EXTENDED COMMON ASSESSMENT TASKS IN VCE MATHEMATICS:VALIDITY, RELIABILITY AND OTHER ISSUES

David Leigh-Lancaster
Board of Studies, Victoria
<david.leigh-lancaster@dse.vic.gov.au>

Ken Rowe
The University of Melbourne
<k.rowe@edfac.unimelb.edu.au>

Since its inception in 1990, the Victorian Certificate of Education (VCE) Mathematics study has incorporated extended school based and assessed common assessment tasks (CATs) in its assessment regime. These tasks have been externally set by panel, graded by teachers and reviewed by various formal mechanisms. Key issues associated with the use of such tasks include validity, reliability, authentication and equity. We discuss several of these issues in this paper.

INTRODUCTION

The assessment regime constructed for the Victorian Certificate of Education (VCE) Mathematics study responded to the body of research and strong direction in mathematics pedagogy that supported the development of a broader assessment repertoire than one based solely or substantially on examinations. This regime was implemented to provide a robust picture of student learning and achievement. A unique feature of the approach taken in the VCE Mathematics study was to construct extended Common Assessment Tasks (CATs) of 2-4 weeks duration over a prescribed period, which would be set externally by panel (consisting of two or three distinct starting points based on a broad theme or context) and required the use of problem solving, modeling, or investigative approaches (see Leigh-Lancaster, 1995). These tasks would then be graded by teachers using set criteria and task-related advice provided on an annual basis. Teachers and students would be required to attest to the authenticity of student work (student ownership of unacknowledged work) and work from each school is externally verified. The first VCE Mathematics study allocated an equal weighting to each of the two school assessed extended CATs (CAT 1 - Investigative Project, CAT 2 - Challenging Problem) and the two examination CATs. This assessment structure had a significant effect on approaches taken to teaching and learning mathematics.

REVIEW

The VCE has 'lived in interesting times', a scenario which in various forms has been played out in the review of senior secondary curricula throughout Australia. The 1990's have been a decade of intense competition for access to positions in post secondary education or employment opportunities, with mathematics occupying a particularly significant area of the curriculum in these 'high stakes' contexts.

In the early years the VCE was reviewed on an almost annual basis (see Eyers, 1990; Northfield, 1991; Brown, Hill & Masters, 1993). Whereas the work of Brown and others found a high level of reliability for school assessed mathematics CATs, several issues were identified in association with the use of this type of school assessed extended CAT, namely, validity, reliability, authenticity, workload and equity (for the latter, see Leder et al., 1995). From 1994 VCE Mathematics assessment consisted of one school assessed CAT and two examination CATs, with each CAT weighted equally. Student (and teacher) workload and authentication had emerged as major issues for the school-based task, and this re-configuration addressed both of these concerns. For the remaining Mathematical Methods and Specialist Mathematics school assessed tasks, both now conducted over a prescribed period of 2 weeks, there was also an associated test, externally set, but teacher-marked according to a marking scheme provided by the Board of Studies.

The VCE was again reviewed in 1997, with a recommendation for assessment to move to a school assessed coursework and examination regime with essentially the same weighting of components. School assessed coursework will not include CATs or other tasks involving significant amounts of work over an extended period or work which is not directly supervised. Rather, it will consist of a collection of smaller, locally devised tasks from a specified selection of task types. The detail of school assessed coursework for the re-accredited VCE Mathematics study 2000 - 3 is currently being elaborated for a statewide information program in the second half of 1999. It is envisaged that this approach will provide for greater local control of teacher and student workload in this component of the assessment regime, and ease the systemic load of authentication requirements, while retaining significant aspects of the earlier CAT based approach within a more flexible framework. The Board is currently working with Cambridge University Press to produce a CD ROM resource, based on these CATs, to support teachers in devising suitable tasks for school assessed coursework.

