STATISTICAL MODERATION OF TEACHER ASSESSMENTS

A report to the Qualifications and Curriculum Authority

John Wilmut and Jennifer Tuson

Commissioned by the Qualifications and Curriculum Authority.
QCA, 83 Piccadilly, London, W1J 8QA. United Kingdom.

Contents

Introduction

Project description

This report is a review of statistical moderation of teacher assessments, undertaken for QCA within the following general objectives.

i) To investigate technical issues relating to the statistical moderation of assessment results, with a particular view to development of a model for the statistical moderation of teacher assessment results based upon student performance in external examinations. This will involve:

- reviewing international evidence and experiences (past and present) concerning methods of statistical moderation designed to achieve comparability of teacher assessment judgements (between students in different schools)
- considering the wider literature on statistical moderation/linking/comparability, to explore potential implications for the statistical moderation of teacher assessment results (this will need to be selective rather than exhaustive)
- identifying exceptions to general statistical principles of moderation, as required to accommodate special circumstances (e.g., anomalies relating to small schools, apparent injustices to individual students)
- highlighting strengths and weaknesses of alternative methods, relating to the technical quality of moderation outcomes and to issues of feasibility/manageability in a UK context (e.g., financial, organisational, cultural, educational or political issues)

ii) To propose a single model of statistical moderation that would best suit the (QCA-defined) notional assessment system. This might require proposing modifications to the notional assessment system to facilitate more effective statistical moderation.

iii) To make recommendations for additional research required to establish the viability of the proposed model.

In relation to ii) QCA proposed focusing attention on two possible scenarios:

a) a GCSE in which assessment was 100% by teacher assessment. In this case, there might be an external reference test or some sort of examination, but this test or examination would function purely as a moderating instrument: the test/examination result would not be aggregated with the teacher assessment result. It would be helpful to consider the nature and implications (in terms of validity/reliability/dependability) of a range of external reference tests/examinations (i.e. aptitude, ability, curriculum).

b) a GCSE in which assessment was by 50% teacher assessment and 50% external examination (giving the opportunity to consider issues arising from the aggregation of TA/external assessment). It might be that an additional monitoring test can be used here and the review should explore its nature and implications. There might be a partial overlap, in terms of assessment objectives, between the teacher assessed component and the external examination (but not a complete overlap)

Acknowledgements

Our thanks for sharing information, mostly by email or telephone, to

Janet Keightley, Fiona Booker and Graham Benger (SSABSA)
Samantha Pickering (NSW Board of Studies)
Claire Robinson-Pope and Madeleine Jenkins (Victorian Curriculum and Assessment Authority)
Graham Baker and Alan Honeyman (Western Australia Curriculum Council)
Julie Bancroft and Brian Nott (Queensland Studies Authority)
Gerben van Lent (ETS Europe)
Jo-Anne Baird, Diana Fowles, and Martin Taylor (AQA)
Elizabeth Gray (OCR)
Eva Baker (CRESST, UCLA)
George Madaus (University of Boston)
Bob Adams (WJEC)
Alastair Walker (CCEA)
Rob van Krieken and Honor Murphy (SQA)
Ann Clegg (Edexcel).

Thanks also to QCA staff for advice and assistance with the work and to Robert Wood and Roger Murphy at CDELL for their inputs and support.

John Wilmut and Jennifer Tuson

Centre for Developing and Evaluating Lifelong Learning, School of Education,
University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham NG8 1BB
Tel: 0115 951 4496; Fax: 0115 951 4475
email: **cdell@nottingham.ac.uk**
website: http://www.nottingham.ac.uk/education/centres/cdell

August 2004

1 Classifying approaches to statistical moderation

The scope of statistical moderation

1.1 In their review of internal assessment Elley and Livingstone (1972) discuss the propositions that teachers, when making assessments can

> ..determine the relative achievement levels of their own pupils with reasonable reliability and validity

but cannot be expected to

> ..determine the standards of achievement reached by their pupils, relative to those of other schools, other years and other subjects groups

and that there is, therefore, a need for a method of moderation. Much later, and in a different context, the Task Group on Assessment and Testing (TGAT) report that heralded UK national curriculum assessment identified moderation as

> the process of bringing individual judgements into line with general standards … and is an essential part of an assessment system
>
> (Department of Education and Science/Welsh Office, 1988)

The methods available fall generally into three broad groups, identified by Cohen & Deale (1977) and Walker (1979a): inspection, statistical and consortium. They described what they saw as the desirable features of schemes of teacher assessment and moderation, namely that

- teachers should be given training, with opportunities for trial marking and discussion
- moderation feedback should be provided
- moderators should themselves have standardisation and monitoring
- moderation judgements should be made without knowledge of teachers' marks or grades
- double moderation is desirable
- sample sizes for moderation should be adequate to ensure comprehensive judgements
- statistical moderation against written papers as criterion should not be used as a sole method, but should trigger inspection.

