ESTABLISHING COMPARABILITY OF YEAR 12 SCHOOL-BASED ASSESSMENTS

Peter W. Hill, Tim Brown, Kenneth J. Rowe
University of Melbourne
and Ross Turner
Board of Studies

Paper presented at the 1994 annual conference of the Australian Association for Research in Education, Newcastle, New South Wales, Australia.
November 27 to December 1, 1994.

Correspondence:
Peter W. Hill, Centre for Applied Educational Research, John Smyth Building, Faculty of Education, The University of Melbourne, Parkville, Victoria 3052, Tel. 03 344 8201.

Acknowledgement:
The cooperation of the Board of Studies and permission to reproduce materials from Board publications is gratefully acknowledged. The views expressed here are those of the authors and are not necessarily the views of the Boards of Studies.

ESTABLISHING COMPARABILITY OF YEAR 12 SCHOOL-BASED ASSESSMENTS

Abstract: At the Year 12 level in Victoria, as in several other state systems, there is widespread recognition of the value of ensuring that final assessments are based on a mix of subject-based examinations and structured school assessments. In particular, school-based assessment is seen as a means of allowing assessment of the broadest possible range of valued outcomes and of improving the validity of final assessments. The use of school assessments within a high stakes environment does, however, raise significant issues regarding the comparability of those assessments. In the past, these concerns were addressed largely through statistical moderation which ensured that the location and spread of school-based assessments within a given school were adjusted to the location and spread of the examination results of students within that school. With the introduction of the Victorian Certificate of Education, statistical moderation was abandoned in favour of a system of verification based on external checking of samples of student work. The system proved inefficient and followed reports by Brown and Ball (1992) and Hill, Brown and Masters (1993), the Board of Studies in Victoria adopted a new system for maximising comparability of school-based assessments. This involves all students taking a General Achievement Test (GAT) mid-way through Year 12. Scores on sub-sections of the test are used to construct composite scores which best predict students' results on school-assessed tasks. A series of three statistical tests are then used to identify schools for follow-up by external reviewers. Where the assessments of a school are within the expected range given their weighted GAT scores, the school's assessments are confirmed. Where there are discrepancies, adjustments are made. This paper describes the new system, presents some preliminary results from applying the new procedures in 1994 and provides a general discussion of the challenges presented in endeavouring to ensure fair and authentic school assessment at the Year 12 level.

The `comparability dilemma'

Ensuring comparability of assessments provided by schools is an important consideration in the Year 12 assessment and certification arrangements of all States and Territories in Australia. In the Australian Capital Territory and Queensland, Year 12 certification is based entirely on moderated school assessments. In all other States and Territories, assessment of subjects or courses recognised for the purposes of entry to tertiary institutions is based on a combination of external examination results and moderated school assessments. For courses not recognised for tertiary selection, assessment typically is based entirely on moderated school assessments.

Assessment and certification agencies face difficult decisions in making use of school assessments of student achievement. On the one hand, they are drawn in the direction of emphasising school assessment as a means of improving the link between the curriculum and assessment procedures, of recognising the achievement of a wide range of valued outcomes not readily amenable to paper-and-pencil examination, and of catering for a more diverse clientele of students entering Year 12. At the same time, they are under increasing pressure to ensure that procedures for ensuring comparability of assessments are demonstrably fair and cost effective in the context of constrained budgets, larger enrolments and increasing reliance by employers and tertiary institutions on Year 12 certification for selection and access to scarce employment and further study opportunities. These pressures draw authorities in the direction of external, examination-style assessment. Not surprisingly, most systems attempt to balance the two approaches and to use both school assessments and external examinations in arriving at an overall assessment of students. In essence, the dilemma of the authorities is that of ensuring fair and authentic assessment in an increasingly high-stakes environment.

Various methods are used by assessment and certification agencies in maximising the comparability of the assessments provided by schools. Broadly speaking, these may be classified under three headings:

1.Ensuring that, within a given subject, schools are teaching and assessing broadly the same content and are using the same criteria to assess performance. This may involve:

∑Common syllabuses or courses and approved work programs.

∑Common assessment tasks and criteria for the award of marks or grades.

2. Establishing mechanisms involving expert or peer review of school assessments. These may take the form of:

∑Moderation committees which establish each year the specific procedures which schools will follow to enhance the comparability of assessments.

∑Consensus moderation, involving teachers meeting in groups to compare sets of school assessments and to either confirm or adjust schools' initial assessments.

∑Inspection of school assessments by externally-appointed moderators who either confirm or adjust the schools' initial assessments.

