stimulation, recording answers electronically while drawing on a wealth of background and supportive material also displayed electronically. This is a test of generic skills applied in year 9, another first in this state, indeed in this country” (EQ, 2005: ii).

The innovative aspects of the computer-based task will be discussed what is highlighted as significant at this point are the developments that relate to the relevance of the electronic medium and multi-literacies, online marking and the cost effectiveness of computer-based assessment.

As one teacher noted:

“Middle school is where students are increasingly becoming disengaged and this can be addressed through the introduction of greater computer use as in the computer-based task.”

Another suggested that:

“For curriculum, assessment, political, pragmatic and financial reasons there is a need to continue with the computer-based assessment. This is, because students engage in it, it can be often marked by the computer, it can be marked on-line when it is not marked by the computer, it’s interactive, it can give kids information and they can actually learn something as part of the process of doing the assessment. It can give instant feedback. ... It doesn’t replace their normal mode of interacting in their life with something that is perhaps less normal.”

It was also stressed that it was possible for students to learn from the computer-based task. When asked if they had learnt anything from completing the tasks, students were more inclined to suggest that they learnt about contour lines from completing the computer-based task whereas those students who did indicate that they had learnt something from the completion of the constructed response referred to the factual content of the questions, sometimes seriously, sometimes flippantly and other times erroneously.

“The computer-based task was also a learning task, the students learn skills but don’t train for a specific task.”

Perhaps the cost effectiveness of the computer-based task was recognised as most significant. To illustrate:

“More cost efficient to include the computer based task because the students are more engaged because of what they do.”

“... it will be cheaper if the computer-based task is followed through ... the children will be more engaged ... because it is what they do now days. That’s how they spend most of their life and they find it easier to use a computer ... than write pen and paper exams.

Online marking was also identified as an important contributor to the cost effectiveness of the computer-based task.

“I think that online marking is definitely the way to go. But I think also that it really needs to ... have more resources so that we are not all stressing out all of the time.”

“Let’s face it; if you have got a computer-based task and you have 4,500 kids doing it, it is not going to cost you anything more to have 15,000 kids doing it. Whereas if you have pen and paper testing even just the printing of the papers costs more. You have also got the issue of how are you going to mark them. If you can mark things on-line you are paying marker time or by script or whatever method you choose ... if you have to have people coming here like this you have got meals, accommodation, travel etc. on top of just the marking. I think the on-line marking is the thing that has really come out of this. We know how to set up a marking guide, to mark papers and we know how much that is going to cost. We haven’t known how much cost is going to be

incurred by the on-line marking until now. All of those scripts were marked over 4 days. That's 4,500 kids 9,000 markings roughly."

## **OPPORTUNITIES**

The many opportunities that have emerged from the 2005 QAT pilot relate to the prospect for innovation, teacher professional development, school-based curriculum, a bank of assessment tasks, reporting of results to the schools, the students and to the system.

### **The Prospect for Innovation**

In the context of the advancement of assessment and testing, Wood (1993) has noted that it is almost impossible to properly scrutinize an innovation and its likely effects before its introduction.

"It does seem that to find out whether innovations actually work or will work, you have to introduce them first..." (ibid: 251).

The 2005 QAT piloted assessment tasks pitched at level.5 and was designed to assess generic skills including the use of ICTs and multi-literacies. Wood has identified some key principles for deciding on the necessity of change. These are useful to analyse the opportunities provided by the 2005 QAT which was essentially an explorative research project.

- Are the changes that are claimed to be innovations actually new?
- Do the innovations actually work? If they actually work, do they work better than what they replace?
- Is there a value in change for its own sake?
- Does the innovation predict better what it is supposed to predict?  
(Wood, 1993)

These questions raise important issues related to the nature and purpose of the 2005 QAT. It was always intended that the QAT would be made up of three standardised assessment tasks in the following different assessment modes:

- Task 1 – interactive and computer-based
- Task 2 – constructed response
- Task 3 – performance-based.

This was to be complemented by the corresponding Teacher Generated Task. Another driving tenet was that the QAT would be intellectually challenging and have connections to the wide world. With the recognition of the importance of the nature of learning and the need to teach learning skills the 2005 QAT does indeed offer an important step towards offering something that is new and of value. As Claxton (2005:2) has indicated it:

"... is actually possible to help young people become better learners – not just in the sense of getting better qualifications, but in real-life terms."

Claxton (ibid) refers to concepts from cognitive science, neuroscience and sociocultural theory to explain that learning is learnable and is very much more complex than schooling has assumed. He predicts that this realisation will help to direct curriculum development and education research in the future. He cautions that 'raising standards' as evidenced through exam results is a limited view of how we should be preparing students for a lifetime of change in the 21<sup>st</sup> century. In analysing the policy implications of his perspective he includes the need to:

- Disseminate good practice in the sense of developing learning-to-learn;
- Review the curriculum to analyse the extent to which the different stages offer a coherent programme for the development of 'learning power';
- "Coach beginning teachers in how to vocalise the processes of learning, to 'learn aloud', and to model effective learning" at initial teacher education level and to use this perspective to inform national teacher training agencies and teaching councils;
- Encourage parents to collaborate with schools in developing their children's 'learning power';
- "Develop new assessment instruments that enable students, their teachers and parents to keep track of their developing learning power so that they can feel a growing sense of achievement, not just in passing tests, but in becoming steadily more resilient, resourceful and reflective in the face of real difficulties". (Claxton, 2005: 4).

