# An Evaluation of System-wide Assessment of Problem Solving at Year 12 by Report and Related Test

*Barry McCrae*

University of Melbourne

*In 1994, nearly 5000 Victorian Certificate of Education students undertook a new, two-component Problem solving CAT (Common Assessment Task). The students had two weeks to prepare a written report on the solution of a problem and five days later had to sit for a related test. Both the problem and the test were centrally set and school assessed. This study reports the findings of an external evaluation of the CAT.*

## Introduction

The Victorian Certificate of Education (VCE) is awarded to students who satisfactorily complete the final two years (years 11 and 12) of secondary schooling. In order to gain credit for a one-semester unit within the VCE, students must satisfactorily complete each work requirement in the unit. Each unit of VCE Mathematics has *problem-solving and modelling* as a work requirement; this is defined in the Study Design as 'the creative application of mathematical skills and knowledge to solve problems in unfamiliar situations, including real-life situations' (Board of Studies, 1994, p. 7).

The assessment of levels of performance in year 12 VCE subjects (i.e. unit 3 and 4 sequences) is via a number of Common Assessment Tasks (CATs) directly linked to the work requirements. For the first four years of the VCE, from 1989 until 1992, one of the CATs in each of the Mathematics subjects was a *Challenging problem*. Students were given two weeks to solve one of three or four centrally-set problems and to prepare a written report of approximately 1000 words on their solution. Schools graded their own students' reports, according to a common set of criteria, but these assessments were subject to a statewide verification process.

The Challenging problem CAT was suspended at the end of 1992 because of growing concerns that many students were being unfairly advantaged in the preparation of their reports by assistance from parents, tutors or friends, even if this assistance was acknowledged in the reports. The credibility of CATs such as this which are not done under test conditions depends to a large extent on whether a student's work can be *authenticated* with confidence. As discussed in detail elsewhere (Stephens and McCrae, 1995), this not only means attesting that the student is the *author* of the work but also that the student *understands* what is in the report.

During 1993, a two-component *Problem solving* CAT was trialled at year 11 in selected volunteer schools as a possible successor to the Challenging problem CAT. In the first part of the task, students were required to prepare a report on a challenging problem over a period of ten days. The second part consisted of a test, conducted shortly after the due date for completion of the report, which required the students to solve a related, but not identical, problem. The rationale for the test was that it would provide evidence of the authenticity of students' reports, since it would show whether the students understood the mathematics they used in solving the original problem.

Following the success of this trial, it was recommended that a Problem solving CAT, consisting of a report and a related test, should be one of the three CATs in the highest-level year 12 mathematics subject, *Specialist Mathematics* (see Stephens, 1994). Accordingly, for their CAT 1 in 1994, approximately 5000 *Specialist Mathematics* students were

required to attempt to solve one of three problems set by the Board of Studies, the government body responsible for the VCE, and to prepare a written report of between 800 and 1200 words on their solution. The report had to be completed within a designated two-week period and five days later the students had to do a one hour test. Three tests were prepared, one relating to each problem, and students had to attempt the test that corresponded to the problem they had tackled.

The report and the test were both graded by the school using guidelines provided by the Board of Studies. The final grade was obtained by combining the report and the test grades in a 60:40 ratio. If the grade of the report was much higher than the grade on the test, the Board required the school to interview the student to review the authenticity of the student's report. If the student could not convince the interview panel that he/she understood the content of the report, the grade of the report was reduced to the grade of the test. Disciplinary procedures were implemented if it appeared that the student was not the author of the report. No further action was taken if the grade on the test was higher than the grade for the report.

The administrative arrangements and timeline for *Specialist Mathematics* CAT 1 were as follows:

*Monday 25 July*

Cat 1: Problem solving task *Student booklet* distributed.

*Friday 5 August*

Completion date for written report; *Solution notes* sent to teachers, but not to be discussed with students before test.

*Wednesday 10 August*

CAT 1: Test; *Marking scheme* sent to teachers.

*Friday 12 August*

Schools notified of acceptable discrepancies between report and test scores; interviews to be conducted in all other cases as soon as possible.

*Friday 26 August*

Deadline for entering school marks for report and test into Board of Studies computer.