3. Statistical moderation, whereby the distribution of each school's assessments for a given subject is adjusted to conform to that of the external examination in that subject or to a specially developed reference test such as the Australian Scholastic Aptitude Test. This may involve:

∑linear transformation of school assessments in which the marks

allocated by the school are adjusted to be equal to the mean and standard deviation of the marks obtained by the students in that subject in the external examination or reference test;
∑curvilinear transformation of school assessments based on centroid or percentile scaling (to handle situations where school assessments are distinctly non-normal) using marks obtained by the student within a given subject on the examination or reference test to re-scale their school assessments.

Most comparability procedures involve a combination of 1 and 2, or 1 and 3, of the above approaches. In those States and Territories that have dispensed with subject-specific external examinations and also in other States and Territories in the case of non-tertiary courses or subjects with relatively small enrolments, comparability is pursued primarily through common curricula and assessment and expert or peer review mechanisms. In States and Territories where Year 12 assessment includes a significant component of externally-set, subject-based examinations, a combination of common curricula and assessment, and statistical moderation predominates.

Comparability procedures involving moderation committees, consensus moderation or visits by externally-appointed moderators, enable professional judgments to be made about the relative standards of work as assessed by different schools. Work of a particularly high or low standard can be identified and assessed accordingly. Such procedures also have the merit of involving teachers in discussing standards of work and thus contribute to enhanced professional awareness of standards. They have a number of shortcomings, however, of which the following are noteworthy:

∑They are resource intensive in terms of both time and the dollar costs associated with employment of moderators, teacher release from classes and travel expenses.

∑The potential exists for significant inflation of grades as teachers seek to enhance students' prospects of a place in higher education (Karmel, 1985, p.94).

∑Achieving consistency in standards among moderators or among teachers within moderation panels is difficult and considerable effort is required to maintain acceptable levels of inter-rater reliability (Brown and Ball, 1992). This problem increases as the size of the candidature and the number of schools involved increases.

∑Where the review process allows dialogue between the reviewer and the reviewed, the potential exists for undue influence on the reviewer to accept the school's assessments of students (Brown and Ball, 1992).

Statistical moderation avoids some of the problems of expert or peer review procedures. In particular, it is an inexpensive solution to the comparability problem for those States and Territories that have retained external examinations, it is not subject to problems of inter-rater reliability, and it is manifestly `objective'. An important feature of current statistical moderation procedures, as far as teachers are concerned, is that the scaling of school assessments preserves the order of merit of students as assessed by the school and the relative spacing between students. What changes is the level and spread of school assessments. Although offering certain advantages, statistical moderation has a number of disadvantages, of which the following are probably the most serious:

∑The examination score may not always be a valid moderator variable, particularly in cases where the school assessment has been specifically designed to measure outcomes which are not or cannot be assessed through the external examination (Masters and Hill, 1988). This applies particularly to subjects involving school-based assessment of outcomes requiring very different kinds of abilities from those required in the examination. For example, where the school assessment focuses on practical activities, performances, or extended research, the correlation between the school assessment and the examination is likely to be modest and the examination may not be a particularly valid indicator of the level of performance of students on the outcomes assessed at the school level.

∑Where examination scores are used as the moderator variable, statistical moderation encourages schools to focus all their efforts on maximising scores on the examination to the neglect of school assessment. This is because the distribution of school assessments is adjusted to conform to the distribution of examination scores, irrespective of the actual level of performance on school-based assessment activities. There is thus no incentive for the school to put significant effort into the school assessment, since, regardless of the actual standard of performance, students' assessments will automatically be adjusted to coincide with their performance on the examination.

∑Statistical moderation is problematic when applied to schools with small subject enrolments (Masters and Hill, 1988). This is because the magnitude of the adjustments made to the results of the students within the school is unduly influenced by the inclusion or exclusion of one or two individuals with high or low results. In practice, small enrolments (1-10 students) in a subject within a school are very common at the Year 12 level, so the `small n' problem is by no means a trivial matter. It can be minimised by requiring schools with small subject enrolments to combine with other schools to form `pseudo' schools which jointly submit their assessments for the purposes of statistical moderation. Unfortunately, this tends to place additional burdens on small, isolated schools who find it time-consuming and expensive to conduct joint meetings to establish comparability of assessments prior to submitting them for statistical moderation.

The above analysis suggests that neither of the two main approaches to achieving comparability of school assessments is likely to be entirely satisfactory. The purpose of this paper is to describe a new system that makes use of all three methods for achieving comparability and combines them in ways which attempt to maximise the advantages of each while minimising their disadvantages. The system was implemented for the first time in Victoria in 1994. This paper is therefore a preliminary report on the new procedures for verifying school assessments within the Victorian Certificate of Education (VCE).