Moreover, moderation procedures should be adapted to meet subject needs and assessment procedures and should not be constrained to fit the requirements of the method itself. Thus, a performance that does not naturally generate permanent evidence may require the moderator to be present or may require a recording to be made. This is not an area where consortium moderation is easy to manage (and it may be undesirable) and it may be difficult to find an external calibrator of sufficient relevance to enable statistical methods to be used. Equally, an expensive form of moderation such as the use of consensus meetings may not be justified for a small teacher-assessed component or where an acceptable statistical method is available.

1.2 In a more recent profile of moderation methods Daugherty (1997) discussed a typology that related these to a range of attributes: whether they were bureaucratic, contributed to professional development, took a lot of time, were costly, impacted on assessment process and impacted on assessment products or outcomes. Statistical moderation was seen as highly bureaucratic and having a major impact on assessment outcomes, making zero contribution to professional development and having zero impact on the assessment process, and taking small amounts of time and being cheap to operate.

1.3 Elley & Livingstone's propositions, which derived from their review of the relationships between internal and external assessment in several countries, led them to suggest a number of approaches to moderation, although their review of these approaches focused very strongly on statistical methods. Their classification of these methods is still relevant today, although some of the techniques now available to be used have become more sophisticated. All of the methods involved the use of an *external calibrator* (they called it a 'reference test'), that is, a measure that is independent of the teacher assessment but that may lay some claim to be an appropriate basis upon which to moderate it. Their classification (generalised from the specific New Zealand applications that they were principally discussing) involved the use of

- examination results
- aptitude and intelligence tests
- tests of general achievement and developed abilities
- what they called 'omnibus' tests of achievement
- item banks.

1.4 This range of possibilities will form the basis of the first part of our discussion although we will group the methods together under slightly different headings. Sometimes statistical moderation may involve the use of more than one of Elley and Livingstone's methods (that is, a calibrator may comprise several components) or a quality control procedure may use some statistical method alongside non-statistical approaches. We illustrate the methods with examples drawn from various periods and places and offer evidence of their effectiveness, where this exists.

Clarifying some basic concepts

1.5 Before starting this discussion we need to do some clarification. First, we know that terms other than teacher assessment are used on the ground, either interchangeably or alternatively – these include *internal assessment*, *coursework assessment*, *school-based assessment*, *school assessment* and so on. In a public examination context teacher assessment is taken to be the assessment that a school or college submits to the awarding body. This may be based on curriculum-based tasks set by the teacher (perhaps within an externally supplied framework or set of criteria) and also coursework, which is usually taken to mean that the tasks (with their accompanying assessment criteria) have been set externally. The former implies that work that may be spread over a period, perhaps at the teacher's discretion, whereas the latter is normally done within a specified limited window.

1.6 However it is arrived at, teacher assessment is then subjected to a process of moderation. If statistical moderation using an external calibrator is to be used aggregated teacher assessments need to be represented as a score; if this is not the case the assessments have to be transformed into a score. Generally speaking, statistical moderation is not viable when the result of the teacher assessment is on a binary (such as a pass/fail) scale, and a short grade scale (such as a five-point scale) is unlikely to support a satisfactory moderation process.

1.7 Unlike other moderation methods, statistical moderation deals with these scores and not directly with the evidence from which they have been generated. The nature of the evidence may influence the validity and reliability of the teachers' marks (for example, whether the teacher assessment is based on a single task or multiple tasks, and how these have been defined and managed) but the technical attributes of the moderation process itself are not dependent on the nature of the evidence, although the outcomes may be determined by it.

1.8 It is important to contrast *calibrating* with *monitoring.* Calibrating is the universal application of statistical adjustments to teacher assessments, on the basis of the external calibrator. Monitoring may be the use of statistical comparisons in a broad-brush way (perhaps via comparison of means and slopes of regression lines) to flag up possible lack of comparability or in order to monitor centre or awarding body longer-term behaviour (Wolf, 1996). Both are active roles (one in relation to the student and one in relation to the centre or awarding body) but evidently calibrating requires a more powerful instrument or battery of instruments to do what it has to do than is the case for monitoring.

Requirements for an external calibrator

1.9 There are some requirements that will attach to all types of external calibrator, all of which must be capable either of providing a basis for the adjustment of individual teacher assessment scores or of providing an alert that will trigger some further investigation. An instrument of this type might be a single calibrator or some combination of calibrators.

