Statistical Moderation and Social Moderation around Australia

Gabrielle Matters
Australian Council *for* Educational Research

Abstract

With the increased emphasis on teacher judgment in high-stakes assessment has come the role of social moderation in ensuring that student performances of equivalent standard are recognised as such before subject results can be recorded on a certificate with comparability guaranteed. With the widespread use of an index of overall achievement as the basis for offering university places has come the role of statistical moderation in ensuring that the results of different assessments (different subjects, different sites) are on a common scale before being combined to produce a tertiary entrance rank. Social moderation and statistical moderation are two of a variety of techniques for linking results of one assessment to those of another (Linn, 1993). And there are variations within each of these techniques in practice. Australia has eight different systems for senior curriculum, assessment and certification. The federal government has recently proposed the introduction of a single Australian Certificate of Education. The process of exploring a way forward has included an analysis of similarities and differences in current arrangements − including ways of validating teacher judgments and 'scaling' subject-group results. This paper outlines the theoretical underpinnings of statistical moderation and social moderation, and describes applications of these − form and purpose − in various Australian states/territories.

Introduction

Sometimes it is necessary to make the results of one test or set of assessment tasks comparable to those of another. This paper describes two situations in which comparability is desired; namely:

- Validating teacher judgments (in the case where school-based assessment is operating)
- Putting results onto a common scale (in the case where it is necessary to combine results in different subjects).

There are at least five different approaches to linking results from different assessments (Linn, 1993:83). The approach taken depends on the purpose being served. This paper describes two[1] approaches:

- Social moderation
- Statistical moderation.

Across the eight education systems in Australia, examples can be found of matches between the abovementioned purposes and approaches (forms). Of the three matches (Y) in the table below, this paper focuses on two of them (those in bold type), and alludes to the third one.

[1] Linn (1993) mentions three others: Equating, Calibration, and Prediction.

PURPOSE	*FORM*	
	Social moderation	Statistical moderation
Validation	Y	Y
Scaling	N	Y

It is important to make these distinctions early in the discussion because, in some conversations in some places, the verb ‘to moderate’ is used interchangeably with the verb ‘to agree’. But it is not as simple as that. Moderation is a *set* of processes designed to ensure that standards are applied consistently across teacher–assessors and across schools. The set of processes in social moderation is different from the set of processes in statistical moderation, and they are described below.

Outline

The structure of this paper is as follows. First, I define statistical moderation and social moderation and describe scenarios for the application of each form of moderation. Second, I give the background to the existence of eight different arrangements for curriculum, assessment and certification in Australia, highlighting two of the significant differences in assessment. Third, I use these differences as the basis for illustrating how comparability is desired and, therefore, where moderation is necessary. The illustration is in the form of a narrative set in the various states and territories of Australia. Fourth, I summarise, in tabular form, the application of different forms of moderation to serve different purposes across the country. The final section is not a conclusion but, rather, a glimpse into research being commissioned by the federal government into the comparability of standards across Australia.

Definitions and examples

In statistical moderation, which is sometimes referred to as scaling or anchoring, comparisons are made between results provided by different sources (e.g. teachers) or between results in different subjects (e.g. English, mathematics and history). Statistical moderation is used to adjust scores to make them ‘comparable’ (Linn, 1993:84). It is assumed that statistical moderation will remove from the scores the effect of group membership.

One example of statistical moderation in action is the use of a commonly applied (standardised) test to adjust for between-subject and/or between-school differences in school-based assessments. In this scenario, tests and assessment tasks are set and marked locally by teachers, and the standardised test is administered under controlled conditions and scored centrally. In another, slightly difference scenario, school-based assessments are scaled using external examination results.

Social moderation, which is also called consensus moderation, auditing, and verification, performances on distinct assessments are graded using a common framework and interpreted in terms of a common standard (e.g. the quality of student responses to assessment tasks in School A and School B are interpreted in terms of

the same statewide standards). It is assumed that performances of individual students and schools will be compared to a single set of statewide standards.