Background to the new procedures

The background and rationale for the changes that have occurred in relation to the moderation of school assessments in Victoria are described elsewhere (Hill, Brown & Masters, 1993; McGaw, et al., 1990; Victorian Curriculum and Assessment Board, 1987, 1988). Prior to the introduction of the Victorian Certificate of Education, comparability of school assessments at the Year 12 level in Victoria was ensured primarily through the use of statistical moderation for Higher School Certificate (Group 1) tertiary entrance subjects (see Cropley, 1981), and through a variety of other procedures, particularly consensus moderation, for non-tertiary entrance subjects.

There had been considerable dissatisfaction surrounding the use of statistical moderation, particularly on account of the way in which it discouraged schools from giving proper attention to school-assessed `options' work, and because it led in many cases to very substantial adjustments to school assessments, particularly within schools with small enrolments. On the other hand, there was a powerful body of opinion that believed that the consensus moderation procedures that had been developed primarily for non-tertiary entrance subjects and which had gained substantial support among teachers, were insufficiently rigorous and too resource intensive to provide a more general solution to the comparability problem.

In responding to the recommendations of the Blackburn Report (Blackburn, 1985) and to the statement by the then Minister for Education on Future Directions in Post Compulsory Schooling (Cathie, 1986), which called for a new and common credential able to meet the needs of the great majority of young people, the response of the Victorian Curriculum and Assessment Board was to develop a new approach to assessment based on common assessment tasks combining a mix of external and school-based assessment, and a new approach to comparability of school assessed common assessment tasks based on what were termed verification procedures.

Verification of school-assessed common assessment tasks as implemented in the early years of the new VCE was a multi-stage process (see Victorian Curriculum and Assessment Board, 1992, for a full description). This process included a formal verification meeting early in the year attended by representatives from all schools to provide an overview of the verification process, to clarify assessment criteria for the award of grades, to ensure common approaches to the interpretation and application of criteria, to view samples of work from the previous year and to undertake other activities directed at assisting schools in providing comparable assessments of student work. In the second stage, schools administered common assessment tasks and rated student performance on an eleven point scale (UG, E, E+, D, D+, C, C+, B, B+, A, A+) on these tasks using centrally prescribed criteria. Upon completion of initial school assessments all work assessed at the lowest and highest level of performance (UG and A+) and two pieces from each school selected at random at all other grade levels (E to A) were submitted to verification panels to an external panel chairperson who either confirmed the schools' assessments or called for schools to submit all of their students' work and if appropriate change the grades awarded by schools. The final stage involved a November verification meeting to review the outcomes of the verification process and make proposals for adjustments in subsequent years.

With the full implementation of the new VCE in 1992, it became apparent that there were problems with the VCE verification process and substantial criticisms were voiced in the public media and elsewhere. This led to a report on VCE verification by Brown and Ball (1992) who concluded that the process was insufficiently reliable and should be discontinued. Brown and Ball also made an important recommendation regarding alternative mechanisms for verifying school-assessed common assessment tasks. They noted that a key deficiency of previous approaches to statistical moderation was that they automatically adjusted the grades of students. They proposed the use of the external examination in each subject to check on the reasonableness of school's assessments and to make no adjustments to the marks of schools if they fell within an expected range. This would then enable a greater focus on the assessments of those schools that were significantly higher or lower than expected.

This proposal was incorporated into a further set of recommendations in a subsequent report by Hill, Brown, & Masters (1993) which were adopted by the new Board of Studies and form the basis of the current procedures as described below. In brief, the process followed by the Board is as follows. The distribution of assessments submitted by a school is compared with that which could be expected on the basis of their scores on an external reference test known as the General Achievement Test (GAT). If the school's assessments are within a specified tolerance band, the school's assessments are automatically confirmed. If they fall outside the tolerance band, the school's assessments are subject to re-marking by two external reviewers. This re-marking in turn leads to a decision either to confirm the school's assessments or to adjust them.

The process specifically does not make use of subject-based examinations as the moderator variable since these are usually taken towards the end of the year when there is no time for re-marking by external reviewers. The GAT is therefore administered mid-year and prior to the submission of school-assessed common assessment tasks.

The most significant feature of the new procedures is that no school's assessments are subject to statistical adjustment. Scores on the GAT are used solely for the purpose of checking on the reasonableness of schools' assessments and of identifying schools for follow-up. Adjustments are made only on the basis of expert review of schools' assessments. This means, for example, that schools that achieve better-than-expected results on the school-assessed task may have their assessments confirmed if the external reviewers agree with the schools' assessments. This in turn encourages schools to attach the same importance to school-based assessment as to external assessment. The new procedures are outlined in further detail below.