1.10 Elley and Livingstone identified three general requirements for such calibrators; these were that

- the calibrator results should correlate positively with teacher assessed scores in each subject in which it is used
- the calibrator should be such as to minimise undesirable backwash effects on learning and teaching
- the calibrator should appear to teachers and students to be appropriate in the context of each subject in which it is used.

1.11 In their *Moderating Instrument Research Project* Nuttall and Armitage (1983) looked at a mix of subject-specific and broad tests and other information about students (such as previous performance and age) to be used as an alerting device. They extended Elley and Livingstone's requirements by suggesting that a calibrating instrument should also

- operate equally across all approaches to teacher assessment that are to be moderated
- be free of particular bias (such as with respect to gender or ethnicity)
- be cost effective in development and administration.

In addition to these two sets of requirements the appropriateness of a particular method of statistical moderation, incorporating an external calibrator, will have to be judged in relation to

- its effectiveness in making adjustments to teacher assessed scores across the range of centres, candidates and subjects where it will be used
- its transparency and acceptability as a method including the demands that its use places on centres.

We shall also need to comment on the capacity of statistical moderation to support the development of teachers' professional competence in assessment.

1.12 It is clear that adjustments may be made to teacher-assessed scores for an individual centre irrespective of the correlation for that centre between the external calibrator and teacher-assessed scores. But the value of the correlation for any subject at any centre and on any occasion will indicate the suitability (and power) of the calibrator in providing dependable adjustments in which there is stakeholder confidence. In this respect the concern is with the individual candidate: it is not sufficient to have a high level of confidence in the average adjustment across a centre if there are large anomalies in the treatment of some individuals whose final grades will therefore be affected. The value of the correlation will be determined by

- the extent to which there are shared attributes assessed in the external calibrator and the teacher assessment
- the existence of centre-specific variables (such as the particular treatment of a coursework topic or a particular interpretation of a marking criterion)
- the extent of biases affecting sub-groups assessed in the external calibrator and teacher assessment
- the sizes of error variances in either external calibrator or teacher assessment.

1.13 When we observe a particular correlation it is generally impossible to say which combination of these factors has affected it and by how much. Moreover, correlations might vary widely between centres for the same subject and on the same occasion and it may be naïve to expect much more than very general consistency. This means that it may be unrealistic to set acceptable fixed boundaries on correlations although it has generally been accepted that their values will be higher in science, mathematics and similar subjects than in art, English and similar subjects. Moreover, there is a correlation 'ceiling' caused by measurement errors in *both* teacher assessments and the external calibrator that will vary widely from subject to subject but that may make it increasingly difficult to lift correlations above about 0.75 in some subjects without having extensive and very costly controls (such as multiple blind marking) over component reliabilities[1].

1.14 Smith (1978) does make the point that, in an examination setting, it is less important to know the overall correlation across all candidates than to know the weighted average of

[1] This is based on a limited amount known data from past and current examinations where, for example, reliabilities in English examinations have been quoted in the range 0.7 to 0.9 (quoted in Newton, 2003).

the correlations measured on a centre-by-centre basis. It is also important to know whether any centre correlation falls a long way below that average or whether the average itself fluctuates significantly from year to year. However, an external calibrator that generates a low correlation across all candidates will not generate a high average centre correlation, so that the general suitability of any calibrator can be determined from its overall correlation with the variable being predicted.

Requirements for teacher assessment

1.15 There are, of course, some requirements for the teacher assessment that is to be moderated using a statistical method and these will emerge in later sections of this report, particularly in Sections 4 and 5. There are some fundamental issues that we attempt to address; these include

- whether there is a minimum weight for a summative teacher assessed component below which moderation is neither appropriate nor cost-effective
- whether there are differences in the content of the work done by students that affect the outcome of the moderation process
- how far the structure and style of the assessment criteria or marking schemes affect the outcomes of the moderation process
- what levels of internal centre moderation should be required and whether this may need to be monitored or supported
- whether the provision of support may be combined with a moderation procedure as, for example, in the use of an item bank
- the extent to which the choice of moderation system is determined by technical requirements and limitations and how far it is a consequence of wider requirements for the summative assessment process
- whether post-assessment moderation is an appropriate process in any particular education system.

1.16 It is impossible to separate any technical discussion from these wider issues and modern summative assessment systems that are increasingly operating on a mass scale in very high stakes climates lead us to consider in much greater depth the relationships between the learning, assessment and certification processes and the social, political and economic environments in which they operate. Thus, for example, at the technical level we can ask whether there are situations in which we would not want to moderate. So while Smith (1978) made the point that it is

> probably just as unwise to place all one's faith in the moderating instrument and to adjust candidates' internally assessed marks in strict accordance with performance in it as it would be to accept the internally assessed marks without applying any kind of moderating technique.
>
> (Smith, 1978, p.26)

we might wish to ask under what conditions moderation as a process does not serve wider educational goals and statistical moderation in particular is incompatible with a set of requirements that we have for teacher assessment. In this context authors such as Strachan (1995), describing the approaches operated by the New Zealand Qualifications

Authority (see Section 5) are first describing in a detailed way their requirements for assessment in the classroom and only once those are established are they going on to discuss what forms of moderation are consistent with these requirements.

