School-based assessment in VCE Mathematics: Ten years on

Barry McCrae, The University of Melbourne
John Dowsey, The University of Melbourne
Max Stephens, Victorian Board of Studies

Since the introduction of the Victorian Certificate of Education in 1989, assessment in VCE mathematics subjects has included school-based tasks undertaken over an extended period of time. This paper reports on an evaluation of the latest of many changes that have been made to these tasks since their introduction and examines whether they have retained their validity for assessing problem solving and investigative project work. Student use of powerful technology is identified as an emerging issue in their future conduct.

History of Change

Since the introduction of the Victorian Certificate of Education (VCE) in Victorian secondary schools in 1989, assessment in VCE mathematics subjects has comprised a mix of school-based assessment tasks undertaken over an extended period of time and traditional end-of-year examinations. Two types of school-based Common Assessment Tasks (CATs) have been used: an investigative project CAT and a problem solving CAT.

The history of the problem solving CAT is reviewed in McCrae and Stacey (1997). Until 1992, students were given two weeks to solve one of three or four centrally-set 'challenging problems' and to prepare a written report of up to 1000 words describing their solution. Because of concerns over the authentication of students' reports (Stephens and McCrae, 1995) the *Challenging problem* CAT was suspended for 1993 and replaced in 1994 by a two-component *Problem-solving* CAT. The first component of the new CAT is similar in nature to the original CAT, requiring students to prepare an 800–1200 word report over a two-week period on one of three centrally-set problems. However, the problems are more structured than the earlier 'challenging problems', each one consisting of a sequence of specific questions about a particular situation. Hence, it is clearer to both teachers and students what is required and so it is possible, for each problem, to set a test of the mathematical techniques used in its solution: this is the second component of the *Problem-solving* CAT.

The one-hour test is conducted a few days after the report is due and has a similar, but not necessarily identical, context to the corresponding problem. It contributes 40% to the CAT grade and provides evidence of the authenticity of students' reports. If a student's mark on the test is significantly lower than would be expected according to his/her mark on the report, the student is interviewed about the report by a panel of teachers. If the student cannot convince the panel that he/she fully *understands* his/her report, the report grade is reduced to that of the test. Students who are found not to have *authored* their reports are disciplined.

McCrae (1995) conducted an evaluation of the first (1994) implementation of the *Problem-solving* CAT. He found clear support among the teachers involved for the continuation of the new problem format. There was general agreement that problems structured in this way are accessible to more students and less time-consuming for students, and that they are easier to assess reliably than the more open problems characteristic of the former *Challenging problem* CAT. However, the teachers tended to disagree with the suggestion that the closed format increased their confidence in the authenticity of their students' reports and tended to agree that it reduced the task's validity as a measure of problem solving ability.

The teachers were firmly of the opinion that the inclusion of the test/interview process improved both the public credibility of the CAT and its credibility amongst the students concerned. They were unsure, though, whether it enabled a more valid assessment of problem solving ability. The requirements that students must submit with their report any draft material and a bound log book containing all working notes, were considered much more important factors than the test/interview process in allowing teachers to feel confident about authenticating their students' reports.

McCrae recommended that in future each problem include a question requiring some aspect of the solution to be generalised to improve the task's validity as a measure of problem solving ability and to ensure less clustering of marks near the top of the scale. Since then, all problems have included a 'generalisation' component. McCrae also noted that test questions should be sufficiently similar to the corresponding problem questions to enable them to be answered by each student using the techniques *which he/she used* in solving the problem.

Stacey and McCrae (in press) analyse the changes which have taken place in the nature of the problems set in the problem solving CATs from 1989 to 1997. They judge that the trend has been towards problems that require less creativity in their solution, are more familiar in nature and offer less opportunity to generalise results. They conclude that, although the CAT is a less valid assessment of problem solving than originally, it still has sufficient validity to support its retention, especially since its existence continues to encourage creative problem solving work in earlier years.

The history of the investigative project CAT, until the end of 1993, is reviewed by Stacey (1995). Students had to submit a 1500–2000 word report of a project they had undertaken over a period of four weeks on a centrally-set mathematical theme. In the first few years there was considerable scope for students to shape their own investigations. They were able to define the path of their investigation, as well as the mathematical techniques to be used, to determine what data or information they needed and to obtain it themselves by experiment or from an appropriate source.