It is Claxton's latter recommendation that is useful in explaining the relevance and value of the 2005 QAT. The assessment of generic skills applied in year 9 proved possible and was reported relative to an A-standard performance. The A grade was described in terms of desirable features of the responses in the task.

The implication is that in the statements of essential learnings at key junctures generic skills and attributes will need to be identified **as well as** the domain-specific knowledge, skills, understandings and dispositions. There needs to be coherence across the key junctures and these skills will need to be incorporated into the assessment tasks so that students and teachers understand their value and importance in learning.

### *Assessment of Generic Skills*

What is new and significant is the possibility of reporting processing skills of transforming information and ideas. To illustrate, in addition to exhibition of knowledge:

"A-grade students: ....

Extract information from prose, diagrams, maps and symbolic text; clarify it and transform it to display meaning in multiple media.

Discern patterns and relationships in verbal, pictorial and symbolic text (alone and in combination); make significant decisions and

judgements, operationalise these into accurate representation and products.”

These processing skills relate to learning-to-learn or in Claxton’s terms ‘learning power’. In a teacher’s own words;

“... the QAT ... gives us ... items that gave the children the opportunity to demonstrate their skills, their confidences and the interconnectedness of information. I think that the criteria for those tasks are going to be really useful and I think that if the common assessment task can have that as a starting point I think we are ... headed in the right direction.”

Many teachers emphasised the importance of generic skills involved in transforming ideas and information and the impact of the QAT in raising teacher awareness of how these skills need to be addressed.

“... need for kids to be taught and guided in (skills) like analysis ... synthesis ... evaluation and critique. If there is a focus on the knowledge base only without putting the knowledge to use in some context they may not be teaching the processing.”

In a similar vein another teacher commented:

“... this (2005 QAT) has been a real eye-opener, that we are purely looking at ideas and thought patterns and understanding. We are looking at abilities and literacies and abilities to use information and make a response ...”

### *Impact of QAT*

Given the investigative nature of the project and the timing of the evaluation, during the development and implementation phase, the impact of the 2005 QAT at the school level, after results were reported, was not included in the evaluation design and constitutes a major limitation. It is therefore not possible to provide answers to all of Wood’s questions.

- Does the innovation have a better effect on the teaching or the learning, however we choose to measure or observe this?
- Is the innovation more acceptable to experts (theorists and professionals in teaching and testing) and also to the lay public?
- Is the innovation more efficient, that is, does it give the same or similar results with less effort, or with less time or money devoted to administration and scoring?
- Does the innovation involve less training, few specialists to produce, administer, score and interpret? (Wood, 1993)

From the interview data it is, however, possible to suggest how teachers perceived the impact of the 2005 QAT on teaching and learning, its

acceptability to some experts and the efficiency and effectiveness of the marking operation.

In terms of teaching and learning, teachers suggested that the QAT had motivated a number of them to share their learning and insights with staff for the purposes of reforming current practices. To illustrate:

“Teachers may have heard about learning styles at school but when it ... comes to teaching so that we teach children to organise information and to think, teachers don’t have those skills. ... For secondary teachers to ... focus on ... thinking strategies and organising information is really confronting, as well, ... their own personal literacy. The science teachers and maths teachers, they teach science and maths, they don’t teach ... reading and writing that is not their job. But actually it is their job because science and maths have their own literacy ... the QAT ... demonstrates that need, that pulling that together, the fact that it tested the transformation thinking really pulled that together. ... I feel more confident to lead that (connection of curriculum areas through generic skills).”

#### *Innovative Nature of the QAT*

In terms of whether the innovative nature of the 2005 QAT has been acceptable to experts, it was certainly the case that during the development phase an internationally recognised expert in computer-based assessment was consulted and recommended that the concept of a computer-based task was a good one that extended the fairly common practice of testing in multiple-choice format into new territory.

Teachers too who were interviewed responded positively to the computer-based task indicating their support for innovations, such as this, that enhanced the students’ motivation and that are genuinely intellectually challenging.

Some of the teachers who were involved in the marking suggested that to increase the take-up of ICTs in schools, the system could make more use of the online marking. It was felt that a start had been made with the marking of the computer-based task. Online marking proved acceptable to teachers and some recognised the efficiency and effectiveness of this innovation. For example:

“Online marking has been a valuable lesson that has proven to be cost effective.”

“Teachers have been able to mark student responses more efficiently out of school hours than in school marking.”