2 Varieties of external calibrator

2.1 Following Elley and Livingstone's classification we now examine a selection of examples of external calibrators and comment on evidence of the use of each. Not all have been used for statistical moderation purposes but all might be used or are typical of a type of instrument that might be considered.

Aptitude tests

2.2 Aptitude tests have generally been developed as predictors of future performance rather than as calibrators in a statistical moderation environment. They do have the potential to be used for moderation since they can clearly be applied across a range of occasions and subjects and should be equally applicable to all candidates.

2.3 The early 1970s saw an attempt to develop a Test of Academic Aptitude (TAA), which could be used to supplement A level grades for the purposes of selecting students for university admissions (Choppin *et al*, 1973; Choppin and Orr, 1976). It consisted of two parts, verbal (TAA-V) and numerical (TAA-M). For TAA-V, the correlations with university subjects (measured by end of first year performance) varied from -0.13 (Economics) to 0.30 (Psychology), while for TAA-M the variation was from -0.07 (Economics) to 0.22 (History). Not only were these correlations very low - after all, a correlation of 0.30 explains only 9% of the variation in a normal bivariate regression model - they 'bounced' around excessively (that is, were not reasonably constant from occasion to occasion). The TAA never got off the ground.

2.4 The TAA was a precursor of another test of developed aptitudes (a term used to describe tests not directly related to what has been taught in school), the International Test of Developed Abilities (ITDA). This was a test developed at ETS as a measure suitable for the selection of college entrants around the world. Some work has been done with it in this country (Fitz-Gibbon and Vincent, 1994). When combined with a vocabulary test it has produced correlations with A level arts and science subjects ranging from 0.20 to 0.50. These correlations are better than the TAA achieved, but they are still variable (between subjects and between occasions) and they are still rather low.

2.5 Both the TAA and the ITDA are *low relevance monitors* (Murphy, Wilmut and Wood, 1996) and therefore lacking in power when used to 'correct' grades for variations in the general ability of different candidatures because of the low and variable nature of the correlations obtained. Whilst they may be suitable for detecting group differences they do not provide a sufficient basis for the certification of achievement or selection of individual candidates. Thus, Elley and Livingstone (1972) reported modest predictive correlations for the Scholastic Aptitude tests used in the USA and Canada, suggesting that these, too, would not serve the purpose of individual certification.

2.6 More recently a report from the NFER on behalf of the Sutton Trust reviews all the aptitude testing initiatives in modern times including quite recent introductions in Sweden and Israel. Although it is not concerned with the use of these tests as calibrators in an examination situation it does provide confirmation that aptitude testing methodology is likely to be found wanting in predicting achievement at university. One of its conclusions reads:

> Evidence from the United States shows that overall prediction is modest, and can vary considerably between different institutions. This may well result from the extent to which colleges provide support for students who have difficulties adjusting to university.
>
> (McDonald, Newton, Whetton and Benefield, 2001)

2.7 Although tests of this kind can operate across a wide range of subjects, and it may be possible to develop one or more that would function within general qualifications, they may always suffer from some low correlations and will be difficult to justify as moderating instruments, since they lack face validity in relation to many subjects.

General ability tests

2.8 This is a group of tests that may be used as a reference for calibration in the context of a specific curriculum provision or qualification, rather than for prediction. They are not generally subject-specific and the two examples cited below were developed and used for standards comparability purposes rather than the certification of individuals.

NFER Test 100

2.9 In the early 1970s there had been interest in the Schools Council in using a general ability or reference test for use in monitoring the comparability of examination grades. The so-called Test 100 had been developed for other purposes[2] but was widely used in connection with the studies of the feasibility of a common 16+ examination in the early 1970s. In 1974 Willmott reported the results of correlating Test 100 with CSE and GCE grades.

- For CSE groups in 1973 he reported correlations with subjects varying from 0.20 (Art) through 0.39 (English) to 0.65 (Mathematics) with a mean of 0.42.
- For GCE groups in the same year the range was 0.24 (Art) through 0.47 (English) to 0.61 (Mathematics) with a mean of 0.45.

He also reported quite different correlation patterns for boys and girls.