One example of social moderation in action is the use of peer review – teachers attend meetings to ensure that statewide standards in a particular subject have been interpreted and applied consistently across schools. In this scenario, like the previous one, schools develop their own sets of tests and assessment tasks in reference to a common content framework (or syllabus). Marking of student work depends heavily on professional judgments of teachers and a system of checks and verification. In another, slightly difference scenario, teacher judgments are reviewed by a panel of their peers.

The Australian context

Australia is made up of six states and two territories. All eight jurisdictions issue senior secondary certificates at the end of Year 12, the final year of schooling in Australia. The states are: New South Wales (NSW), Queensland (QLD), South Australia (SA), Tasmania (TAS), Victoria (VIC), and Western Australia (WA). The territories are: Australian Capital Territory (ACT) and Northern Territory (NT).

The notion of a federation of states is fundamental to understanding the existence of different systems across the country. Under the Australian Constitution, the states have responsibility for education (schools). The Constitution Star, the large star that stands aside from the Southern Cross on the Australian flag, has seven points representing the six states and one territory that existed when Australia became a federation of states in 1901. The emergence of public (government) education systems, which was preceded by some grammar schools and religious-order-owned schools, had begun before then − in the 1870s.

The Senior Secondary Certificate of Education is referred to by local titles at the state and territory level as follows:

ACT *ACT Year 12 Certificate*[2]
NSW *Higher School Certificate* (HSC)
NT *Northern Territory Certificate of Education*[3] (NTCE)
QLD *Senior Certificate*[4][5]
SA *The South Australian Certificate of Education*[6] (SACE)
TAS *Tasmanian Certificate of Education*[7] (TCE)
VIC[8] *Victorian Certificate of Education* (VCE)
WA *WA Certificate of Education*[9] (WACE).

The federal government recently proposed the introduction of a single Australian Certificate of Education in pursuit of greater consistency in senior secondary arrangements for curriculum, assessment and certification, more comparable student results across jurisdictions, and clearer and more consistent standards of student achievement. In a country with a relatively small and homogenous population, it is

[2] Also a separate *Tertiary Entrance Statement*
[3] Based on procedures of the Senior Secondary Assessment Board of South Australia (SSABSA)
[4] To be replaced by the *Queensland Certificate of Education* in 2008
[5] Also a separate *Tertiary Entrance Statement*
[6] Currently under review
[7] Also a separate *Tertiary Entrance Statement*
[8] Plus the *Victorian Certificate of Applied Learning* (VCAL)
[9] To be replaced by the new WACE by 2009

difficult to argue with the rationale[10] for national consistency. It is not the aim of this paper to discuss the pros and cons of a single certificate but, rather, to focus on some significant differences between jurisdictions in two aspects of the existing arrangements − procedures for ensuring comparability of standards in reported results and procedures for combining results in different subjects.

Different assessment arrangements

Two of the most significant differences in assessment arrangements across the country are: one, in the underpinnings of assessment and standards; and, two, in the calculation of rank orders for university selection. These differences (and many of the others that exist) are grounded in the history of the states/territories and their education systems and in the different sets of compromises that have had to be struck with their respective stakeholders.

Underpinnings of assessment and standards

There are differences in key assessment practices in the senior secondary years, with variations in the balance of external examinations and school-based assessments across and within states and territories.

In this discussion, the term external assessment is reserved for subject-specific examinations set by a body external to the school, as exemplified by the HSC in NSW. Such examinations are devised to assess student achievement in a particular subject, whether by objective-type or by conventional written, oral or practical questions. All the questions refer to a syllabus that has been defined by a group of educators (including teachers and/or examiners).

Internal (school-based) assessment is devised, constructed and implemented by schools, sometimes based on an official syllabus and accredited work program, sometimes not. Teachers have to be trained to become consistent judges of the quality of student work and there has to be a quality assurance process in place to guarantee comparability of results. A side-effect of such processes is that teachers engage in professional conversations about curriculum, pedagogy and standards. Thus social moderation delivers professional development for teachers, not only in assessment but also in discipline-specific knowledge.

There are no external examinations in the ACT. Queensland has operated a system of externally moderated school-based assessment since the abolition of public examinations (set by the University of Queensland) in 1972. The other jurisdictions have a combination of both, and there have been changes in the relative proportions of external and internal assessment over time. The current arrangements in NSW, for example, are 50% external and 50% internal. SA has 100% internal for Stage 1 (usually Year 11) and 50% internal for Stage 2 (usually Year 12).