The graphical component of the marking scheme, referred to in Chapter Five and elaborated in the *QAT 2005 Reflections* report (EQ, 2005), has been described by the QAT developers and markers as innovative and extremely



important in the context of assessing the generic skills that underpin the construct of processing, and in particular transforming ideas and information.

“For me the graphical component was an important development of trying to find a new way of overcoming difficulties we know in marking response items that are rich - in the sense that they are trying to cope with this interplay of skills - and you want to be able to reward that interplay, not just treat it as the answer is correct or half correct or something ...”

“The marking scheme involved advice from the marking advisor ... (who) was able to clarify any issues that emerged, particularly with the use of the graph to capture the multi-dimensions of the open-ended questions. None of us had used that before. I think it is limiting in the way that we have had to use it but they are trialing it ... There needs to be more clarification and probably better indicators on the graph in how you can divide it up in terms of achievement.”

Other opportunities that emerged include: teacher professional development, possibilities for school-based curriculum development, development of a bank of assessment tasks and reporting of results.

### **Teacher Professional Development**

Teachers and schools valued and gained from participating in the 2005 QAT. Some key areas for support that have been identified for continued development are: teachers' assessment skills and literacy, 'growing an assessment culture', moderation practices, use of ICTs, marking, curriculum development, building student's learning capacity and developing pedagogy.

Teachers involved in the 2005 QAT indicated their enthusiasm for continuing to build their skills in assessment. The initiative entitled "Growing an Assessment Culture" should continue and will be central to supporting teachers in the implementation of the QCAR framework, if it is developmental and maintained.

“There's a need for the status of assessment to be elevated and for an understanding that assessment can make a difference to learning.”

There was some criticism of the cascade model of teacher support and the one-off approach to teacher development:

“I think that the model of professional development is wrong. I think that if teachers aren't doing things in their rooms with their classes with support it just doesn't happen and people go away and that a one day workshop - it's just dead money.”

“... hopefully teachers will become more aware of assessment that may sound somewhat ironic considering the millions of dollars that

have been spent on professional development for teachers to do the assessment and reporting workshops.”

What possibly would be more worthwhile is engaging teachers in action research at the local professional level in collaboration with colleagues from EQ, QSA and universities. Such a model of teacher development that has been successful in the context of developing assessment practice was demonstrated in the research conducted by Black and colleagues (Black, Harrison, Lee, Marshall and Wiliam, 2003). A professional development programme that is ongoing and sustained should parallel the development and trialing phase of the QCAR framework.

Some members of the QAT team described the benefits of being involved in the development of the QAT **and** being involved in teacher professional development programmes for task moderation:

“We have been also giving assessment workshops and things like that over the last year or so. In some ways that diversity has been a real benefit ... If we hadn’t had that experience with the rich tasks and also the experience, particularly with ... growing an assessment culture, then we wouldn’t have had the opportunity to engage with teachers in the classroom and know more about what they are really thinking and doing. Having that opportunity to interact ... don’t get to interact with Year 4 teachers that often.”

During the QCAR framework trial period in 2006 the process of building teachers’ capacity to develop intellectually engaging tasks that are aligned to the essential learnings will commence. Teachers will need to appreciate the value of common assessment tasks and their use in the context of standards-reference assessment as has been realised by teachers involved in the 2005 QAT pilot.

“ ... the marking guide, the training, the professional development this has been a tremendous professional development exercise because they (teachers) are looking at students’ work, not from their own schools and their own little cubby hole. You have a wider range of experience and are seeing how some students or students in general would attack or handle such a response to the stimulus.”

Interviewees who had extensive experience in moderation exercises and training teachers for the marking operation analysed the teacher development benefits, particularly in relation to addressing reliability:

“... reliability comes from a single source explanation of what is embedded in the mark scheme - the meanings and the shared understandings and the rules and restrictions that we all have to apply. Convincing markers that in spite of their gut reaction to some of the rules and the shared understandings that for the convention of the marking operation we’re all going to adopt this. ... it’s different from

what they do at school and they may actually be learning something if they try something completely different.”

Much was gained by teachers from participating in the marker training. For example:

“... clarified assessment and things I had been grappling with. I knew that I had to explain the what and why of assessment. Although teachers set assessment items they don't always understand ... “

“From being involved in the marking I have gained a better understanding of the standards and believe the students at my school are about average.”

### **School-Based Curriculum Development**

Teachers reported the value of their involvement in the marking exercise and the benefits for the students and teachers from their participation in the 2005 QAT pilot. Some expressed the importance in terms of their own learning:

“To involve ourselves in learning, I think is a good thing. It is not an easy thing to do either as a manager of a school to put your school on the line, or as a teacher to always be changing and learning. That is not easy. ... it is a good model for students. Just because we are teachers, we don't know it all, and we try something new ... learning doesn't stop.”

“Lots of ideas came to me during the training. I was thinking to myself here is something I can do. Here is an activity I want to do with teachers. I can link this back to ... when we are designing our integrated units. Here's a way we can put this right in the planning stage. This is what we are planning to do, this is how we are going to assess it, this is what we are going to assess and how we are going to make sure it is taught.”