VALIDITY, RELIABILITY AND OTHER ISSUES

Validity

The validity of extended mathematics CATs has, in general, been strongly endorsed by both mathematics educators and teachers. The CATs are widely regarded as valuable learning activities for students, with an acknowledged close relation between the relevant mathematics study design content, task design and context specification. This is due to several factors including the design brief for the task, the expertise of the setting panels and vetters, extensive vetting process key learning area manager, sitter vetters, study specialist vetters and NESB vetters. Moreover, there is a robust annual discourse on the extended CATs within the mathematics education and mathematics teaching community.

Authentication

Teachers and the Board have been keenly aware of the issue of authentication since the inception of the VCE. This has been dealt with by strict progress consultations, draft, logbook and journal requirements. Both student and teacher are required to attest to the authenticity of any unacknowledged work, with the teachers being required to not consider unauthenticated parts of the CAT for assessment purposes. Schools can require students to attend interviews to establish authenticity of sections of work as appropriate. The requirement for an interview occurs automatically if the student result on the test component compared to the project component is outside a specified discrepancy range. Student results for the CAT may be adjusted as an outcome of these processes. The general feeling amongst teachers, however, is that while these processes work effectively, they require a lot of time and effort, and can produce tension in the student teacher relationship.

Equity

Another important issue, which has received increased attention, is that of equity. Whereas the extended mathematics CATs appear to have provided female students with increased opportunity to demonstrate their mathematical capabilities (see Cox, 1996), there has been a critique of extended tasks on equity grounds based on likely access, or non-access, to resources across socio-economic groups. A notable aspect of this critique is access to supporting technology such as graphics calculators or other 'mathematically able software' such as computer algebra systems. Although the extent to which access to particular technologies may *enhance* the quality of student response within the context of this type of task is not clear (since this would depend on, amongst other things, the effectiveness of such use), differences in access to available technologies are likely to vary in schools both within and across sectors and also with demographic factors. The use of different technologies within all types of mathematics assessment is a pertinent and current issue, with a keen awareness of equity considerations in Board directions on this issue.

Reliability

Reliability analysis is essentially concerned with estimating how well an entity and its constituents have been measured. The assessment tasks of VCE studies provide a rich source of data for this type of statistical analysis. The following material is part of a comprehensive study (Rowe, 1998, 1999c) which places VCE Mathematics assessment within the context of reliability analysis for all VCE studies. Typically, an index of reliability (ρ_{rel}) is expressed in the form of a coefficient that is defined as the ratio of observed variance, σ_o^2 (less the measurement error variance, σ_e^2) to observed score variance, σ_o^2; or more simply, the proportion of observed variance that is true score variance. This coefficient is expressed as:

$$\rho_{rel} = \frac{\sigma_o^2 - \sigma_e^{2\,'}}{\sigma_o^2}$$

where values of ρ_{rel} range from zero, indicating total unreliability, to unity, indicating perfect reliability, or zero measurement error variance.

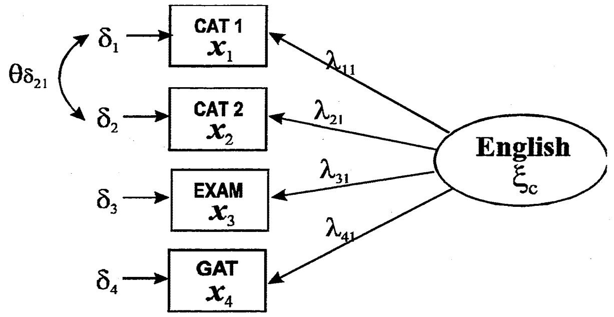
In the case of assessment measures related to the VCE since 1995 student achievement scores for each study have consisted of three measures that have been either two examination scores and a score on one school-based common assessment task, or scores on two school-based assessed CATs and one examination. As part of the VCE assessment procedures, a fourth measure of students' 'ability' has been obtained from their scores on the General Achievement Test (GAT). While student GAT scores are not included in the final achieved study scores, they are used as an administrative, quality control device to identify those school-based assessed CAT scores that appear to be either over-scored or under-scored, in relation to what could reasonably be predicted from GAT scores (see Hill, Brown, Rowe & Turner, 1997).