Common Assessment Tasks

Comparability of school assessments depends to a considerable extent on the degree to which schools teach the same content and assess the same things in the same manner. The VCE is based on a total of 44 studies (subjects) which have been developed to cater for the full range of student aptitudes and interests. Studies are not categorised by the Board of Studies or by universities as either tertiary or non-tertiary entrance subjects. Study designs have been developed which outline the work students are required to complete in order to fulfil the requirements of each semester unit of study.

Assessment of Year 12 level studies (units 3 and 4) is through common assessment tasks. For each subject or study there are typically three common assessment tasks which together are designed to assess the key learning outcomes for that study. In most studies, one of the common assessment tasks is completed under test conditions and externally marked (ie., it has the properties of a public examination), while the other two common assessment tasks are assessed at the school level and subject to moderation. All school-assessed common assessment tasks have been designed so that they arise directly out of the work requirements built into each study design.

Details of each common assessment task are specified in the relevant study. The tasks vary considerably in their form and the level of specification. Some of the more common types of tasks include a report of an investigation or piece of research, an essay, an analysis task, a set of structured questions, a folio of writing or graphic work, a performance, and the creation of a design, product or model. As an example of a relatively open-ended common assessment task, completed over an extended period of time, Table 1 reproduces the description contained in the study design for English for the second common assessment task for that study, namely the Writing Folio. From this specification it will be apparent that the study design prescribes a range of features of the task, including the number and type of pieces of writing and the processes to be followed in authenticating students' work; that is, in ensuring that the work is the students' own.

Criteria for the award of grades

A further way in which the VCE attempts to maximise the comparability of school assessments is to specify the criteria that teachers will use in assessing student performance on common assessment tasks. Rating scales have been developed for each school-assessed task which identify criteria for the award of grades and require teachers to rate students' responses to common assessment tasks against these criteria as `very high', `high', `medium', `low', `very low', `not shown', thus giving a 0-5 point scale for each criterion. By aggregating ratings against each of the criteria a total numerical score is obtained and this score is used to allocate a grade to each student. Figure 1 reproduces the assessment sheet which teachers are required to use for VCE English CAT 2: Writing Folio.

The Board of Studies monitors the way in which the assessment criteria are used by teachers, and the extent to which the rating scales are providing valid and reliable information. To do this, data are gathered from two sources. Qualitative data from teachers and others directly involved in assessing student work is used to evaluate the validity of the criteria and to shape professional development activities and training materials. Quantitative data from a sample of completed assessment sheets is analysed using Partial Credit Analysis, a procedure based on Rasch modelling (Wright and Masters, 1982). This analysis provides information on the measurement properties of the criteria. Criteria are modified as necessary on an annual basis in the light of both the qualitative and quantitative information.

The General Achievement Test

Schools submit their initial assessments of students on each of the school-assessed common assessment tasks to the Board of Studies having ensured within-school comparability of assessments provided by different teachers of the same subject within their school. The Board of Studies then makes use of results of scores of all students on a General Achievement Test (GAT) to perform various checks on the reasonableness of school's numerical assessments of students.

Table 1. Specification of VCE English
Common Assessment Task 2: Writing Folio

Description
Type of Task
The task requires students to present a folio of selected pieces of writing.

Purpose of task
The task is designed to assess students' competence in writing for a range of purposes and audiences.

Details of task
∑Each student will submit a folio of three pieces of finished writing.
∑Each piece should be written for a different purpose, clearly identified by the student on a cover sheet attached to each piece.
∑The three pieces should be written for at least two different audiences, clearly identified on the cover sheet.
∑In addition to individual cover sheets, the students will attach a contents page and list of references used, if appropriate.

Conditions
The task will be completed under the following conditions.

Time
∑Completion date of task: August-September, by a date to be determined annually by the Board.

Access to resources
∑Students may consult any source material or person in preparing the folio.

Setting
Topics for the pieces of writing will be determined by the school in accordance with the `Details of task'.

Authentication
The following authentication procedures will apply.
∑The teacher will monitor the development of the task by sighting plans and drafts of the student's work, and the teacher will record this process.

∑The student is expected to retain appropriate documentation of the development of the task to enable the teacher to attest that the work is the student's own.
∑To assist in the authentication of the student's work, the teacher may consider it appropriate to ask the student to demonstrate his or her understanding of the task. It will contain a statement that all unacknowledged work is the student's own.
∑The work will be assessed only if the teacher can attest that, to the best of his or her knowledge, the work is the student's own.

Assessment

The task will be assessed initially by the school. Details of assessment procedures are contained in the VCE Administrative Handbook(s) published annually by the Board.