[10] To reduce difficulties for students moving between states and territories; assist national reporting on student learning outcomes; identify essential learnings that prepare students for an Australian and global society; utilise to the maximum effect scarce curriculum resources; assist universities to develop teacher programs that are appropriate to all Australian students; reduce the new learning required of teachers who move between jurisdictions; stimulate the development of high-quality resources to support implementation; enable the articulation (and marketing) of what distinguishes Australian education.

Whatever their assessment regime, all systems recognise the value of using a range of assessment methods although different modes of assessment dominate in different jurisdictions. Assessment instruments used include formal examinations, written assignments, projects, practical work, oral presentations, aurals, end-of-semester tests, field work, and the creation of an artefact. The conditions of assessment cover the whole gamut: supervised or unsupervised, point-in-time or continuous, within prescribed dates or not, paper-based or computer-based, open- or closed-book, once-off or revisions allowed, and so on. Sometimes this variation is a function of the nature of the subject, sometimes it is a function of the nature of the assessment regime. For written assessments, the format of the assessment task might be multiple-choice, constructed response or extended response such as written expression or symbolic data manipulation. The wide range of assessment methods used is a response to the diversity of subjects now on offer, many of which do not lend themselves to point-in-time pen-and-paper tests.

Overall, there has been a change in emphasis over recent years with a shift towards assessment instruments that emulate the kind of process-based higher-order tasks thought to represent good practice.

At the present time, no national standards exist. At the state/territory level there has been an attempt to develop more explicit statements of achievement standards but those jurisdictions do not have the same way of expressing the standards for assessment and nor are those standards, however expressed, equivalent from state to state. For example, in NSW, the HSC provides detailed information about students' levels of achievement in relation to explicit standards and the cohort taking each subject. In QLD, standards descriptors for each exit level of achievement are published in the corresponding syllabus document.

In describing the processes for judging the quality of student work at the task/test/examination level and for grading student performance at the certification level, the states/territories use terms such as criteria-based, standards-based, or standards-referenced.

After the McGraw Report of 1996, NSW moved from normative towards standards-referenced assessment and reporting; that is, from assessing and reporting student performance relative to that of other students in the cohort to giving meaning to marks assigned to student work by referencing the image of the work to pre-determined standards of performance.

After the 1978 Review of School-Based Assessment (ROSBA) in Queensland, norm-based assessment was replaced with criteria-based assessment; that is, it changed from assigning grades according to the normal distribution to assigning grades after focusing on the specific nature of a student's actual achievements in relation to specific criteria.

Maxwell (2001:2) collapsed under five headings the various usages of the term 'standards' in relation to educational assessment and reporting. In summary, standards could be:

1. Moral or ethical imperatives (what students should do)
2. Legal or regulatory requirements (what students must do)
3. Quality benchmarks (what is expected of students)
4. Arbiters of performance quality (defining success or merit in student work)
5. Learning milestones (progressive targets for student learning).

Number 4 (defining success or merit) matches the definition of achievement standards (the 'how well' of a student's performance). It is the sense of how well that characterises the standards-referenced/based approach that now operates at the senior level in most jurisdictions. SA, for example, provides criteria for judging performance and performance standards in their Curriculum Statements, Parts I and II, respectively.

Calculation of rank orders for university selection purposes

For some jurisdictions the student's tertiary entrance rank (TER), which has various names across the country, does not appear on the Senior Certificate but on a separate document (such as the Tertiary Entrance Statement). The states/territories have different methods for compiling rank orders for university entrance, although the underpinning principles are similar.

The approach to university entrance in Australia is mainly based[11] on combining results attained by students in senior secondary school. Using success at the senior curriculum as a predictor of success at the tertiary level is not a universal practice. Alternatives include aptitude tests (not necessarily curriculum-based), lotteries, course-specific skills testing (e.g. manual dexterity for example for entry to dentistry), interviews, portfolios, and paying full fees. Some places that do use secondary achievement for tertiary selection take the grade point average without accounting for differences in subject-population characteristics or in intrinsic difficulty of subjects.