At the end of this period, the Department of Science and Mathematics Education of the University of Melbourne conducted an evaluation of the CAT with the support of the Board of Studies. The evaluation was carried out through the use of a questionnaire sent to all of the 24 schools which participated in the 1993 trial of the new assessment task in Year 11, and to another 108 schools selected at random from the approximately four hundred schools teaching *Specialist Mathematics*. Completed questionnaires were received from 15 of the first group of schools (62.5%) and from 65 of the second group (60.2%): an overall response rate of 60.6% involving 953 of the 4963 students who completed the CAT (19.2%).

The questionnaire contained 39 questions covering eight areas: background information, problem attributes, report structure, test attributes, comparing report and test grades, interviews, authentication issues and other issues. Most questions required respondents to select one of five responses representing positions along a continuum of possible opinions (e.g. strongly disagree, disagree, not sure, agree, strongly agree). In addition to the 39 questions, common incorrect answers to some of the test questions were given and the respondents were asked to mark those solutions which corresponded to questions taken from the tests attempted by their students. There was also space at the end of the questionnaire to add comments and 62 of the 80 respondents (65%) commented further on at least one aspect of the CAT.

## Findings

The evaluation revealed general support for the problem formats of a sequence of

closed questions with 77.9% of respondents indicating that this should be a feature of future Problem solving CATs.. There was general agreement amongst the teachers that such problems are more accessible and less time-consuming for students, and easier to assess reliably, than the more open problems that were a feature of the former *Challenging problem* CATs. On the other hand, 48.7% of the respondents agreed that the closed problem format *reduced* the task's validity as a measure of problem solving ability, with 30.8% disagreeing and 20.5% undecided. Further, 46.7% of respondents did *not* agree that the format increased their confidence in the authenticity of their students' solutions and 22.1% were not sure.

One of the problems (Problem 2—Gaussian integers) can be classified as a 'pure mathematics' exercise, whereas both of the alternative problems involve practical applications of mathematics. (The problems and the tests, together with sample solutions, have been published with the other two *Specialist Mathematics* CATs as a resource book (Board of Studies, 1995c) for this year's students.) The 'pure maths' problem was regarded by respondents as the least suitable problem for students to demonstrate their *problem solving* abilities (as defined in the Study Design) and it was attempted by only 138 (2.8%) of the students. However, respondents indicated that students were less likely to have previously encountered similar problems to this one than for the two practical problems and the type of mathematical strategies required by students to solve it were regarded as much more likely to be 'creative' (compared with 'routine')!

Each of the three CAT problems was preceded by a list of mathematical techniques 'which might be required for this task'. Immediately following the list for each problem was the highlighted statement: 'While other prescribed methods are acceptable, the above are considered particularly appropriate, and will feature in the test which will follow this task'. However, despite the warning inherent in this statement, 91% of the questionnaire respondents expected that, compared with the corresponding problem, the test questions would be sufficiently similar to the problem questions to be able to be answered by each student *using the techniques he/she used in solving the problem.* Further, 50% of these teachers (i.e. 45.6% of respondents overall) expected the test to be set in the same context as the corresponding problem. No doubt teachers were influenced in their beliefs by previous descriptions of the test as a 'transfer task' (Stephens, 1994, p.14) and by the fact that the problem and the test had the same context in the sample CAT 1 distributed earlier in the year by the Board of Studies.

It turned out, however, that only test 2 satisfied these expectations. Tests 1 and 3 were both set in a different context to the corresponding problem and, depending on the approach taken to solving the problem, the test questions possibly required students to use techniques in the list that they personally did not use in their solution of the problem. Not surprisingly, therefore, tests 1 and 3 were frequently criticised: Test 1 for specifying that a particular technique be used whereas an alternative approach included in the designated list of techniques was otherwise just as appropriate (question 2), and for testing work not done by students in the corresponding problem (question 3); test 3 for effectively requiring the algebraic solution of trigonometric equations whereas problem 3 had required the numerical solution of equations and the technology used by most students (graphics calculators or computer spreadsheets) was not permitted in the test. Further, the changing of contexts for these two tests was described as a significant disadvantage to ESL (English as a Second Language) students. Test 3

was also criticised for the interrelation of its parts, especially since (unlike tests 1 and 2) the questions contained no inbuilt answers.