Estimating the Reliability of Assessments for VCE Studies

To estimate a reliability coefficient for each VCE study, confirmatory factor analysis was employed using LISREL (Jöreskog & Sörbom, 1998a) under a weighted least squares method of estimation and a listwise method for deleting missing data. In particular, the reliability coefficient was obtained from fitting a one-factor, congeneric measurement model to the standardised data for the four constituent measures (indicators) relevant to each study, based on a scaled variance-covariance matrix (and its asymptotic estimates) using PRELIS (Jöreskog & Sörbom, 1998b). Unlike traditional methods for computing reliability coefficients, this method makes use of factor score regression weights that minimise measurement error in the indicators contributing to each study and hence, maximise the accuracy of the reliability estimates. For explanatory research applications, the use of maximally reliable composite scores is crucial in fitting both single-level and multi-level regression models (Bryk & Raudenbush, 1992; Goldstein, 1995; Rowe, 1999a), as well as in fitting structural equation models (Bollen & Scott Long, 1993; Holmes-Smith & Rowe, 1994; McDonald, 1996; Rowe, 1999b).

This sort of model is illustrated diagrammatically for the VCE English composite (x_c), as shown in Figure 1. In this case, the assessment of VCE English since 1995 has entailed scores on two school-based assessed CATs (CAT 1 and CAT 2), one examination CAT, and a score on the GAT – each measured with error (δ_1, δ_2, δ_3, δ_4).

Figure 1
One Factor, Congeneric Measurement Model for VCE English



In matrix format, Figure 1 shows the regression ofx_i (i = 1, ..., 4) on ξ_c where the elements λ_{xi} are the partial regression coefficients of ξ_c in the regression of x_i on ξ_c, namely:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \\ \lambda_{41} \end{bmatrix} \begin{bmatrix} \xi_c \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \end{bmatrix}$$

$$\text{or} \quad x_\iota = \lambda_{xi}\xi_c + \delta_I$$

[1.1]

The model assumes that the vector of observed indicators (ξ_i) are measured independently, and implies that the covariance matrix of the ξ_i's is of the form:

$$\Sigma = \lambda_{xi}\lambda_{xi}' + \Theta_\delta$$

[1.2]

where Θ_δ is a diagonal matrix with elements $\theta_{\delta\iota}$ indicating the variances of the measurement errors δ_i (i = 1, 2, 3, 4). It should be noted from Figure 1 that a covariance has been specified between the unique measurement errors for CAT 1 (δ_1) and CAT 2 (δ_2), namely $\theta_{\delta 21}$. In such instances where the assessment procedures for VCE studies entail two school-based assessed CATs, the assumption of independence of the measures is not tenable. Hence, to adjust for what is commonly known as *method effect* (i.e, non-independent measures from the same source, or repeated measures from the same source), it is essential that the errors of measurement among such indicators are specified to co-vary. For those VCE studies assessed by two examination CATs and one school-based CAT, as is the case with VCE Mathematics studies, such covariances among the measurement errors are not warranted since all constituent indicators are measured independently. From the parameters of equation [1.2] the reliability (ρ_{rel}) of a composite (ξ_c) is given as

$$\rho_{rel} = \frac{w_C'\left(\hat{\Sigma} - \hat{\Theta}_\delta\right)w_C}{w_C'\hat{\Sigma}w_C}$$

[1.3]

where $\hat{\Sigma}$ is the estimated variance-covariance among the factor 'loadings' for the vector of congeneric indicators (ξ's), $\hat{\Theta}_\delta$ is a vector of unique measurement error variances among those indicators, and w_c is a vector of factor score regression weights that maximise the reliability of the composite. For specific details of these well-established procedures, see Brown (1989), Fleishman and Benson (1987). The rationale for this approach to computing composite variables and their reliabilities has more recently been detailed in Hill et al. (1993, 1996a) and Rowe and Hill (1997, 1998).