“How we use that test information to identify gaps in what we are doing and also gaps in our skill base ... if there is a gap in what the children have (achieved) it will identify a gap in what we do ...”

Those teachers from remote or rural schools commented:

“Our children are very insulated in that they don't travel, they don't do lots of things, many of them don't take part in community or other activities ... What they get, they get from us, but we need to make it good and that worries me, because if we don't have those skills our children are going to be seriously disadvantaged.”

“The professional development needs are multi-pronged ... curriculum design and development delivery model that encourages teachers to look at what they are teaching and how they are teaching it. ... thinking

processes that encourage teachers and students to think wider - hopefully communities. That is probably the biggest and hardest one because in many small closed communities this is a real issue.”

“... it’s interesting to read the responses of children who focus very much on facial expressions, they are sad, they are stern, they are cranky ... as well as the transsexual and metro-sexual. They swapped clothes because they are transsexuals. I guess our children wouldn’t have even thought about that.” (This teacher was marking the constructed-response task, section three entitled, American Gothic – Not!)

These comments allude to the importance of the cultural context and can be understood from socio-cultural and situated theories of learning. Such insights help to heighten our awareness of the need for teachers to understand the theoretical underpinnings of the proposed changes to assessment practice. The remote and rural settings of Queensland in which assessment takes place, and curricula are developed and implemented, need to be understood. Too often these rich cultural contexts are neglected by policy developers who give teachers the false impression that the implementation of curriculum and assessment policies can be implemented in a homogeneous manner without considering the disparities that exist. It will be important for those responsible for the development and implementation of the QCAR framework to be informed of the developments in learning theories and how these relate to curriculum development and assessment. Teachers too will need to be kept informed if they are to understand how and why they will need to change their practices.

### **Bank of Assessment Tools**

There are plans for the establishment of an assessment bank for use by teachers for Years 1-10. It is intended that exemplary assessment strategies and tools for teachers will be included together with those strategies and tasks developed by schools. In addition assessment tasks developed by QSA for common assessment against standards by teachers to be undertaken in Years 4, 6 and 9 will include administration, marking and reporting guides and student work samples. Assessment tasks developed by schools, and approved by QSA as of similar quality to the common assessment tasks, will also be made available. Models of schools’ good assessment practices will also be included in the assessment bank (QSA, 2005b).

The resources produced by the QAT team as illustrated in the *QAT 2005 Reflections* report are examples of the quality needed. This is because teachers will respond to principles and practices that they can relate to and that are grounded in their own contexts. Teachers will not change their practices based on research or evaluation evidence alone, they need

“... examples of implementation, by teachers with whom they can identify and from whom they can both derive conviction and confidence

that they can do better, and see concrete examples of what doing better means in practice.” (Black & William, 1998: 15- 16).

Teachers who participated in the 2005 QAT identified the data bank as one of their needs and an opportunity to share good practice:

“... need to set up a unit, a task bank of assessment items and tasks”

“TGTs need some exemplars with model answers that include explanations of how marking criteria are used”

“I think a lot of the stuff there, it would be a shame if it wasn’t released as a teaching tool, some of the contour stuff could be used to teach contours ...”

It will be important to ensure that the materials for the items that are included in the bank of tasks, illustrate an appropriate match to the intended purpose of the task. For example in the 2005 QAT pilot some teachers questioned the selection of stimulus material for the constructed-response task, section three entitled, American Gothic – Not! Teachers wondered whether an appropriate choice for the Year 9 students had been made given the stereotypes associated with the word Gothic which they suggested could have distracted students from the task. As has been suggested due to lack of resources the constructed-response was not trialed.

For a common assessment task to be developed there will need to be a strategic approach adopted as suggested by one of the QAT team members.

“I would make a commitment to setting up a unit ... you can’t just construct individual tests (tasks) , ... you need a task bank, you need someone to nurture the expertise for that task bank and then to grow it before you ever start producing tests. Anyone can ... produce a single test, but can you do it within twelve months for the next one, and the next one and the next one. There has to be a commitment for a lot of structure and the expertise is pretty sparse ... We keep using the same people in Queensland because they are the only people who can do it. Nurturing new talent takes time. So it’s not that it can’t be done, it’s that the commitment for resourcing has to be there.”

## **Reporting of Results**

When asked how results should be reported, teachers and members of the QAT team indicated, that because of the innovative nature of the tasks and the fact that this was a pilot, reporting should be approached with caution.

“I think it needs to be reported carefully ... because if the students haven’t taken it seriously and they have done poorly and you compare them to other people in the state it could leave a negative impression about what they have achieved or haven’t achieved in the exam. As a teacher you don’t trust statistics because they can be manipulated.”

“I wouldn’t like material on individual students to be presented because I think teachers might label them too early. I would do it by ‘of your students the range was this, and the mean was this’ or something of that nature...”