- For the 1973 GCE group correlations for boys ranged from 0.20 (Art) through 0.49 (English) to 0.60 (Maths) with a mean of 0.44.
- The corresponding figures for girls were 0.29 through 0.56 to 0.62 with a mean of 0.50.

2.10 The combination of some low correlations and the gender differences meant that Test 100 was a low-relevance monitor for many subjects with the possibility that adjustments to teacher assessed scores would carry at least some gender bias. As it stood it certainly was not suitable for routine use as an external calibrator for adjusting teacher assessments and was never used for this purpose. The conventional wisdom since 1980 has therefore been that, in the context of the scrutiny of public examination standards,

[2] Test 100 was devised by NFER as a successor to its scholastic aptitude test CP66 which was used in the national comparability exercises carried out by Nuttall and Skurnik for NFER in 1965 - 67. CP66 and (later) Test 100 were used by researchers to link their results to NFER published data and to compare performances (such as in one or more schools) with national norms. Test 100 was then widely used in the 16+ CSE/GCE Feasibility Studies of the 1970s to check the alignment of grades on the two examination scales (see Nuttall, 1971; Skurnik & Hall, 1969)

general ability reference tests may perhaps best be seen as a cost-effective screening device: best used to discover where further scrutiny of standards may be worthwhile and to support professional judgments about the desirability and extent of remedial action (Nuttall, 1971; Willmott, 1980).

UCLES work on a calibration test

2.11 More recently Dexter and Massey of UCLES have reported on the use of a 'Calibration Test' for checking the comparability of the International GCSE (IGCSE) with GCSE. This is quite interesting because estimated correlations between this general ability test and examination grades were found to be quite high, often rivalling those found for subject-based tests of this length (note that the correlations reported in this study were between the test and examination results and that teacher assessment was not involved). For instance the correlations with the six most popular GCSE syllabuses 'sampled' were 0.74 (a Mathematics syllabus with coursework), 0.68 (Geography), 0.66 (English), 0.63 (Science), 0.61 (French) and 0.56 (History); these were not untypical. Higher correlations were observed with other syllabuses and a reasonable correlation (0.53) was even recorded with GCSE grades for Art and Design. Reliability (coefficient alpha for Form A) was estimated as 0.9 (Massey *et al*; 1998).

2.12 Yet Dexter and Massey too found significant gender-related differences in correlations and felt obliged to point out that

> ….the need to control for gender as well as general ability in applications involving the Calibration Test is thus very clear.
>
> (Dexter and Massey, 2000)

2.13 Because it is rare to come across an attempt at a calibration test it is worth quoting how Dexter and Massey described their test.

> The Calibration Test is not intended to break new psychometric ground. For the most part it uses item types often found in similar tests - requiring verbal, numerical and spatial reasoning. It is also speeded: students are asked to work quickly and it is not anticipated that all will find it possible to finish within the allotted time. The intention is to provide a brief, cost effective and reliable group test of general ability, readily available for large-scale use in research monitoring standards and investigating equivalence.

2.14 Although the evidence from this work does not relate directly to the moderation of teacher assessment and the test is used for group comparisons rather than as a calibrator it is sufficiently promising to suggest that some advance has been made over the use of an instrument such as Test 100. Perhaps a concerted effort to maximise the applicability of the test across subjects whilst controlling bias (particularly in relation to gender and ethnicity) would pay off.

Subject-based tests

2.15 We distinguish here between subject-based tests that have been developed specifically for the purpose of acting as an external calibrator, and are used as such over many occasions, and examinations which are part of the structure of the qualification and which are generally used on one occasion only. Although its use may present problems of security and may invite undue teaching to the test, there is an obvious attraction in using a calibrator that has proven characteristics and that largely eliminates between-occasion effects. However, because such tests impose an additional workload on all concerned they have had limited use.

2.16 Where they have been tried, subject-based tests have, unsurprisingly, generally performed better than aptitude and ability tests. The moderating instrument (or external calibrator) described by Nuttall and Armitage (1983) included subject tests alongside what they called a 'broad test' (that was more like the curriculum-based test described below) and also included a variety of other predictors, including past performance in relevant qualifications and the age of the candidate. This was in the context of a study of the moderation of the TEC level III science units, and the moderating instrument was designed to act as a national yardstick of standards and to pinpoint cases where the grades awarded to particular student groups were out of line with national standards. Nuttall and Armitage pointed out the significant resource implications in developing and using subject-based tests for at least each major subject (let alone for all subjects), so that the resources required to support the incorporation of this option in a moderating instrument would have to be carefully evaluated. They also highlight the difficulties using tests of this kind, on top of the demands of the assessment requirements for a particular qualification, of getting students to take them seriously.