Results of Reliability Analysis

An excerpt from the results of the reliability analysis for 53 VCE studies undertaken between 1995 and 1998 are given in Table 1, which includes the assessment type in order of occurrence during the year (E for examination CAT and S for school-assessed CAT) as well as a mean reliability estimate over four years. The data has been sorted on this last category with those studies with higher reliability (and smallest measurement error variances) placed towards the top of the table.

Table 1
Excerpts of Reliability Coefficients from 53 VCE Studies: 1995-1998

Study	Assess. Type	1995	1996	1997	1998	Mean
1. Mathematical Methods	SEE	0.945	0.948	0.950	0.953	**0.949**
2. Accounting	ESE	0.947	0.950	0.947	0.949	**0.948**
3. Specialist Mathematics	SEE	0.945	0.949	0.940	0.952	**0.947**
4. Psychology	SEE	0.936	0.937	0.941	0.941	**0.939**
5. Biology	ESE	0.934	0.938	0.942	0.936	**0.938**
6. Chemistry	ESE	0.937	0.939	0.933	0.940	**0.937**
7. Physics	ESE	0.928	0.939	0.940	0.942	**0.937**
8. Physical Education	SSE	0.941	0.936	0.901	0.912	**0.923**
9. Economics	SSE	0.920	0.922	0.908	0.907	**0.914**
10.Further Mathematics	SEE	0.907	0.920	0.896	0.917	**0.910**
13.English	SSE	0.877	0.890	0.890	0.893	**0.888**
16.Info-Tech-Info-Systems	SSE	0.858	0.837	0.876	0.898	**0.867**
43.Art	SSE	0.788	0.807	0.834	0.836	**0.816**
Means (all 53 studies)		**0.843**	**0.850**	**0.849**	**0.855**	**0.849**

Comments on the Reliability Analysis

Of the 53 VCE studies examined, 20 were assessed on the basis of two examinations and one school-based assessed CAT, while 33 studies were assessed via one examination and two school-based assessed CATs. Twenty six studies (49%) had mean reliability coefficients in excess of 0.85, ranging from a high $\rho_{rel} = 0.949$ for Mathematical Methods, to $\rho_{rel} = 0.852$ for Materials and Technology, of which 11 studies were assessed via two examinations and one school-based CAT, and 15 studies were assessed via one examination and two school-based CATs.

In general, reliability coefficients $\rho_{rel} > 0.85$ are indicative of those VCE studies in which the assessment criteria for each of the constituent elements are both clearly specified and uniformly applied by assessors and examiners. Such studies are characterised by assessment components that are more likely to be similar in terms the nature of the assessed tasks, yielding high inter-correlations among their component scores. To illustrate this phenomenon, Table 2 gives an excerpt of corresponding data from the 53 studies in 1998. For example, the squared multiple correlation values, R^2, for the two examination components (CAT 2 Facts, skills and applications task - multiple choice and short answer format, and CAT 3 Analysis task - extended response format) of Mathematical Methods in 1998 were 0.864 and 0.922, respectively, and for the school-based assessed CAT 1 Investigative project, $R^2 = 0.626$. While this CAT required students to demonstrate knowledge of core content related to the Mathematical Methods course, there is greater assessment emphasis on student written communication and presentation skills of related mathematical working compared to the examination CATs.

VCE studies with reliability coefficients $\rho_{rel} < 0.85$ are more likely to have assessment criteria that are less clearly specified and not as consistently applied by assessors and examiners. Such studies are often characterised by assessment components that are more likely to be less similar in terms the nature of the assessed tasks, often yielding low inter-correlations among their scores, such as those found in the latter part of Table 2.