“There is always a tension between the supply of information and knowing it is going to be used well and the supply of information that is going to be used badly. ... they have to think of the consequences. It depends on what you believe you are reporting. If you believe that the QAT is giving a snapshot of a student’s performance over a sample of some things they do, you will probably think differently about the reporting too if you were to believe that the QAT itself gave a comprehensive and full picture of the student’s situation.”

Care was also needed, given the move towards the inclusion of some external assessment in terms of the common assessment tasks of the QAT. Some students may be more prepared than others for this form of assessment. Other factors that impact on the approach taken with reporting relate to:

- The purpose of the assessment (e.g. provide state-wide data on student achievement at Year 9);
- The connection between the assessment task and the curriculum;
- Teachers’, parents/,carers, and students’ expectations.

For these reasons the importance of explaining what the results relate to in the report was emphasized.

“I guess the first thing is to make clear what it is we believe that the result is reporting on. ... secondly is the issue of – should it be individual, should it be schools, should it be statewide? I see this as a policy decision.”

“... QAT, important to report on the task not just the result, the implication for a common assessment task is that it will be important to explain to teachers, parents, carers and students what the result relates to, that is, what it is reporting. This is a policy decision at state level for common assessment tasks and school level for teacher generated tasks.”

“... report the results in the best way so that our community can understand them. ... it needs to be meaningful ...”

“The key purpose of the QCAR framework is to align the syllabus related component of school curriculum (the intended learning outcomes) with the assessment of student achievement and with reporting of student achievement.” (QSA, 2005b: 4) As highlighted throughout this evaluation the standards will be central and will provide the language for bringing together curriculum, assessment and reporting.

In the context of the QCAR framework quality data on student achievement will be provided to schools, teachers, parents and school authorities for improving teaching and learning (QSA, 2005b: 5). This data will be used to report and analyse school performance. By 2008 a common format for reporting student achievements of the Essential Learnings in English, Maths, Science and associated capabilities will be developed and schools be guided in how to report on student achievement to parents twice yearly using a five point results scale.

Some interviewees were aware of the anticipated benefits of system level data:

“Reporting of results allows students and schools to get a feeling of state-wide standards in a quality sense that is relative to the standards and the aspiration where it has been met by some students and where students are relative to that.”

“With the identification of essential learnings there will be a need at the school level to rethink how to restructure the curriculum. If provided with tasks of the quality of the QAT that include the need for students to use their thinking skills and ideas then this will be helpful for teachers.”

“This being a trial, I think any kid who has been involved in this deserves to have a certificate that basically says they did either very well, not bad, average or thank you for your participation. I think it is a bigger picture in terms of what we get out of this trial in terms of reporting.”

Students who participated in the 2005 QAT received a certificate that provided a statement of results. Students received an A-E grade for the computer-based assessment, the paper-based assessment and an overall grade, three grades in all. A verbal descriptor for an A-standard performance on the total QAT was given identifying relevant knowledge, processes and skills involved to achieve this standard. In addition, graphical information was provided illustrating the percentage of the grades awarded based on all task items (overall), computer-based items and paper-based items. Details pertaining to the grading and reporting processes involved have been presented in the *QAT 2005 Reflections* report, (EQ, 2005).

## **CHALLENGES**

If these were the opportunities and the strengths that emerged from the 2005 QAT pilot; what then were the major challenges? These range from issues concerning validity, the trialing of common assessment tasks to addressing the conditions needed for sound implementation.

## Validity

A system for statewide monitoring of student achievement can fulfil the purposes of analyzing trends in performance over time, recognizing patterns of achievement in terms of progress, or lack of progress, and providing information that promotes discussion about the areas of relative strength and weakness across the curriculum.

As has been indicated in this evaluation the nature and format of the tasks of the 2005 QAT were judged by some students and teachers as assessing worthwhile skills and processes. In developing common assessment tasks it will be important that QSA develops tasks that:

- address important aspects of the target domain of the essential learnings;
- are likely to motivate students;
- can act as models for assessment tasks to be developed by classroom teachers.

If a narrow focus is adopted in the common assessment tasks then teachers, parents and the community will focus on the areas assessed rather than the full spectrum of essential learnings that will include: cross-curricular aims, generic skills and attributes and domain-specific knowledge, skills, understandings, dispositions and values.

The QAT was designed as a 'dipstick approach' to assessment in that the students chosen for assessment would provide data on the range of scores achieved for a sample of Year 9 students on tasks that were pitched at Level 5 for the processing construct of transforming ideas and information.

As suggested by Crooks, Kane and Cohen (1996: 283):

"The goals of national monitoring can be satisfactorily achieved by generalising from a small sample of students to a description of the population. This approach has considerable advantages. For a given expenditure, much more comprehensive and rich data can be obtained from a small sample than from a full national population, thus supporting the extrapolation link without weakening the generalisation link."