However, widespread concerns about the consistency of assessment across the State of very divergent projects, and about the difficulty of determining beforehand the breadth and depth of investigation needed for a good grade, soon led to a reduction in the openness of the CAT. By 1993, students had to select one of three specified 'starting points' within a theme for their investigation. Each starting point consisted of a number of questions intended to provide clear guidance on the expected direction and extent of the investigation, including in what way(s) it might be extended. There was still scope for students to choose what mathematical methods they would employ, but they were no longer expected to obtain their own data. Stacey concluded that, despite the changes, the CAT was still a valid assessment of investigative work in mathematics.

From 1994 to 1996, the format of the investigative project CAT remained the same, though the starting point questions typically became less open to interpretation and to method of solution and continuing attempts were made to constrain the breadth (if not the depth) of extension work. The main objective of this 'tightening up' was to limit the excessive amount of time that students were spending on the CAT in their quest for high grades. The total time spent on the project was not supposed to exceed 10 periods in class and 20 hours out of class, but many schools devoted all class time during the four week period to it and outside of class students were concentrating on the project to the detriment of everything else. The matter came to a head in 1996, with a large and influential group of schools reporting that, in Mathematical Methods 3/4 (the mainstream VCE mathematics subject) students averaged nearly 90 hours on the investigative project CAT.

The response of the authorities to the workload crisis was to reduce the prescribed time period for the CAT from four weeks to two weeks, to shorten the length of the report to 1200–1500 words and to direct the investigations even more closely with starting point questions that were predominantly closed. This overhaul was accompanied by the introduction of a test/interview process, identical to that associated with the *Problem-solving* CAT, to assist with the authentication of students' reports which could be expected to be much more convergent than previously under the new conditions. Thus, in 1997 the Mathematical Methods investigative project CAT had two components: the report and a closely related one-hour test, conducted three days after the report was due and contributing 25% to the final grade.

The format of the investigative project CAT for Further Mathematics 3/4, a lower level VCE mathematics subject, was not altered for 1997—the time period is still four weeks and there is no test component. Excessive workload has never been a major issue in this case as, typically, Further Mathematics students are not competing for selection into courses that require a high Tertiary Entrance Ranking score.

The 1997 Study

The authors conducted an evaluation of the implementation of the 1997 Mathematical Methods 3/4 investigative project CAT, similar to the 1994 evaluation of the *Problem-solving* CAT (McCrae, 1995). The evaluation was carried out through the use of a questionnaire sent to 126 secondary schools randomly selected, using stratified random sampling, from approximately 500 schools teaching Mathematical Methods. Sufficient copies of the questionnaire were provided to each school so that each teacher of Mathematical Methods could independently complete and return it. A total of 92 schools replied to the survey, a response rate of 73%, with 160 individual teacher responses. The response rates for metropolitan schools (72%) and country schools (75%) were about the same (after some follow up with metropolitan schools that did not reply by the due date). The response rates by sector were: Government 70%, Catholic 79% and Independent 78%; one of two TAFE institutes also responded. The number of students that inform the survey total about 3100. This represents just over 17% of the 17 833 students that completed the CAT.

The questionnaire contained 42 questions covering eight areas: background information, project attributes, report attributes, test attributes, comparing report and test grades, interviews, authentication issues and other issues. Most questions required respondents to select from one of five alternatives on an ordinal scale (eg. strongly disagree, disagree, unsure, agree, strongly agree). For data analysis, a corresponding scale of 1 to 5 was used so that an average rating could be calculated. In a few questions, respondents were asked for a numerical response (e.g. the average amount of time spent in assessing each report). Space was left at the end of the questionnaire for respondents to write comments on any aspect of the CAT and just over half (51%) availed themselves of this opportunity.

Students had to choose between two starting points, SP1: *Bushwalking with Kim* and SP2: *Blood flow* (Board of Studies, 1997). Respondents regarded both starting points as suitable for students to demonstrate their abilities in project work, with SP1 (average rating 3.8) marginally more suitable than SP2 (average rating 3.5). Both starting points were seen as requiring a reasonable balance of routine and non-routine work, with SP1 (average rating 2.7) tending to involve more routine work than SP2 (average rating 3.0). *Bushwalking with Kim* was by far the more popular choice, being attempted by 82% of students.

The ten criteria used to assess students' reports were judged to be relatively easy to apply and the five-point scale (very low, low, medium, high, very high) was seen to make it relatively easy to discriminate between the reports. Three-quarters of respondents reported that their students spent more than 25 hours on the investigation and teachers spent an average of 43 minutes per report as first markers. About two-thirds of the respondents also spent an average of 23 minutes second marking reports and some complaints about the workload involved in assessing the reports were registered.