**Table 1:** *Specialist Mathematics* 1994 CAT 1 Statistics

| | *No of candidates* | *Maximum possible mark* | *Median mark* | *Correlation: report/test* |
|---|---|---|---|---|
| *Task 1* | 1674 | | | 0.65 |
| Report | | 30 | 26.5 | |
| Test | | 20 | 15 | |
| *Task 2* | 138 | | | 0.83 |
| Report | | 30 | 27 | |
| Test | | 20 | 15.5 | |
| *Task 3* | 3151 | | | 0.74 |
| Report | | 30 | 26.5 | |
| Test | | 20 | 13.5 | |
| | 4963 | | | |

The closer correspondence between problem and test for task 2 is also probably partly responsible for the much higher overall report/test correlation in this case as shown in Table 1. (It is likely that mainly only very good students attempted task 2 and this would also be a contributing factor to the high correlation.) The Board of Studies provided the overall statistics contained in Table 1. Responses to the questionnaire showed that the correlation between report and test marks for tasks 1 and 3 varied widely between classes, ranging from 0.09 to 0.91 for task 1 and from 0.07 to 0.90 for task 3. However, many of these calculations were based on very small group sizes. None of the 15 respondents who taught students who attempted task 2 were dissatisfied with the correspondence between their students' report and test marks and 13 (86.7%) were satisfied. For task 1, 12 out of 58 (20.7%) were dissatisfied and 35 (60.3%) were satisfied; for task 3, 27 out of 71 (38.0%) were dissatisfied and 23 (32.4%) were satisfied. Respondents were almost equally divided as to whether the combined report and test mark (60:40) was a more valid problem solving assessment for their students than the report mark alone.

Respondents found it easy to apply the ten criteria used to assess students' reports, though the relevance of some of the criteria was occasionally questioned and the solution notes were criticised for not relating key stages in the solutions to the criteria. A feature of Table 1 is the high median mark for each report, indicating a clustering of report marks towards the upper end of the 30-mark scale. For 1995, in order to achieve better discrimination in report grading at the upper end, the previous 4-point scale (High=3, Medium=2, Low=1, Not Shown=0) on 10 criteria has been replaced by a 6-point scale (Very High=5, High=4, Medium=3, Low=2, Very Low=1, Not Shown=0) and guidance has been provided as to when it is appropriate to award Very High, Medium and Very Low on each of the criteria.

For both tests 1 and 3, a significant minority of respondents (17.2% and 24.6%) had trouble at least 50 % of the time in applying the marking scheme provided by the Board of Studies, and even more (19.0% and 32.4%) believed that the scheme was not sufficiently detailed. There was confusion as to whether half marks could be awarded and about the degree of accuracy (number of decimal places) that was required for various answers. Most difficulty arose, however, in determining how many marks to award when a significant mistake was made but subsequent working was otherwise correct. In these situations,

and in the absence of further guidance, most respondents apparently followed the marking scheme 'to the letter' whereas others used their 'professional judgement' and were more generous in their allocation of marks.

This inconsistency is exemplified by the distribution of marks awarded by the questionnaire respondents to the sample solution provided to questions 4 and 5 from test 3. Question 3 required the students to *differentiate* to determine a pair of formulas for use in questions 4 and 5, but in the sample solution incorrect formulas were obtained in question 3 by *antidifferentiation* and these were then used (with some ingenuity, but generally correctly) to answer questions 4 and 5. Considering that the corresponding problem had involved antidifferentiation in its formulation, I believe that an experienced external assessor would have awarded 3 marks out of 5 for question 4 and at least 3 out of 4 for question 5. As can be seen in Table 2, both questions attracted the full range of available marks from the 65 respondents who marked them, with the average marks awarded being on the low side at 2.1 and 1.6 respectively. Table 2 shows that the average marks awarded to the other sample solutions were more in keeping with 'expert' opinion, but it is of concern that in each case some respondents awarded full marks to what were clearly incorrect solutions.