Two situations in which comparability is desired

This remainder of this paper focuses on differences between jurisdictions that relate to the validation of teacher judgments (in the case where school-based assessment is operating) and to the combining of results in different subjects (in the case where a statewide rank order list of overall achievement is used for selection to university).

Comparability of reported results

Comparability of reported results means that standards are applied consistently across sites (schools, regions) and across judges (teacher–assessors) so that student performances of equivalent standard are recognised as such (e.g. assigned the same grade). Thus, social moderation can be an appropriate response to the reliability challenge that invariably accompanies internal assessments. In social moderation, sometimes called consensus moderation, auditing or verification, the judgments of those teachers are ratified by others within the 'guild of professionals' − other teachers and moderators. Ratification of teacher judgments can occur through teacher meetings (e.g. all teachers in the ACT are able to attend moderation meetings), panel meetings (e.g. there are district and state panels in QLD), and visitations from a central office from a central office (e.g. as in SA, which also uses peer review). All versions of social moderation demand the development of a consensus on definitions of standards and on the performances that meet those standards (Pitman, O'Brien, & McCollow, 1999).

[11] Although there is an ever increasing multiplicity of pathways

Tertiary entrance ranks

One way of making the results of one assessment comparable to those of another before adding them together (to get a statewide ranking) is statistical moderation. WA, for example, applies statistical moderation to the numerical school assessments (teacher judgments) for tertiary entrance examination subjects.

Statistical moderation transforms every set of school marks onto a common scale. The process of putting different sets of results (assessment in different subject in different schools) on a common scale is called scaling. It is necessary because the populations selecting different subjects are not necessarily of the same general ability. Thus, through scaling, students are neither advantaged nor disadvantaged by the combination of subjects studied.

The common scale is provided by a measure that is common to all students in that subject-group – standardised examination marks. NSW scales against 'other-subject results'; that is, the common measure for scaling results in one subject is taken to be students' performances in their other subjects. An iterative process (sometimes called inter-subject scaling) is used to ensure that the distribution of students' results in a particular subject is aligned with the distribution of those students' results in their other subjects).

The technique of using external examination marks as a yardstick against which the achievement of each subject-group can be compared is only one technique. Another is to administer an anchor test to all students involved, thus providing a different yardstick against which the achievement of each subject-group can be compared. Whether the common measure is the external examination or an anchor test, it is the teacher(s) of the subject in each school who determine the rank order of students within that particular group.

Standardised testing and scaling

There are three examples in Australia of using an anchor test for scaling: the ACT Scaling Test (AST), Victoria's General Achievement Test (GAT), and the Queensland Core Skills (QCS) Test. The process of scaling involves a linear transformation in which the scaling parameters (measures of location and spread) are derived directly from a common scaling test such as the AST, QCS or GAT. The 'equivalence model' sets the mean and standard deviation (or mean and mean difference) of each set of school assessments to that group's mean and standard deviation (or mean difference) on the common scaling test.

The ACT and QLD are similar in that they are the only states/territories in which there are no external examinations. They are similar in that they both use social moderation for validating teacher judgments. They are similar in that they both use an omnibus test to produce the scaling parameters needed to 'iron out' differences between subject-groups and/or school-groups (both in the case of QLD) by transforming the distribution of school assessments to match the distribution of test scores. The scores can then be combined to produce a statewide rank order list for use in university selection. One difference between the ACT and QLD is that QLD records a student's individual test grade (A to E) on the Senior Certificate as well as using group scores for scaling purposes.

VIC is similar to the ACT and QLD in that it has a general achievement test (GAT) but it is different in that is has external examinations (as well as some school-based

assessments). The GAT is used to scale school assessments before they are combined with examination scores for inclusion in the Tertiary Entrance Score. It has other functions as well; for example, to check the accuracy of examination marking −the examination is reassessed if there is a significant difference between a student's examination score and predicted score.