Respondents agreed that the test was very closely related to the investigation in both cases (average rating 4.6 for both). However, the tests were criticised for favouring students who had used particular approaches in their investigations and it would appear that the setting panel needs to be more careful in this regard in the future. Just over half of the respondents felt that the level of difficulty of the tests was 'about right', with all but a handful of the rest believing that the tests were too easy.

On average, respondents believed that both tests discriminated reasonably well between levels of understanding of the mathematics employed in the students' reports (average ratings SP1: 3.1, SP2: 3.2). However, each of the five possible responses from 'not well' to 'very well' had a solid group of supporters. The explanation for this volatility lies in accompanying comments such as "The test discriminated well between two standards of students but in the top group there was not much of a spread of results," an observation which is confirmed by the histograms of scores reproduced in Figure 1. This failing occurred because, in both cases, the hardest parts of the investigation were not tested. This deficiency should be remedied in future years as it also brings the test's effectiveness as an authentication instrument into question.

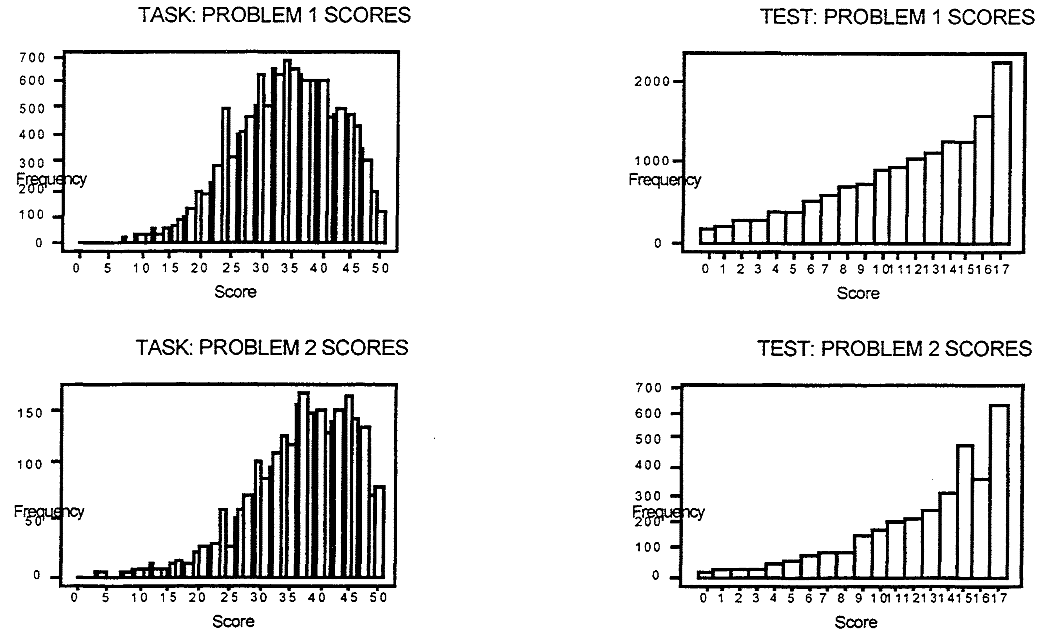


Figure 1. Histograms of report and test scores for Mathematical Methods CAT 1, 1997.

For both starting points, respondents were well satisfied on average with the correspondence between their students' performances on the report and the test (average ratings SP1: 3.6, SP2: 3.8). The population report/test correlations were 0.665 for SP1 and 0.616 for SP2. These are lower than might be expected, given the close relation between investigation and test, and probably suffered from the test deficiencies already identified with some respondents expressing views similar to the following comment:

> My students generally performed better on the test than they did on the actual project .... The test really only tested the routine elements of the question and did not ask any 'thinking' questions to discriminate [between] those who had mastered the routine maths and those who understood the more complex concepts involved.

Respondents advised that interviews were necessary in 173 cases for SP1 (6.9% of those who did that starting point), with 16 (9.2%) of those interviewed not having their reports authenticated. The corresponding figures for SP2 were 25 (4.4%) and 6 (24%). In total, 198 (6.4%) of the students covered by the sample had to be interviewed and 22 (11%) of them—about 0.7% of the student sample—did not have their reports authenticated. Respondents were generally well satisfied with the outcomes of the interviews (average rating 4.6).