2.17 Nuttall and Armitage explored the effectiveness of the various components of their moderating instrument by dropping them, one at a time, from the multiple regression and examining the efficiency of detection of what they called 'deviant cases' (these were student groups where, on the basis of alternative moderation procedures, there was known to be a departure from national standards beyond a threshold that normally warranted an adjustment). They concluded that

- the broad test, on its own or with just age, would be able to detect only about a third of the deviant cases
- the specific test, on its own or with just age, would be somewhat more efficient (detecting 60% of the deviant cases) as would prior performance plus age but without any test
- the best predictions came from a combination of prior performance and one or other of the tests and age, detecting almost 90% of deviant cases.

Consultations with practitioners did suggest a greater confidence in the use of external tests than measures of prior performance, whose use was seen to be less convincing despite the fact that they contributed to a significant increase in the detection of deviant cases.

2.18 Although Nuttall and Armitage did not identify a very clear advantage in using a subject specific test over using a broad test the advantage of the former would probably be that the correlations with teacher assessed scores would be higher than those on a general test (such as were developed in Dexter and Massey's exercise) and probably higher than those obtainable when using the examination as external calibrator. This is because we would expect that the commonality between teacher assessment and external calibrator that were both subject-specific would be greater than with a general test and that the psychometric characteristics of specially designed subject tests would be much better controlled than is possible with examinations. On the other hand, on cost and public relations grounds, it would be very difficult to justify the use of subject-based tests where there are already examinations available, but the use of a subject-based test as a moderating instrument where there is 100% teacher assessment may be a perfectly feasible option.

2.19 This instrument continued to be used primarily as a *monitoring* device to check on standards and alert BTEC to those centres, programmes and modules that needed investigation. It clearly had potential for use in any system that sought to accredit centres in the conduct of teacher assessment and Nuttall & Thomas (1993) envisaged that it might evolve into multi-level modelling based on centre characteristics. However, it is quite a big step from this type of use to one where general or subject-based tests, perhaps with other predictors, are widely used in, say, GCSE. Although there may be no technical reason why special tests should not be used, the increase in the assessment burden may not be acceptable and the higher levels of detection of the 'deviant cases' was only achieved with the insertion of a prior performance predictor into the regression. For GCSE it might be extremely difficult to find a prior performance predictor that would command confidence.

Examinations

2.20 Candidates' scores on a written examination provide the most obvious external calibrator against which teacher assessment in that subject may be moderated statistically. These scores appear to have the advantages of

- being readily available to the awarding body
- requiring no more effort on the part of centres than is already involved in the assessment process
- being specific to the subject in question.

2.21 We would expect the correlations between examination and teacher assessment scores to be higher than for aptitude and general ability tests but not as high as for subject-based tests. Elley and Livingstone (1972) report a range of median correlations for several school certificate subjects taken across between 11 and 13 schools as lying between 0.63 (physics) and 0.77 (chemistry), which is not untypical of correlations that have appeared in other contexts. However, they don't tell us how much variation there is between schools and the list does not include many arts or performance subjects. Correlations found in other contexts include: 0.75 (mean across 93 subjects) in New South Wales; between 0.35 and 0.77 for English language and between 0.44 and 0.86 for mathematics in Nigeria; above 0.86 in mathematics, between 0.35 and 0.75 in Chinese language and between 0.74 and 0.95 in English language in Hong Kong.

2.22 If examinations were to be the external calibrator it is likely that a good deal of the error variance in the correlations would be attributable to the examination as well as that attributable to the teacher assessment, if only because of the several components that comprise examinations. That much is known from studies of re-marking exercises (Wilmut, Wood and Murphy, 1996) and is an argument for using a fixed calibrator that is not subject to the variations that examinations suffer from year to year. We might begin to understand the properties and behaviour of a fixed calibrator in a way we never can with examinations.

2.23 There is, however, a substantial history of the use of external examinations as calibrators, particularly developed in the UK by the Joint Matriculation Board following its introduction of teacher assessed examination components in the 1960s. Smith (1978) and Forrest (1981) provide reviews of the Board's practices up to that time. Their requirements for external calibrators have already been identified and there is a more detailed discussion of the methods of statistical moderation that they used in Section 4 below.

2.24 JMB's statistical moderation at that time covered a wide range of subjects. Other boards experimented with the method but did not use it as extensively or routinely as JMB, and there was a widely held view that its disadvantages outweighed its advantages; the result was a general preference for methods of moderation using inspection with the statistical analyses, based on the examination, used as a monitoring or alerting device; the arguments were clearly presented by Murphy (1981), reflecting the earlier work of Cohen & Deale (1977) and Wilmut (1977) and now reflected in the current practices of many UK awarding bodies (see Section 6). In some other countries the calibration of teacher assessment using the examination as calibrator was adopted, and some of these applications are discussed in Section 5.