Crooks et al.(1996) go on to suggest that if a sampling approach is chosen then it will lower the stakes associated with an external assessment or examination programme and limit the possibility of reporting data on student or school performance, thus reducing incentives for teachers to distort the results by 'teaching to the tests'.

In the planning and development of the 2005 QAT a major challenge was for the task developers to address these threats to validity (ibid). In the context of the QCAR framework these threats will also need to be addressed and are summarised here. Crooks' model provides a framework for the validity of assessment uses and interpretations. For the developers of common



assessment tasks this model could help to facilitate efforts to address the threats to validity. It is stressed from the outset that validation can only take place if the intended purposes of the assessment tasks are understood. The appropriateness of the tasks to those purposes will be central to determining the strength of each link in the assessment chain (see figure 1).

Figure 6 Threats to validity: the Eight Stage Model (reversed) (Crooks, Kane & Cohen, 1996)

Link	Threat
<b>8. Impact</b> on the student and other participants arising from the assessment process, interpretations and decisions.	<ul style="list-style-type: none"> <li>• Positive consequences not achieved</li> <li>• Serious negative impact occurs</li> </ul>
<b>7. Decision</b> on actions to be taken in light of the judgements.	<ul style="list-style-type: none"> <li>• Inappropriate standards</li> <li>• Poor pedagogical decisions</li> </ul>
<b>6. Evaluation</b> of the student's performance, forming judgements.	<ul style="list-style-type: none"> <li>• Poor grasp of assessment information and its limitations</li> <li>• Inadequately supported construct interpretation</li> <li>• Biased interpretation or explanation</li> </ul>
<b>5. Extrapolation</b> from the assessed domain to a target domain containing all tasks relevant to the proposed interpretation.	<ul style="list-style-type: none"> <li>• Conditions of assessment too constrained</li> <li>• Parts of the target domain not assessed or given little weight</li> </ul>
<b>4. Generalization</b> from the particular tasks included in a combined score to the whole domain of similar task (the assessed domain).	<ul style="list-style-type: none"> <li>• Conditions of assessments too variable</li> <li>• Inconsistency in scoring criteria for different tasks</li> <li>• Too few tasks</li> </ul>
<b>3. Aggregation</b> of the scores on individual tasks to produce one or more combined scores (total score or subscale scores).	<ul style="list-style-type: none"> <li>• Aggregated tasks too diverse</li> <li>• Inappropriate weights given to different aspects of performance</li> </ul>
<b>2. Scoring</b> of the student's performances on the tasks.	<ul style="list-style-type: none"> <li>• Scoring fails to capture important qualities of task performance</li> <li>• Undue emphasis on some criteria, forms or styles of response</li> <li>• Lack of intra-rater or inter-rater consistency</li> <li>• Scoring too analytic</li> </ul> <p>Scoring too holistic</p>
<b>1. Administration</b> of assessment tasks to the student.	<ul style="list-style-type: none"> <li>• Low motivation</li> <li>• Assessment anxiety</li> <li>• Inappropriate assessment conditions</li> <li>• Task or response not communicated</li> </ul>

These threats to validity will also need to be addressed by teachers in developing teacher generated tasks and should be explained and be included in the professional development programme.

### The Need for a Trial

The importance of trialing the assessment tasks prior to administration has been acknowledged by many teachers and developers. Despite the expertise on the team there is always a need to trial the assessment tasks. In the case of the 2005 QAT the timeline was too short so there was no trial of the

questions. This impacted on the validity of some questions. If a common task is to be developed then it must be trialed with students prior to administration. This includes students with special needs.

Such acknowledgements of this principle are illustrated in these comments:

“You would hope that there would be a trial ... because of the background of the people involved (in the development of the constructed –response) and their experiences in testing ... fairly confident that we could get things to a pretty good standard. Trialing would have just taken that guesswork out of it. We would have known. ... At least the trial for the computer-based one gave us pretty good information. ... confidence to go ahead ... open to disaster if we didn't know ... about whether the machines in schools could handle that sort of software and whether kids would understand the instructions and ... how to play with a paint package before they got to the real thing - being able to use the facilities of the software available really.”

“We had a very short time line and a small group of people and it would have been a lovely luxury to have more time for panelling final items, knowing what we know now about what students have done with questions. It would have been nice to have trialed them but we didn't have that luxury. Some of the ways that they handled questions had to be picked up in modifications to marking schemes, so it would have been lovely to trial things. Trialing costs money.”

“We were unable to see how well the marking guides were able to stand up because of the lack of funding for a trial.”

“A trial is needed to make sure that the wording conveys what is intended, the presentation does not lead to confusion and invalid responses and that the setting out of the paper is not too complicated and can be easily followed.”

“Make sure a wider group is involved in the selection of the items for the constructed response rather than just the developers. Constructed response for the QAT seemed to be too in-house. There is a need for teachers across the state to be involved in the development of the items and the selection of materials.”

“The limited funding prevented a trial that meant that modifications to marking schemes couldn't be carried out prior to administration. If the common assessment task is going to be a state-wide assessment then it will need to be trialed.”