Students must provide evidence of progress on their investigation and demonstrate that it is their own work by having at least one consultation with their teacher, based on a draft of their report or their log book contents, during the period allowed for the investigation. Consultation with students was rated as the most important factor in enabling teachers to authenticate students' reports, with 70% of respondents regarding it as an essential condition. The submission of log books, containing all working notes, (average rating 4.0) and draft material (average rating 3.9) were judged as very important, whilst the existence of the test (average rating 3.7) and the possibility of interview (average rating 3.6) also were regarded as important, but not as essential as the other factors.

There was general agreement that the report plus test format improves public credibility of the CAT (average rating 4.0) and, to a marginally lesser extent, agreement that the CAT's credibility among VCE students also is improved (average rating 3.8). There was a strong tendency to agree that the new format for the CAT assists teachers to authenticate authorship of reports (average rating 3.6), but a tendency to disagree that it reduces the amount of collaboration that might otherwise occur (average rating 2.7).

Students may 'consult any source material or person' (Board of Studies, 1996, p. 60) in preparing their project report, but are expected to acknowledge all assistance provided and how it helped them complete the task. It is clear from a number of similar written comments that many teachers believe that, despite the presence of the test/interview process and other conditions designed to provide evidence of authenticity, the CAT remains open to abuse by students receiving unacknowledged assistance from sources such as fellow students, teachers, family members, private tutors and commercial seminars. The issue of acknowledgment aside, many respondents expressed concern that students who could not access such forms of assistance because of their family's socio-economic situation were unfairly disadvantaged.

Respondents clearly agreed that structuring the starting points so that the questions required specific mathematical responses makes the project accessible to more students (average rating 4.1) and generally agreed that this should be a feature of future investigative project CATs (average rating 4.0). They tended to agree that it reduces the amount of time that students might otherwise spend on the task and increases the reliability with which they can assess students' reports (average rating 3.5 in both cases). They were unsure as to whether it increased their confidence in the authenticity of their students' reports (average rating 3.1) and tended to disagree that it reduced the task's validity as a measure of students' ability at investigative project work (average rating 2.7).

Some respondents commented to the effect that the redesigned CAT was no longer really an investigative project, but more a problem solving activity akin to the *Problem-solving* CAT. Nevertheless, these respondents, like the group as a whole, tended to agree that the new format enabled a more valid assessment of students' abilities in project work. This apparent contradiction may be seen as an indication that teachers' concerns for their students' results and welfare outweigh all other factors. One respondent observed that it was a "much better CAT period than last year .... Students were less anxious, had more sleep and stayed in better health than last year."

Use of Powerful Technology: An Emerging Issue

The use of technology emerged as an issue in the 1994 evaluation of the *Problem-solving* CAT. Students had relied heavily on the aid of graphics calculators or computer software (such as graph plotters and spreadsheets) in solving one of the problems, but this technology was not allowed in the test. In early 1995, it was announced that students would be allowed to use graphics calculators in the end-of-year VCE mathematics examinations from 1997. Later, it was decided that also from 1997 one of the *Problem-solving* CAT problems, and the related test, would assume that students had access to a graphics calculator.

The use of technology in the Mathematical Methods 3/4 investigative project CAT has been actively encouraged since 1995. At first (1995), this encouragement was restricted to using a graphing package but in subsequent years it has been extended to include the use of a graphics calculator, a computer spreadsheet or 'other computer packages' (Board of Studies, 1997, p. 5). It is doubtful that the excessive workload issue, which caused the CAT's redesign for 1997, would have arisen without the availability of such technology which enabled students to generate graphical representations and analyses not otherwise within their capabilities.

Respondents to the 1997 survey reported that most students had easy if not constant access to a graphics calculator (average rating 3.9) and made reasonable use of it if they did have access (average rating 3.2 for both starting points). Access to computer software such as graph plotters and spreadsheets was quite reasonable (average rating 3.6), though not as constant as graphics calculator access, and tended to be more frequently used than were graphics calculators (average ratings SP1: 3.7, SP2: 3.6). A significant part of this preference for using computer software can be attributed to its ability to produce a better quality product than graphics calculators for direct inclusion in students' reports.