Curriculum-based tests

2.25 In the context of the modern curriculum there is a good deal of common ground followed by all students at a particular level either because of a common core of subjects or because of a common core of skills that are contained in all subjects. This provides a more secure basis for using a curriculum-based external calibrator that is based on this common ground since, generally speaking, the more variance that is shared between teacher assessment and external calibrator the more promising a reference test becomes:

> …a situation in which a larger proportion of the age-group was following a greater number of common subjects would make reference testing a more defensible proposition.
>
> (Wolf, 1996)

2.26 If the curriculum that is being assessed is organised around some core subjects or is made up of groups of modules that comprise a subject, with very limited exchanges of modules between subjects and some tightly controlled optional modules, then we will have a system that tends to a large amount of common material (that is, common content and/or common skills) across candidate entries within the certification of any subject. We may then be starting to satisfy the condition outlined by Wolf and it is possible that existing experience with basic and key skills tests may provide a starting point for this approach (but see the later discussion in Section 5).

2.27 In such a situation a curriculum-based calibrator that reflected and exploited that commonality might do a better job than aptitude or general ability tests; it would have a lot in common with the broad test devised by Nuttall and Armitage. The Queensland Core Skills test is an achievement test of this type rather than being an intelligence or aptitude test. It assesses the commonalities within the senior curriculum and aspects of educational performance common across a range of senior secondary studies by testing 49 Common Curriculum Elements that are the threads of the Queensland senior curriculum. Taken together, these are said to constitute a reasonable sample of the higher-order thinking skills expected in an educated senior student, including extended written expression involving complex analysis and synthesis of ideas, short written communication involving reading comprehension and basic English expression, basic numeracy involving simple calculations and graphical and tabular interpretation and solving complex problems involving mathematical symbols and abstractions.

2.28 The test comprises multiple-choice items, short-response items and a writing task. It contains stimulus material containing visual and spatial material as well as numerical, verbal, tabular and graphical content. We are told that

> .. as an almost invariable rule, QCS test performance reflects other achievement information: i.e. if a subgroup performs badly at school they perform less well on the QCS as well.
>
> (Nott, 2004)

There is some detailed further discussion of the operation of this test in Section 5 below.

More complex uses of statistical moderation

2.29 There is clearly some scope for combining tests into a more complex external calibrator that will function better in moderating teacher assessment. The effect is to increase the predicted variance in the regression of teacher assessed scores on scores from the external calibrator, although the resources needed to do this may be unacceptably large.

2.30 We have already seen that Nuttall and Armitage (1983) investigated a range of variables in developing their moderating instrument and ended up with a broad and a subject-specific test as the main components. They said that relevant information about a candidate or a measure of prior performance in the relevant area will, if added into the regression, always improve the prediction, although the gain may sometimes be small and the effort required too great or too intrusive to be worth doing. It may also be difficult to find measures that can be applied equally to all candidates and statistical moderation loses much of its attraction if it becomes expensive and time-consuming to use or is seen to introduce new inequalities.

2.31 The Australian State of Victoria uses its own general ability test, the GAT, in tandem with examination results to improve the teacher assessment - external calibrator regression equation. It uses it rather sparingly on a more or less opportunistic basis; that is, to capitalise on whatever variance exists with a particular teacher assessed subject or study as described in Section 5 below. A question here would be about the propriety of cashing in on, as it were, *windfall* variance and whether those enhanced regression equations would hold another time. As it happens, the Victoria authorities say that the GAT contributes little to the prediction (Robinson-Pope, 2004), which means that the rest of the external calibrator (the examination marks) are doing the calibrating job most of the time.

2.32 We should note again that current practice in the UK is largely to use statistical methods in tandem with other moderation methods. Formerly the statistical moderation procedure was seen as the basis for most adjustments in some Boards and moderators only intervened with inspection methods if a particular centre fell outside certain pre-determined parameters. Increasing experience from the mid 1960s with teacher assessment and with inspection and consensus moderation systems opened up a vigorous debate between the advocates of the different methods that eventually led to a reduction in interest in statistical methods (and a reduction in research into these methods *in public examination settings*) so that moderation by inspection (perhaps on a sample of centres or candidates) has become the more standard first step which can lead to statistically-based adjustments (almost always based on linear scaling) where centres are seen to be ‘in line’ with moderators’ expectations, but to a full moderation by inspection where this is not the case. This approach does have the advantage of greater transparency with a reduced risk of anomalies going unnoticed, and is capable of providing a degree of support for those teachers and centres that need it most.