Reproduced from Board of Studies (1993, 9-10) with permission

Details regarding the GAT are published by the Board (Board of Studies, 1994a). The test consists of four papers, namely two writing papers each taking 30 minutes and two multiple-choice papers, each taking 90 minutes. The first of the writing papers requires students to respond in a factual way to a question related to some written and/or graphical information. The second of the writing papers requires students to express their views on a social issue and to present reasons and arguments to support their opinions. In the writing papers the emphasis is on using well organised and accurate prose and on communicating clearly and effectively. The multiple-choice papers each consist of 50 multiple-choice questions drawing on the areas of: 1) mathematics, science and technology, and 2) humanities, arts and social sciences. These items do not test specific knowledge but rather are a test of those general skills considered to underlie VCE studies.

In 1994, all Year 12 students completed the GAT mid-way through the year and prior to schools submitting the results of student performance on any of the common assessment tasks. Following marking of students' responses, three scores were computed for each student, namely written communication (G1), mathematics, science, technology (G2), and humanities, arts, social sciences (G3).

Statistical procedures for identifying schools for follow-up

Details of the statistical checks undertaken to identify schools for follow-up are contained in a technical bulletin published by the Board (Board of Studies, 1994b). The aim of the statistical checks is to identify unexpected patterns of results given students' scores on the GAT. In particular, the checks aim to identify schools with unexpectedly high or low scores on the school-assessed common assessment task, or scores that are unexpectedly bunched together, or spread out. Statistical checks cannot be relied upon to identify unexpected patterns of results among schools with very small enrolments, nor among schools for which there is significant missing data. Thus, schools with less than five students are removed from the statistical checking procedure and subject to automatic follow-up, as are schools for which more than 20 per cent of students do not have useable GAT scores.

For the remaining schools, three tests are used. The first involves fitting a two-level regression model to the data and estimating the magnitude of the school-level residual term in the model. This residual term is the primary indicator of whether a school will be subject to follow-up and external marking. Two other tests are also used to identify schools for follow-up with unexpected patterns of results not identified by the first test. The first is a test of dispersion and the second a test of location.

Before undertaking these three statistical tests, for each school-assessed task, a weighted combination of the three GAT scores (G1, G2 and G3) is computed as:

EMBED Equation.2 [1]

where EMBED Equation.2 is the weighted composite GAT score for each student i, in school j, on common assessment task k, and b1k, b2k, b3k are level 1 (student) regression coefficients obtained by using a two-level (student and school) regression model to regress scores on the school-assessed task (EMBED Equation.2) on the three GAT scores for the population of students for whom data are available. This means that for each school assessed task, scores on the GAT are combined differently and in a way which maximises the correlation with the school assessed task. For example, in the case of the Writing Folio in English, the above procedure led in 1994 to the following weights being used to combine the three GAT scores:

Written expression (G1).314
Mathematics, science, technology (G2).065
Humanities, arts, social sciences (G3).281

These weights reflect the nature of the abilities underlying the Writing Folio. A very different pattern of weights would be expected, however, for other school assessed tasks.

A check is also made on the correlation of the predicted school assessment obtained from the weighted combination of the three GAT scores, adjusted for the differences in intercepts among schools (EMBED Equation.2), with the actual score provided by the school (

EMBED Equation.2). If the correlation between these two scores falls below .45, no further statistical checks are undertaken and all work for that task is subject to follow-up. In the case of the Writing Folio in English, a value of r = .705 was obtained in 1994.

The first of the statistical tests involves estimating school-level residuals obtained from fitting a two-level regression model to the data. This test makes use of recent methodological advances in the development of computer routines for fitting regression models to hierarchically organised or multi-level data (for a comparative review of currently available software packages, see Kreft, de Leuuw & Kim, 1990). The model used is a two-level model which assumes that the actual scores of individual students as assessed by the school can be predicted by a knowledge of their expected level of performance as indicated by a weighted combination of their scores on the GAT, and of the identity of the school providing the initial assessment. The model can be written out in two parts. First, the relationship between the school-assessed task and the weighted GAT score can be expressed as:

embed Equation.2 [2]

defining:
embed Equation.2 as the score on the school-assessed task for student i in school j for task k;
embed Equation.2as the intercept for school j, task k;
bkas the slope of the regression line for predicting the embed Equation.2 for task k;
xijkas the weighted GAT score for student i, in school j, for task k; and
eijkas the residual or unique contribution of student ijk.

Second, the relationships between the scores at the level of schools can be expressed as:

embed Equation.2 [3]

defining:
akas the mean of means for all schools (constant term) on task k; and
ujkas the residual or unique contribution of school j on task k, beyond that explained by the constant term ak.

Combining the above into a single equation gives:

embed Equation.2 .[4]

This model is fitted to the data for each school-assessed task using a generalised least squares procedure developed by Prosser, Rasbash & Goldstein (1993). A check is made on the reasonableness of the assumption that the coefficient bk should be fixed at the same value for all schools. In 1994, this assumption was found to hold across the full range of school-assessed tasks.