**Table 2:** Sample Solution Marks

| | *Test 1 Q2* | *Test 2 Q6* | *Test 2 Q7* | *Test 3 Q3* | *Test 3 Q4* | *Test 3 Q5* |
|---|---|---|---|---|---|---|
| *Maximum mark available* | 10 | 5 | 5 | 2 | 5 | 4 |
| *'Expert' assessor's mark* | 7 | 4 | 4–5 | 0 | 3 | 3 |
| *Respondents' marks* | | | | | | |
| Number | 51 | 13 | 13 | 65 | 65 | 65 |
| Minimum | 4 | 2 | 2 | 0 | 0 | 0 |
| Maximum | 10 | 5 | 5 | 2 | 5 | 4 |
| Mode | 8 | 3 | 5 | 0 | 2 | 0 |
| Median | 7 | 3 | 4 | 0 | 2 | 2 |
| Mean | 7.1 | 3.2 | 4.2 | 0.1 | 2.1 | 1.6 |
| Standard deviation | 1.41 | 0.90 | 0.93 | 0.37 | 0.97 | 1.34 |

More than three-quarters (76.6%) of the respondents felt that the 5 day gap between the deadline for the report and the test was appropriate, with 19.5% advocating a shorter gap of between 0 and 3 (but usually 2) days. The common, 'sliding-scale' gap allowed between the report and test marks before an interview was required was regarded as appropriate for task 2 by all respondents and by a clear majority of respondents (78.6% and 72.5%) for tasks 1 and 3. Exactly half (40) of the schools who responded to the survey had to conduct interviews: 20 of them for task 1 (representing 33.9% of the schools that had students who tackled this task) and 29 for task 3 (41.4%), with 9 schools having to interview for both tasks 1 and 3. A total of 28 out of 378 task 1 students (7.4%) and 56 out of 548 task 3 students (10.2%) were interviewed in the respondent schools. One school had to interview 7 of its 26 task 3 students (26.9%) and its sole other (task 1) student. No task 2 students had to be interviewed in the respondent schools.

Only two students, both task 1 students from the same school, did not have their report grade confirmed as a result of their interview. The teacher of these two students, like all but 3 of the respondents, was satisfied with the outcome of the interviews he/she conducted. However, a number of respondents commented or inferred that many of the interviews were caused by the lack of correspondence between the problem and the test, especially for task 3, rather than any

real evidence questioning the authenticity of the student's report.

Respondents overwhelmingly agreed that the inclusion of a related test in the CAT, and the introduction of the interview procedure for reviewing authenticity, improved both the public credibility of the CAT and its credibility amongst Year 12 students in comparison with the former Challenging problem CAT. However, there was no consensus as to whether the new arrangements will produce a more valid assessment of problem solving ability (34.6% agree, 14.1% unsure, 51.3% disagree), reduce the likelihood of cheating (28.2%, 21.8%, 50.0%) or assist in identifying instances of cheating (26.9%, 28.2%, 44.9%). The requirement that students must submit with their final report any draft material and a bound logbook containing all working notes, was considered much more important than the test/interview process in enabling teachers to feel confident about authenticating students' reports. This latter finding suggests, though, that teachers may have equated authentication predominantly with 'authorship' and insufficiently with 'understanding'.

## Discussion

As a result of the evaluation, the following recommendations were made to the Board of Studies concerning the conduct of *Specialist Mathematics* CAT 1 in the future:

- Each problem should involve some opportunity to *generalise* solutions to improve the validity of the problems as measures of problem solving ability and allow better discrimination between students.
- The requirements of the report, its word limit and the assessment criteria, should be reviewed in the light of the difference in format between the 1994 problems and the 'challenging problems' of former years.
- The solution notes should relate key stages in the solutions to the criteria used in assessing the report.
- The actual purpose of the test, and its degree of relationship to the corresponding problem, should be made clear to all concerned (i.e. students, teachers and the setting panel).
- It is important to ensure that the tests are seen to be of comparable difficulty, have a similar amount of inbuilt 'help' and are fair to ESL students.
- Careful consideration needs to be given to the implications for the test of setting a problem that encourages or requires a high amount of computer usage.
- More guidance needs to be provided to teachers on the application of the test marking schemes.
- A clear statement should be published concerning the purpose of the interview that is required for those students whose test mark is much worse than their report mark.