“Give the paper to teachers of Year 9 students and get teacher feedback on the appropriateness in terms of level, the language and whether it is appropriate for Year 9, whether the stimulus materials are

something that Year 9s would be interested in and is something that they have focused on.”

### **Conditions for Implementing Assessment Programmes**

As explained in the methodology of this evaluation while the QATs were never designed to fulfil a high-stakes role, in the context of the QCAR framework and the plans for statewide assessment of the essential learnings, it has been important to consider the essential conditions to sound implementation of high-stakes educational assessment programmes. These include:

- protection against high-stakes decisions based on a single test;
- adequate resources and opportunity to learn;
- validation for each separate intended use of the high-stakes assessment;
- alignment between the assessment and the curriculum;
- the validity of the passing scores and achievement levels;
- appropriate attention to students with disabilities;
- appropriate attention to language differences among examinees;
- opportunities for meaningful remediation for examinees who fail high stakes assessments;
- careful adherence to explicit rules for determining which students are to be tested;
- sufficient reliability for each intended use and
- ongoing evaluation of intended and unintended effects of high-stakes testing (AERA, 2000: 2-5).

The latter four conditions cannot be evaluated because the data collection phase for the evaluation stopped prior to the distribution of the reports (QAT statement of results). However, the former seven conditions will now be discussed in relation to the 2005 QAT and the implications for action in the context of the QCAR framework.

The concept of a Queensland Assessment Task that incorporates a constructed-response (paper and pencil test), a computer-based task (featuring a range of response items – short answer, pictorial, diagrammatic) and a performance-based task complemented by a teacher-generated task constitutes a comprehensive assessment system that can enable more valid assessments of student achievement. Assessments based on a thoughtful process, grounded in multifaceted body of evidence are more valid and can lead to continuing improvement in teaching and learning. This was a major challenge for the QAT team given the changing policy context and shortage of resources. However, such a comprehensive approach to assessment should be sustained to prevent high-stakes decisions being made on the basis of a single test.

The 2005 QAT pilot was under-resourced and therefore focused on the development of the innovative and futuristic aspects of the computer-based task. This provided valuable opportunities for learning. What is important in

the context of the QCAR framework is that adequate resources and the opportunities for learning are provided and sustained.

Validation for each separate use of the common assessment tasks can only take place if the intended purposes of those tasks are understood. As is apparent from the 2005 QAT experience not all teachers and students were clear about the purposes of the assessment tasks nor were they certain about how the results were to be reported and used. It will be necessary in the QCAR framework context to provide teachers, students, parents and carers with a clear understanding of the intended uses of the various tasks. This is an important consideration for the improvement of teaching and learning.

What has become clear throughout the evaluation has been the lack of alignment between the proposed approach to assessment and the current curriculum. As many teachers interviewed for this evaluation have indicated a clearer understanding of what students need to know and be able to do, and what they should be given the opportunity to learn, will help them focus their assessment practices. Some confusion exists regarding the alignment between assessment and the curriculum. The plans outlined in the *Technical Paper Queensland Curriculum, Assessment and Reporting Framework, Stage 1* (QSA, 2005b) details how alignment aims to be achieved.

The validity of the passing scores and the achievement levels for the 2005 QAT have been addressed in the *QAT 2005 Reflections* report, (EQ, 2005) and will be referred to here in summary. The marking of the QAT involved the use of computer technology and markers who were teachers. The computer-based task required the use of two different kinds of marking processes. Four of the six questions were scored by computer algorithms and the other two questions were marked via a website by teachers across the state. A 'central venue process' was undertaken for the marking of the constructed response. The nature of the constructed response items allowed for a wide variety of responses. In the marking process this variability is regulated by assigning one of a limited set of meaningful grades to each response. A student's response to each question was marked a minimum of two times and, if necessary, three times using an acceptable minimum standards marking scheme specific to the question. Markers were monitored during the marking operation (ibid: 2).

Conditions that were not met by the QAT developers were the attention to students with disabilities and language differences. The resource implications need to be addressed by the developers of the QCAR framework because as is evident from the experience of the QAT although these needs were identified and planned for the funding and time constraints prohibited the development and implementation of the support for the students with special needs which includes language differences.

## IMPLICATIONS FOR ACTION

Many of the implications for action have been emphasised throughout the evaluation but some important priorities are summarised here in the conclusion.

### *Alignment of the curriculum and assessment*

Inform teachers of the essential learnings and the standards this is an immediate priority. A clearer understanding of what students need to know and be able to do, and what they should be given the opportunity to learn, will help teachers focus their assessment practices. The standards will provide a common frame of reference for making judgments about the quality of student work **and** the progress of student learning, while providing a common language for reporting.

### *Validity and reliability*

Address the threats to validity and reliability. Assessments based on a thoughtful process, grounded in a multifaceted body of evidence are more valid and can lead to continuing improvement in teaching and learning. Such a comprehensive approach to assessment prevents high-stakes decisions being made on the basis of a single test. The validity of the passing scores and the achievement levels of common assessment tasks need to be addressed.