Having fitted the two-level regression model of equation [4] to the data for a particular school-assessed task, the focus of interest is on the estimate of the residual term for each school (embed Equation.2), because this indicates how discrepant a school's scores are from that which would be predicted by the model. These residuals are standardised by dividing them by the square root of their respective comparative variances and schools are sorted on the basis of the magnitude of these standardised school-level residuals. Large negative residuals indicate schools with assessments below expectation and large positive residuals indicate schools with higher than expected assessments. Schools with large positive or negative standardised residuals are identified for follow-up.

Thereafter two supplementary statistical tests are undertaken. First, a test of dispersion is used to ascertain whether the spread of scores provided by the school for a given task differs significantly from the spread of the corresponding weighted GAT scores. The test used is Pittman's (1939) test for equality of variances, given by the formula:

embed Equation.2 [5]

defining:
vdjkas the test statistic for equality of variances for school j, task k;
embed Equation.2 as the variance of the scores on school assessment for school j, task k;
embed Equation.2 as the variance of the weighted GAT score for school j, task k;
njkas the number of students in school j with both a school assessment and a GAT score for task k; and
rjkas the correlation coefficient between the school assessment and the weighted GAT score within school j, for task k.

The second test used is a test of location to ascertain whether the mean of the scores provided by the school for a given task differs significantly from the mean of the corresponding weighted GAT scores. The test used is the well-known test for equality of means given by the formula:

embed Equation.2 [6]

defining:
mdjkas the test statistic for equality of means;
embed Equation.2 as the mean school assessment of school j on task k;
embed Equation.2 as the weighted mean GAT score of school j, task k;
njk as the number of students in school j with both a school

assessment and a GAT score for task k; and
embed Equation.2 as the standard deviation of the differences between the school assessments and the weighted GAT scores for school j, task k.

In determining cut-off scores separating those schools that will be followed up with external marking and those whose assessments will be confirmed, decisions are made on the basis of practical significance in terms of the effect on the final grade awarded to students rather than on statistical significance. For example, in the case of the Writing Folio in English, a decision was made in 1994 to follow up 92 out of 513 schools. This means that the assessments as provided by 82 per cent of schools were confirmed, while the assessments of the remaining 18 per cent of schools were subject to the follow-up processes described below.

In addition to the schools identified for follow-up on the basis of the statistical comparison between school assessments and GAT scores, a small number of schools are selected for audit purposes. These schools are randomly selected from those with negligible school-level residuals from the multi-level modelling process.

The follow-up process

The follow-up process involves independent re-marking of a sample of student work from each school identified for external checking, comparing school assessments with the marks of external reviewers, and on the basis of this comparison either confirming a school's initial assessments, or continuing to re-mark all remaining work and determining final assessments on a case-by-case basis using pre-determined rules.

Schools identified for follow-up for a particular assessment task submit all student work to the Board of Studies for re-marking. In the case of schools with up to 20 students, all of the work is re-marked independently by two reviewers selected by the Board for their expertise in the study. Where there are between 20 and 50 students, a random sample of 20 is selected for re-marking. For schools with more than 50 students, a random sample of 30 is selected.

The school's assessments are compared with the average assessments of the two external reviewers. Two statistical checks are undertaken. The first considers the location of the school's assessments through a test of equality of means equivalent to the test described earlier (ie., equation [6], but with the mean of the two external reviewers' assessments being used in place of the GAT score). The second compares the spread of the school's assessments with the spread of the two external reviewers' assessments, and is derived from the test for equality of variances described earlier (equations [5]). For this test, the test statistic vdsjk(1) or vdsjk(2) is calculated as follows.

If the variance of the sample assessments in a school is greater than the variance of the average of the external reviewers' assessments for the sample, which will tend to occur under the extreme assumption that the school assessment is the product of a single assessor, then the test statistic vdsjk(1) for school j and task k is given by the formula:

EMBED Equation [7]

defining:
embed Equation.2 as the variance of the school assessments for school j, task k;
embed Equation.2 as the variance of the combined external reviewer scores for school j, task k;
njkas the number of students in school j with a school for task k; and
rsjkas the correlation coefficient between the school assessment and the average of the external reviewer assessments for the for school j, and for task k.

If the variance of the sample assessments in a school is less than that of the average of the external reviewer assessments for the sample, which may occur under the extreme assumption that all of the school assessments are the product of two or more assessors, then the test statistic vdsjk(2) for school j and task k is given by the formula:

EMBED Equation [8]

The test statistics statistic vdsjk(1) and vdsjk(2) are compared to the appropriate critical value of the t-distribution with njk-2 degrees of freedom.