Table 2
Excerpts of Reliability Data from 53 VCE Studies: 1998

Study	Assess. Type	CAT 1 R^2	CAT 2 R^2	CAT 3 R^2	GAT R^2	ρ_{rel}
1. Mathematical Methods	SEE	0.626	0.864	0.922	0.338	**0.953**
2. Specialist Mathematics	SEE	0.575	0.882	0.912	0.288	**0.952**
3. Accounting	ESE	0.883	0.560	0.897	0.492	**0.949**
4. Physics	ESE	0.891	0.486	0.858	0.505	**0.942**
5. Psychology	SEE	0.568	0.864	0.868	0.627	**0.941**
6. Chemistry	ESE	0.888	0.484	0.856	0.457	**0.940**
7. Biology	ESE	0.859	0.531	0.852	0.635	**0.936**
8. Further Mathematics	SEE	0.401	0.827	0.824	0.473	**0.917**
9. Physical Education	SSE	0.444	0.446	0.890	0.559	**0.912**
10. Economics	SSE	0.485	0.508	0.881	0.524	**0.907**
13. English	SSE	0.640	0.660	0.786	0.683	**0.893**
16. Legal studies	SSE	0.584	0.577	0.814	0.558	**0.882**
37. Art	SSE	0.609	0.437	0.697	0.492	**0.836**
43. Music-History & Styles	SSE	0.624	0.662	0.490	0.356	**0.823**
Mean (all 53 studies)						**0.855**

The results which have in part been presented in this paper, indicate that the VCE assessment procedures have performed effectively in terms of reliability. For the 53 studies with the largest student enrolments between 1995 and 1998, the mean reliability coefficient was 0.849, indicating an improvement from 0.843 in 1995 to 0.855 in 1998. This is a strong result given the diversity across study types and content, in particular in the case of the VCE Mathematics studies. However, the magnitude of a reliability coefficient is not necessarily commensurate with validity – both content validity and criterion-related validity (see Allen & Yen, 1979). While it is possible to have a highly reliable assessment that lacks validity, a valid assessment that has low reliability is of limited value. The content validity of assessment tasks – including face validity and logical validity – may only be established via a rational analysis of their content based on professional judgment, albeit by consensus, which, in the case of VCE studies, is more properly determined by the respective accreditation and assessment panels.

SUMMARY

The preceding presentation indicates that extended common assessment tasks can be incorporated within a study to provide valid assessment as part of a highly reliable assessment regime. Quality tasks which are relevant, challenging and stimulating for both students and teachers alike can be devised by strongly constituted setting panels on a sustainable basis, with supporting appropriate and effective authentication and verification procedures in place. However there is a substantial commitment of resources and effort required by all concerned - the system, teachers and students. When this type of assessment takes place in a 'high stakes' environment, issues of workload, authentication and equity become highly significant, both in terms of perception and practice, especially when the same type of assessment is an integral part of five or six subjects studied in the final year of secondary education. It is timely for the Board of Studies to move on in terms of the school assessed component of its VCE studies. For Mathematics, the experience with extended CATs has provided a strong resource base for school assessed coursework in the re-accredited Mathematics study 2000- 3.

REFERENCES

Allen, M.J., & Yen, W.M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

Bollen, K.A., & Scott Long, J. (1993) (Eds.). *Testing structural equation models.* Newbury Park, CA: Sage Publications.

Brown, R.L. (1989). Using covariance modeling for estimating reliability on scales with ordered polytomous variables. *Educational and Psychological Measurement, 49,* 385-398.

Brown, T, Hill, P & Masters, G. N.(1993). *Fair and Authentic School Assessment.* Carlton, Vic:BOS.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park: Sage.

Cox, P. (1996). SEX and CATs: Findings from a detailed statistical Analysis. In Mathematics Making Connections *(Eds Forgasz, Jones, Leder et alia) Mathematical Association of Victoria Conference Handbook,* Brunswick.

Eyers, V et alia. (1990). *Assessment in the Victorian Certificate of Education.* Melbourne, VCAB.

Fleishman, J., & Benson, J. (1987). *Using LISREL to evaluate measurement models and scale reliability. Educational and Psychological Measurement,* 47, 925-939.