If results are to be used to compare schools or districts or if changes in results are to be monitored over time policy will need to make explicit which students are to be assessed and under what circumstances students may be exempted from the assessment. The policies need to be implemented consistently across the state to assure the validity of the comparisons of the results. In addition, the reporting of the assessment results should accurately portray the percentage of students who are exempted.

Establish processes to develop consistency and comparability of teacher judgments at the local and district levels. With the emphasis on standards as statements that indicate different levels of quality of performance, teachers will need to meet regularly to discuss work for moderation and reliability purposes. This will involve teacher discussions focused on the exemplification of standards in student work. Sufficient reliability will be required for each intended use of the results. The accuracy of the results will need to be examined to evaluate if they support each intended interpretation. Teachers, principals, students, parents and carers need to know the intended uses of the common assessment tasks; an important consideration for the improvement of teaching and learning.

### *Teacher support*

Current teacher support strategies will need to be maintained. Provide adequate resources and opportunities for teacher development and valuable learning so that innovative research and development can progress. Teachers will need to understand the changes in policy directions, the characteristics of the QCAR framework, the reasons why such changes are

being introduced, the quality and practicality of these changes. To help teachers prepare for implementation there will need to be clarity about the key issues that they will confront and some understanding of the complexity of these issues. Priorities for professional development include the use of standards, addressing threats to validity, teacher generated tasks, moderation, marking, reporting and the use of assessment data to inform teaching and learning. Encourage the growth of teacher professional learning communities both within and across schools. Provide a bank of assessment tasks that includes evidence of student work illustrative of the standard achieved.

#### *Equity and fairness*

Give appropriate attention to students with special needs and language differences. Where the quality of student work or progress is lacking provide opportunities for meaningful remediation. Provide explicit rules for determining which students are to be assessed to ensure careful adherence.

#### *Strategic leadership and implementation plan*

Provide a coherent implementation plan for the QCAR framework and disseminate widely to the education community. Encourage the development of a dedicated team approach in districts and schools to support and monitor the implementation at the local professional level. At the central level nurture new talent to ensure there is sufficient capacity and to provide the necessary infrastructure for implementation during the trial phase.

To help establish coherence across the policy making arena reinstate the work of the Assessment and Reporting Framework Implementation Committee or a similar reference committee to provide guidance and support throughout the implementation and to inform policy leadership. This will also help to maintain 'organisational memory' and capitalize on the wealth of conceptual and intellectual development achieved to date.

#### *Ongoing evaluation, research and development*

The 2005 QAT pilot focused on the development of the innovative and futuristic aspects of the computer-based task the research, gains achieved need to be sustained and progressed. Engage in ongoing evaluation of intended and unintended effects of the QCAR framework, in particular the common assessment tasks. Collaborate with schools, districts and academics for research and development purposes. Provide funding for research to continue. Disseminate the findings which will include those about the intended and unintended consequences of the changes implemented.

### **Conclusion**

A major component of successful change is ongoing inquiry and reflection. This will require collaborative research projects with QSA, EQ and universities. Reforms such as the implementation of QCAR Framework can benefit from 'a critical friend' to evaluate and to monitor the various phases of initiation, development, implementation and evaluation of the changes. Some possible areas for collaborative research and development include:

- *Equity and Fairness Issues*  
With the release of the 'Disability Standards for Education 2005' it will be important to ensure that the QCAR framework incorporates these standards. Research and development will help to identify how best to adapt assessment to support the student with special needs.
- *Moderation and Teacher Judgment*  
It will be important to explore the ways in which teachers make judgments using standards-referenced assessment in order to inform policy about how these judgments ought to be made.
- *Ongoing Monitoring and Evaluation*  
As was apparent in the evaluation of the 2005 QAT important insights and understandings of the intended and unintended consequences of policy change need to be understood to provide strategic policy leadership.
- *Teacher Support Strategies.*  
The development of professional learning communities for the implementation and evaluation of the changes introduced by the QCAR framework should be researched to ascertain the level of resources and the policy support required.
- *Computer-based Assessment*  
Important developments have been achieved in the development of computer-based assessment for the 2005 QAT and should be continued. Important lessons were learnt for future computer based assessments and emerging technologies have the capability to reduce and simplify the deployment to an insignificant cost. This potential should be harnessed and explored for possible policy implications.

The 2005 QAT has the potential to alert teachers, principals, parents, carers and the community to the important skills of learning. It also constitutes an application of multi-modal assessment with computer-based tasks, constructed response, performance and teacher-generated-assessment. It will be of interest and use to those who were not involved in the pilot as it has provided important insights into other exciting dimensions of the assessment arena. That is, the assessment of key generic skills and the use of computer-based assessment. These findings are useful to inform the implementation of the QCAR framework and to illustrate stimulating innovative steps for possible directions in the development of assessment in the state of Queensland.