If the difference between the school and external assessments is statistically non-significant for each of these two tests, then all of the school's assessments are confirmed. If the difference is statistically significant for either test, then all of the remaining work from that schools is re-marked independently by two external reviewers and final assessments for each student are determined.

The assessments of the external reviewers are first adjusted for any inconsistency. This adjustment is based on pair-wise comparisons of scores awarded by different external reviewers for the same pieces of work. A single external score is calculated from the two reviewers' scores. In cases where there are unacceptable discrepancies between the two external reviewers' assessments or between the school and the external reviewers' assessments, the single external assessment is derived as a result of a discrepancy process involving further re-marking. The final assessment is an equally-weighted combination of the school assessment and the adjusted external assessment.

Conclusions

The above procedures were implemented for the first time in 1994 and at this stage there has been insufficient time to evaluate them fully. It is nevertheless possible to draw some tentative conclusions and identify a number of issues for future consideration.

First, the new arrangements have significantly reduced the workload and costs associated with the previous verification procedures and enabled a sharper focus on a smaller number of schools where there is a priori evidence that the application of resource-intensive moderation processes is justified. Under the previous approach there was follow-up of student work in all schools. Under the new approach, follow-up by external reviewers is invoked only when there is evidence of potential discrepancies. In practice, the new procedures led in 1994 to about half the number of pieces of student work requiring re-marking by external markers of the corresponding figure for 1993. Approximately 85 per cent of schools had their assessments confirmed as a result of one or other of the statistical checks. For the remaining 15 per cent of schools whose assessments were subject to external review, some assessments were raised, some were lowered, and some were unchanged.

Final proportions in each category for 1994 are not yet available.

Second, there is evidence that a general achievement test such as the GAT can validly be used across the spectrum of VCE Year 12 studies. Correlations between observed and predicted school assessments in 1994 were quite high and in many cases higher than correlations between school assessments and scores for other common assessment tasks within the same study. Somewhat unexpectedly, this applied to practical subjects such as the Arts and various Technology subjects. This presumably reflects the way in which VCE studies have been developed and the attention given to ensuring that all studies cater for the full range of aptitudes and interests, emphasise generic competencies and involve a mix of theoretical and applied learning. The subjects with low correlations between observed and predicted school assessments were almost exclusively Languages other than English in which there were atypical distributions of school assessments. The magnitude of the correlations across all school-assessed tasks is summarised in Table 2.

Table 2. Magnitude of Correlations Between Observed and Predicted School Assessments

Correlation	Frequency	Per cent
0.50 and below	7	6
0.51 - 0.54	3	2
0.55 - 0.59	19	15
0.60 - 0.64	26	20
0.65 -0.69	41	32
0.70 - 0.74	26	20
0.75 and above	5	4

Alongside this evidence regarding the validity of the GAT for the purposes of identifying schools for follow-up there was further evidence that a number of test items included in the GAT in 1994 were inappropriate and of a kind unrelated to VCE studies. This probably reflects the extremely short time frame available to generate and trial the test items. Writing items that assess the more generic study skills underlying Year 12 studies and which are perceived to have validity in terms of the Year 12 curriculum will clearly be an ongoing challenge for the test developers.

Third, the statistical checks appeared to be effective in identifying those schools where follow-up of assessments was appropriate. Those with a knowledge of the outcomes of implementing comparability procedures over a number of years reported that many of the schools identified for follow-up were those that had also been identified under previous procedures as schools with discrepant patterns of results.

Fourth, the statistical tests based on the two-level regression model appeared to unaffected by sample size and there was no tendency for schools identified for follow-up to be those with either small or large subject enrolments. In addition, estimates of school-level residuals appeared to be sufficiently stable to be used in instances where there were as few as five students within a school with valid data. This latter property is of particular significance given the large number of schools with small subject enrolments.

Fifth, difficulties did arise in applying the new procedures to a small number of school-assessed tasks (mostly in languages other than English) in which students' work was assessed as all being of a high standard. Given that it is the policy of the Board that in all studies the criteria for the award of grades should discriminate different levels of performance, the solution to this problem may lie not so much with the moderation procedures as with the design of the common assessment tasks and the accompanying criteria. On the other hand, this may be unrealistic in the case of assessing native speakers in a language other than English.