Question 42 of the survey (see Figure 2) presents a partial solution by a supposed student author to one part of *Bushwalking with Kim*. The solution uses the symbolic manipulation capabilities of the Texas Instruments TI-92 calculator to perform most of the algebra involved. It demonstrates that the impact of powerful technology, such as computer algebra systems, on VCE mathematics assessment will not be confined to the substantial impact it is going to have on the examination CATs (McCrae, 1996). (As another demonstration, McCrae (in press) shows that *Bushwalking with Kim* can be tackled very effectively with dynamic geometry software such as The Geometer's Sketchpad and Cabri-geometry.)

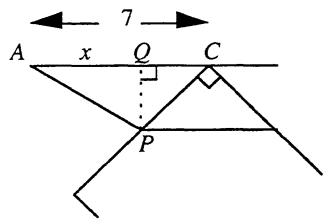
Question 42

Consider the following partial solution to part **c.** of Starting Point 1. In answering this question, **assume that you have no reason to doubt** that the student author **did** perform the derivations on his/her TI-92 calculator.

a. Would you in the current circumstances have
- **A.** authenticated the work irrespective of the student's known strength at calculus and algebra
- **B.** authenticated the work if you knew or could establish (eg by interview) that the student understood all the steps involved, without necessarily being able to perform them by hand
- **C.** authenticated the work if you knew or could establish (eg by interview) that the student could perform all the steps involved by hand
- **D.** *not* authenticated the work irrespective of the student's known strength at calculus and algebra

b. If all students had access to a calculator like the TI-92, do you believe that such work should be
- **A.** authenticated irrespective of the student's known strength at calculus and algebra
- **B.** authenticated if you knew or could establish (eg by interview) that the student understood all the steps involved, without necessarily being able to perform them by hand
- **C.** authenticated if you knew or could establish (eg by interview) that the student could perform all the steps involved by hand
- **D.** *not* authenticated irrespective of the student's known strength at calculus and algebra

Partial Solution to part ***c.***



Let x km be the distance from A to Q as shown on the diagram.
Then, by Pythagoras in ΔAQP,

$AP^2 = x^2 + (7-x)^2$ (since $QP = QC$)

$= 2x^2 - 14x + 49$ where $x \le 7$ and $2(7-x)^2 \le CD^2 = 49$

That is, by TI-92 screen dump 1, *[Screen dumps have been omitted]*

$AP^2 = 2x^2 - 14x + 49,\ 7 - 7/\sqrt{2} \le x \le 7.$

Using $t = \frac{d}{v}$, Kim's travelling time, T hours, is given by

$$T = 2\left\{\frac{\sqrt{2x^2 - 14x + 49}}{v_b} + \frac{7-x}{kv_b}\right\} = \frac{2\left\{k\sqrt{2x^2 - 14x + 49} + (7-x)\right\}}{kv_b}$$

By screen dumps 2–6, this has a minimum at $x = \dfrac{7\left(\sqrt{2k^2-1}+1\right)}{2\sqrt{2k^2-1}}$

and by screen dumps 7–9, the minimum time is $\dfrac{7\left(\sqrt{2k^2-1}+1\right)}{kv_b}$.

Figure 2. Question 42 on the Mathematical Methods CAT 1 questionnaire.

The first part of Question 42 asks respondents under what conditions (if any) they would authenticate the student's work in the current circumstances of limited access to such technology. The second part is identical to the first part except that respondents were to assume that all students had access to a calculator like the TI-92. The idea for this question came from Roberts' (1997a) account of the dilemma he faced in assessing the *Problem-solving* CAT of one of his students in 1996. The student had obtained the general case solution to the last part of the problem apparently by using a computer algebra system to solve an equation he had formulated. When interviewed, the student was unable to explain the steps involved in the solution process—though he could explain the equation, and so this part of his report was not authenticated. Stacey (1997) and Pierce and Roberts (1997), all from universities, responded to Roberts and argued that the student's work could have been authenticated, but Roberts (1997b) was not convinced.

The responses to Question 42 leave us in no doubt that most VCE mathematics teachers would support Roberts. A majority (56%) of respondents indicated that in the current circumstances they would have authenticated the student's work only if they knew or could establish (e.g. by interview) that the student could *perform* all the steps involved by hand. Just over half as many (29%) would have authenticated the work if they knew or could establish that the student *understood* all the steps involved without necessarily being able to perform them by hand. Only a small number (9%) of respondents would have authenticated the work irrespective of the student's known strength at calculus and algebra, or not authenticated it irrespective of those strengths (6%).