Goldstein, H. (1995). *Multilevel statistical models.* London: Edward Arnold.

Hill, P.W., Brown, T., Rowe, K.J., & Turner, R. (1997). Establishing comparability of Year 12 school-based assessments. *Australian Journal of Education, 41 (1),* 27-47.

Hill, P.W., Holmes-Smith, P., & Rowe, K.J. (1993). *School and teacher effectiveness in Victoria: Key findings from phase 1 of the Victorian Quality Schools Project.* Centre for Applied Educational Research, The University of Melbourne.

Hill, P.W., Rowe, K.J., Holmes-Smith, P., & Russell, V.J. (1996a). *The Victorian Quality Schools Project: A study of school and teacher effectiveness.* Report to the Australian Research Council (Volume 1 – Report). Centre for Applied Educational Research, Faculty of Education, The University of Melbourne.

Holmes-Smith, P., & Rowe, K.J. (1994). *The development and use of congeneric measurement models in school effectiveness research: Improving the reliability and validity of composite and latent variables for fitting multilevel and structural equation models.* Paper presented at the 7th International Congress for School Effectiveness and Improvement, The World Congress Centre, Melbourne, January 3-6, 1994.

Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika,* 36, 109-133.

Jöreskog, K.G., & Sörbom, D. (1998a). *LISREL 8: Version 2.20.* Chicago: Scientific Software International, Inc.

Jöreskog, K.G., & Sörbom, D. (1998b). *PRELIS 2: Version 2.20.* Chicago: Scientific Software International, Inc.

Leder, G, Rowley, G & Brew, C. (1995). Second Language Learners: Help or Hindrance for mathematics Achievement ? *Proceedings of the Regional Collaboration in Mathematics Education (ICMI) 1995*, Monash University.

Leigh - Lancaster, D. (1995). The Pragmatics of Change and Innovation - reflections on the Victorian Certificate of Education (VCE) Mathematics Study Design 1990 - 1994. *Proceedings of the Regional Collaboration in Mathematics Education (ICMI) 1995*, Monash University.

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

McDonald, R.P. (1996). Path analysis with composite variables. *Multivariate Behavioral Research, 31*, 239-270.

Northfield, J. R. (1991). *Meeting the Challenge: An Evaluation of the VCE Pilot Program after Two Years.* Monash University.

Rowe, K.J. (1998). *VCE Data Project (1994-1998): Concepts, issues, directions & specifications.* A research and evaluation project conducted under contract with the Board of Studies, Victoria. Centre for Applied Educational Research, The University of Melbourne.

Rowe, K.J. (1999a). *Multilevel analysis with MLn/MLwiN: An integrated course.* The 15th ACSPRI Summer Program in Social Research Methods and Research Technology, The Australian National University. Melbourne: Centre for Applied Educational Research, The University of Melbourne.

Rowe, K.J. (1999b). *Advanced structural equation modeling with Interactive LISREL: A Thematic integrated course.* The 15th ACSPRI Summer Program in Social Research Methods and Research Technology, The Australian National University. Melbourne: Centre for Applied Educational Research, The University of Melbourne.

Rowe, K.J. (1999c). *Reliability of assessments for VCE studies: 1994-1998.* A research and evaluation project conducted under contract with the Board of Studies, Victoria. Centre for Applied Educational Research, The University of Melbourne.

Rowe, K. J., & Hill, P.W. (1997). Simultaneous estimation of multilevel structural equations to model students' educational progress. Paper presented at the tenth International Congress for School Effectiveness and Improvement, Memphis, Tennessee, January 5-8, 1997.

Rowe, K.J., & Hill, P.W. (1998). Modeling educational effectiveness in classrooms: The use of multilevel structural equations to model students' progress. *Educational Research and Evaluation, 4 (4)*, 307-347.

Disclaimer

The views expressed here are those of the authors and do not necessarily represent the views of the Board of Studies.