Sixth, the new procedures appear to have generated a mixed reaction from schools. On the one hand, the majority of schools had their initial assessments confirmed and of those subject to follow-up, many had their assessments adjusted upwards. Among these schools there appeared to be considerable satisfaction with the process. Overall, there were fewer criticisms of the process itself by schools in 1994, with most regarding the introduction of the GAT and `blind' re-marking by two external reviewers as providing more rigour and `objectivity' than previous procedures. On the other hand, there was considerable dissatisfaction amongst other schools, particularly those whose initial assessments were adjusted downwards. This dissatisfaction was reported extensively in the local media and encompassed a range of objections to the processes used by the Board. The most salient criticisms appear to be directed not so much at the statistical processes used to identify schools for follow-up, nor at the processes for external review, but rather at the magnitude of the adjustments for individual students and the extent to which the rank order of students, as assessed by the school, changed significantly as a result of the follow-up process. In response to these criticisms, the Board has announced that it will review its current procedures, particularly the methods used to adjust schools' assessments where follow-up results in significant discrepancies between the school's assessments and those of the two external reviewers.

Seventh, there is evidence that those appointed as external reviewers employ stricter standards when acting as reviewers than they use when assessing their own students' work. Some work has been undertaken by the Australian Council for Educational Research to explore this problem and to develop methods for adjusting for the `harshness' or `leniency' of individual markers. Further work is required, however, to develop procedures which maximise comparability among those appointed as external reviewers and to ensure that the standards applied by these external reviewers are consistent with those employed across the State by teachers.

Overall, the new procedures appear to provide a way of valuing schools assessment and giving schools considerable professional responsibility for assessments within a quality assurance framework, making use of statistical techniques not to control school assessments but rather to help target moderation efforts. At the same time, the new procedures generate information which readily exposes problems of reliability and validity in contrast to previous procedures in which automatic adjustments took place and little further analysis was carried out. And certainly problems do exist. In particular, in terms of traditional indices of reliability, objective testing in quantitative subjects tends to result in more reliable assessments than do school-assessed tasks involving extended activities such as researching a topic, preparing a folio or designing and making something. On the other hand, concerns for reliability must be balanced by concerns for validity. Comparability of school-assessed work can never be an exact science and will always involve judgement and constant attention in striking an appropriate balance between competing demands. Maintaining this balance is assisted when there are procedures that routinely provide statistical checks on the effectiveness of the process.

References

Blackburn, J. (Chair) (1985). Ministerial Review of Post Compulsory Schooling: Report volume 1. Melbourne: VGPO.

Board of Studies (1993). VCE Study Design: English Units 3 and 4. Melbourne: Board of Studies.

Board of Studies (1994a). More About the GAT: The General Achievement Test 1994. Melbourne: Board of Studies.

Board of Studies (1994b). General Achievement Test Technical Bulletin. Melbourne: Board of Studies.

Brown, T., & Ball, S. (1992). A report on the VCE verification process. Melbourne: Victorian Curriculum and Assessment Board.

Cathie, I. (1986). Future Directions in Post Compulsory Schooling: A Statement by the Hon. Ian Cathie, M.P., Minister for Education. Melbourne: F D Atkinson Government Printer.

Cropley, M. (1981). Statistical Moderation: A Guide. Melbourne: Victorian Institute of Secondary Education.

Hill, P.W., Brown, T., & Masters, G.N. (1993). Fair and authentic school assessment: Advice to the Board of Studies on verification, scaling, and reporting of results within the VCE. Melbourne: Board of Secondary Education.

Karmel, P. (Chair) 1985. Quality of Education in Australia. Report of the Quality of Education Review Committee. Canberra: Australian Government Publishing Service.

Kreft, I.G., de Leeuw, J., & Kim, K.S. (1990). Comparing four different statistical packages for hierarchical linear regression: GENMOD, HLM, ML2 and VARCL. Centre for the Study of Evaluation, UCLA.

Masters, G.N. & Hill, P.W. (1988). Reforming the assessment of student achievement in the senior secondary school. Australian Journal of Education, 32 (3), 274-286.

McGaw, B., Eyers, V., Montgomery, J., Nicholls, B., & Poole, M. (1990). Assessment in the Victorian Certificate of Education: Report of a review commissioned by the Victorian Minister for Education and the Victorian Curriculum and Assessment Board. Melbourne: Victorian Curriculum and Assessment Board.

Prosser, R., Rasbash, J., & Goldstein, H. (1993). ML3-E - Software for three-level analysis (Version 2.3). Multilevel Models Project, Institute of Education, University of London.

Victorian Curriculum and Assessment Board (1987). Developing the Victorian Certificate of Education. Melbourne: Victorian Curriculum and Assessment Board.

Victorian Curriculum and Assessment Board (1988). Assessment for the Victorian Certificate of Education. Melbourne: Victorian Curriculum and Assessment Board.

Victorian Curriculum and Assessment Board (1992). Verification Manual 1992. Melbourne: Victorian Curriculum and Assessment Board.

Wright, B. D., & Masters, G. N. (1982). Rating Scale Analysis. Chicago: MESA Press.