As to what they would have done if all students had access to a TI-92, there was practically no shift in the attitudes of respondents, with the corresponding percentages being 56%, 29%, 10% and 5%. This should not be taken as indicating that teachers are unconcerned about equity of access to technology at present, but rather it indicates their strong conviction that students should continue to be able to do their own algebra in future. It is difficult, however, to see how this view can prevail as access increases to computer algebra systems that can show each step in a solution for students to copy: the authentication burden on teachers would be immense. If the school-based CATs are to continue in their current form, then it is more likely that the use of powerful technology, such as computer algebra systems and dynamic geometry software, will need to be accommodated in a similar manner to that which has occurred already with other technology, such as plotting software and spreadsheets.

Concluding Remarks

The use of school-based assessment tasks in the VCE has given rise to several concerns over the ten years of their existence. The principal areas of concern have been the consistency and reliability of assessment across schools, authentication of student work and 'cheating', student and teacher workload, fragmentation of teaching and learning, and the usefulness of the results for selection into tertiary courses (Hill, Brown & Masters, 1993). The two school-based mathematics tasks, an investigative project CAT and a problem solving CAT, have both undergone significant changes during this period in response to these concerns and, in their present forms, appear to satisfy the criteria for school-assessed tasks fixed by the recent Committee of Review on the Victorian Certificate of Education (1997, recommendation 25).

In both cases, the price that has been paid has been the sacrificing of much of the 'openness' that characterised the original tasks. This is particularly true of the Mathematical Methods 3/4 investigative project CAT which has evolved into a task that is essentially indistinguishable in character from the *Problem-solving* CAT. In the forthcoming reaccreditation process, careful consideration needs to be given to the future role of project work in VCE mathematics, including its assessment.

Finally, there has been increasing use of technology throughout the history of the school-based assessment tasks. This trend will continue as more powerful technology, such as computer algebra systems, becomes readily accessible to students. The effect this will have on the assessment of the CATs needs addressing in the wider context of its impact on VCE mathematics curriculum and assessment in general.

References

Board of Studies. (1996). *Mathematics study design (1997–2000).* Carlton, Victoria: Author.

Board of Studies. (1997). *Mathematical Methods Units 3 and 4 Common Assessment Task 1: Investigative project.* Carlton, Victoria: Author.

Committee of Review on the Victorian Certificate of Education. (1997, December). *Enhancing their futures* (Report of the Committee of Review on the Victorian Certificate of Education). Melbourne: State of Victoria.

Hill, P., Brown, T. & Masters, G. (1993, November). *Fair and authentic school assessment* (Advice to the Board of Studies on verification, scaling, and reporting of results within the VCE). Carlton, Victoria: Board of Studies.

McCrae, B. (1995). An evaluation of system-wide assessment of problem solving at year 12 by report and related test. In B. Atweh & S. Flavel (Eds.), *Proceedings of the 18th Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 409–415). Darwin, Northern Territory, Australia: Mathematics Education Research Group of Australasia.

McCrae, B. (1996). The use of calculators in VCE examinations: looking ahead. *Australian Senior Mathematics Journal 10* (1), 65–71.

McCrae, B. (in press). Modelling using dynamic geometry software. In P. Galbraith, W. Blum, G. Booker & I. Huntley (Eds.), *Mathematical modelling: teaching and assessing in a technology rich world.* Chichester, U.K.: Albion Press.

McCrae, B. & Stacey, K. (1997). Testing problem solving in a high-stakes environment. In E. Pehkonen (Ed.), *Use of open-ended problems in mathematics classroom* (pp. 34–48). Helsinki: Department of Teacher Education, University of Helsinki.

Pierce, R. & Roberts, L. (1997). In defence of computer algebra. *Vinculum 34* (2), 19.

Roberts, N. (1997a). On the use of computer algebra in project work. *Vinculum 34* (1), 9.

Roberts, N. (1997b). Further thoughts on computer algebra at VCE. *Vinculum 34* (2), 20–21.

Stacey, K. (1995). The challenge of keeping open problem-solving open in school mathematics. *Zentralblatt für Didaktik der Mathematik 95* (2), 62–67.

Stacey, K. (1997). Computer algebra: the coming challenge for the mathematics curriculum. *Vinculum 34* (2), 16–18.

Stacey, K. & McCrae, B. (in press). Assessing problem solving: give and take. *The Mathematics Educator.*

Stephens, M. & McCrae, B. (1995). Assessing problem solving in a school system: principles to practice. *Australian Senior Mathematics Journal 9* (1), 11–28.