The purpose of the interview has been clarified in the Board of Studies' (1995a) documentation for the 1995 *Specialist Mathematics* CAT 1. This includes the following rationale for interviews:

> *Some students may not perform as well on the test as their reports might have indicated. This can happen for a number of reasons, and there is no automatic assumption that students are at fault if this occurs.*
>
> *Students will be given a second opportunity to display their understanding of the problem they have solved in the form of an interview. During this interview a student will be asked a series of questions which will enable him or her to display his or her understanding of the problem. (p. 27)*

This makes it clear that the main purpose of the interview, like the test, is to ascertain the student's *understanding* of their report, not just its *authorship*.

The relationship between the two components of the CAT, the problem and the test, is discussed in detail in Stephens and McCrae (1995, p.14). It is argued there that, since the test's credibility as a measure of report authenticity relies on it being primarily a transfer task, the test should be designed so that a student can answer the questions using the techniques *that he/she employed to solve the problem*. If it is necessary for students to use specific techniques in the test that they did not use in solving the problem, then not only is the test's validity as an authenticity measure brought into question, so is the policy of combining the report and test marks to obtain a meaningful assessment of problem solving ability.

The fact that the test mark contributes *at all* to the final grade for the CAT was queried by a number of the questionnaire respondents and has recently been challenged by the mathematics teachers' association (Mathematical Association of Victoria, 1995). The decision to combine the report and test marks in a 60:40 ratio appears to have been a compromise negotiated with the Board of Studies which had to be convinced that the proposed Problem solving CAT would not attract the public criticism (regarding undue assistance to students) that had dogged the former Challenging problem CAT during its brief lifetime. The allocation of a 40% weighting to the test effectively reduces the amount of true school assessment in *Specialist Mathematics* to 20%, compared to one-third in most other year 12 VCE subjects (including the other two mathematics subjects). Like all (completely) school-assessed CATs, the combined mark is subject to the Board of Studies' (1995b) statistically-based review procedures for school-assessed CATs.

Of more fundamental concern, though, is the constraining effect that the existence of the test requirement has on the degree of 'openness' that the problem can possess. Instead of a Problem solving CAT, students in the other two year 12 mathematics subjects do an *Investigative project* CAT—a school-assessed project based on a centrally-prescribed theme and done over a designated four-week period. Stacey (in press) observes that many compromises were made to arrive at a workable model 'but the end result is still an authentic [open problem-solving] task'. In my opinion, the problems in the 1994 *Specialist Mathematics* CAT 1 were too 'closed' to enable it to be regarded as an authentic assessment of problem solving. To improve the CAT's credentials, future problems should at least allow some opportunity for generalising solutions. The clear preference of teachers for the continuance in the future of the 1994 problem format suggests that concerns for their students' results outweighs their commitment to the valid assessment of problem solving.

## Acknowledgement

The expert assistance of Margaret Kendal in preparing the questionnaire and analysing the responses is gratefully acknowledged.

**References**

Board of Studies. (1994). *Mathematics study design*. Carlton, Victoria: Author.

Board of Studies. (1995a). *VCE administrative handbook 1995: Part 2. School assessment*. Carlton, Victoria: Author.

Board of Studies. (1995b, March). Review procedures for school-assessed CATs [GAT Special Issue]. *VCE Bulletin*, pp. 2–3.

Board of Studies. (1995c). *VCE official sample CATs 1995. Mathematics: Specialist*. North Blackburn, Victoria: HarperSchools.

Mathematical Association of Victoria. (1995, April). Policy Committee. *Common Denominator*, p.5.

Stacey, K. (in press). The challenge of keeping open problem-solving open in school mathematics. *Zentralblatt für Didaktik der Mathematik*.

Stephens, M. (1994). Developing a problem solving assessment task in the VCE. *Vinculum, 31*(1), 7–17. Melbourne: Mathematical Association of Victoria.

Stephens, M. and McCrae, B. (1995). *Assessing problem solving in a school system: Principles to practice* (Preprint No. 2/95). Melbourne: Department of Science and Mathematics Education, University of Melbourne.