And there is more

This short paper cannot do justice to the subtle differences between states/territories in conceptualising and operationalising statistical and social moderation. Nor can this short paper provide a complete picture of jurisdiction. Details can be found at the ACACA (Australian Curriculum Assessment Certification Authorities) website http://www.acaca.org.au. Details of the content and construct of the AST, GAT and QCS can also be found at that website. The table below summarise the history and uses of the three standardised tests. There is reference to ASAT (the former Australian Scholastic Aptitude Test) that was used by the ACT, QLD and WQ as a scaling test in the late 1970s through the 1980s.

	AST	*GAT*	*QCS*
State/Territory	ACT	VIC	QLD
Name of test	ACT Scaling Test	General Achievement Test	Queensland Core Skills Test
Used ASAT before	Yes	No	Yes
When developed in present form	1992	1993	1991
Individual results reported	No	No	Yes
Scaling device	Yes	Yes	Yes
Validation device	No	Yes	No

Summary of the application of different forms of moderation to serve different purposes across Australia

The following table summarises the application of techniques of social moderation and statistical moderation to senior assessment and certification across Australia for the purposes of validating teacher judgments (in the case where school-based assessment is operating) and putting results onto a common scale (in the case where it is necessary to combine results in different subjects and/or different schools).

Purpose	*Form*	*Technique*	*Examples*
Validation of teacher judgments	Social	Panels; Teacher meetings	ACT, QLD, SA, TAS, WA
		Visitation	SA, WA
	Statistical	Using external examinations	NSW, SA, TAS, VIC, (WA)
		Using other measures	NSW (other subjects) VIC (GAT)
Putting results onto a common scale	Statistical	Using a standardised test	ACT (ACT Scaling Test) QLD (QCS Test)
		Using external examinations	VIC (GAT)

NB: The contents of the above table are subject to further verification processes.

Comparability of standards across Australia

This conference paper has described some of the processes used to ensure comparability of standards *within* each of the eight States and Territories. It would be an interesting exercise to compare standards *across* Australia, *between* the eight jurisdictions.

The federal government has recently requested tender proposals for the provision of a 'Comparative Study of Selected Subjects for the Year 12 Certificate' (Department of Education, Science and Training (DEST), 2005). According to DEST (2005:2):

> At present there are significant differences across states and territories in the content, curriculum and standards of senior secondary school subjects. Recent debate has highlighted the need for greater national consistency. Arguments for greater consistency in education are based on four key themes: greater comparability of curriculum and student achievement standards, mobility of families, employer expectations and Australia's standing in the international education market.
>
> The current diversity in senior secondary curriculum structures and assessment regimes across Australia may be indicative of differing standards of achievement.
>
> …
>
> The study will investigate, compare and report on the content, curriculum and standards of study in English (including Literature), Mathematics, Physics and Chemistry for the Year 12 Certificate in every Australian state and territory.

Figgis (2005:27) made the following point about the task of making the senior years of schooling work for every young Australian. It could also apply to the task of making standards comparable across the country.

> The beginning is interesting, especially in that across Australia so many approaches are being tried. There is every reason to believe that the continuing journey will be, like so much in this era of ours of enormous technological, economic and social change, 'white water all the way'. One can foresee, for example, that a relatively small perturbation like the Commonwealth government's interest in a national Year 12 Australian Certificate of Education will have flow-on effects in all systems and schools.

References

Board of Secondary School Studies. (1978). *A review of school-based assessment in Queensland secondary schools* (ROSBA) (Report to the Government of Queensland; Professor E. Scott, Chairman).

Department of Education, Science and Training (DEST). (2005). *Request for tender: Comparative study of selected subjects for the Year 12 Certificate.* Canberra: Author. [retrieved from the Web, 24/12/05]

Figgis, J. (2005). *Changing senior school certificates: a story of visions and revisions*. In AAAJ Consulting Group (Ed.), Dusseldorp Skills Forum, Perth WA.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*(1), 83−102.

Maxwell, G. S. (1999). *Are outcomes standards?* Brisbane: Queensland School Curriculum Council.

McGaw, B. (1996). *Shaping their Future* (Report to the Government of NSW).

Pitman, J.A., O'Brien, J. E., & McCollow, J. E. (1999). *High-quality assessment: We are what we believe and do*. Paper presented at the 25th annual conference of the International Association for Educational Assessment. Bled, Slovenia.