3 Using item banks

3.1 One of Elley and Livingstone's (1972) approaches to moderation involved the use of item banks. This may be regarded as a statistical moderation process that uses an external calibrator but it is based on a provision on the supply side of teacher assessment that forms the basis for moderating teacher-assessed scores. In that respect it differs from other statistical moderation processes that are solely concerned with post-assessment mark processing.

3.2 Improved methods for pre-testing questions (particularly objective items) and calibrating them on the basis of the scores obtained enabled banks of materials to be developed; initial work in the 1960s (see Wood & Skurnik, 1969) was strengthened by the use of a variety of item response models for calibration, and this work has continued, though rather less enthusiastically in the UK than elsewhere. What Elley and Livingstone proposed was that calibrated question materials could be made available to teachers for incorporation in their own assessments. This would provide underpinning for their role as assessors and (providing that the use of the material was reasonably consistent across centres) a basis for statistical moderation based on these materials as external calibrator. Nuttall and Armitage (1983) returned to this idea in their discussion of their moderating instrument, but did not implement it in any of their work. Where teacher assessment is to be based on a large number of curriculum-embedded tasks teachers may find great value in being able to import materials from a bank. However, where these were going to form the basis for moderation this selection might not be able to be entirely in their hands although the increasing sophistication of their availability through the Internet might make this more feasible than in the past. But as Elley and Livingstone point out, the bank would have to be robust and well stocked, with reliable calibration if the materials were to be available on a sufficient scale to support teacher assessment in a large entry subject. We may, however, be getting closer to realising this approach with wider experience of banks and tests that can be downloaded electronically and current work in developing on-screen public examinations.

3.3 This is the only approach to statistical moderation that might seriously counter the criticism that the method does not form any basis for the professional development of teachers. Whether the costs that would be involved are acceptable is not clear, nor is it evident whether the support that it would give to teachers is better or worse than that coming from moderator interventions in a centre. There is also the danger that the tests will be seen as the 'standard' for the whole of the teacher assessment which may become benchmarked to performances on the limited range of objectives covered by the test (see, for example, the discussion on national curriculum assessment in Reeves *et al*, 2001).

3.4 There is also an emerging literature that considers the effects of importing summative assessment devices into teaching programmes, particularly if these are regarded as, in some way, a contribution to formative assessment. Black (2002) reports that teachers in the UK are struggling to make sense of this interface in the context of national curriculum assessment and that there is a general lack of confidence in conducting assessments under these conditions. The danger in using summative assessment materials from a bank solely for moderation purposes might compound this problem although there does not appear to be a problem with the use of the common assessment tasks that are embedded into the curriculum in Victoria, Australia and there are similar developments in Scotland that are outlined in Section 6 below. This type of use of external summative assessment materials embedded in the curriculum needs to be discussed against the background of evidence from the post-16 sector where students and teachers now take a much more focused and pragmatic approach to securing required levels of achievement in curriculum-embedded tasks that contribute to qualifications (see, for example, Ecclestone & Pryor, 2003; Weeden & Winter, 1999; the further discussion in Section 7 below).

4 Technical issues and some research evidence

4.1 In this section we discuss the mechanisms available for conducting statistical moderation, followed by a brief summary of some further relevant evidence relating to the conduct and moderation of teacher assessment. This anticipates Section 5 where we identify the conduct of some specific applications and outcomes of statistical moderation methods.

Different forms of statistical moderation

4.2 The general methods of statistical moderation (whatever type of external calibrator is used) were developed in the 1960s and 70s and are summarised by Smith (1978) into two main categories:

- methods of scaling; Forrest (1981) identified three approaches all of which are based on simpler or more complex applications of *linear regression scaling* and to which (for our present discussion) should be added non-linear methods such as *equipercentile scaling*
- methods of mapping, which we can regard as *rank order scaling.*

These two authors with Wood (1991) discuss these methods at some length.

Linear regression scaling

4.3 All forms of linear regression scaling are based on the use of a simple regression equation in which the mean and standard deviation of teacher-assessed marks are adjusted on the basis of the distribution of marks on the external calibrator and the correlation between the two measures. These are the three scaling methods described by Forrest and they were also discussed by Backhouse (1976) in work on aligning the marks of differentiated examination papers. Reduced forms of this process adjust only the mean of the teacher assessed scores (that is, assume that the standard deviations of the two sets of marks are the same and that the correlation is 1) or adjust the mean and standard deviation (thus assuming only that the correlation to be 1). This last possibility is commonly called *linear scaling*. The rank order of the teacher-assessed marks is preserved in all cases except where the correlation is negative, in which case the rank order will invert. In this form of scaling no transformation is applied to the teacher-assessed marks in order to induce a linear relationship between these and the marks on the external calibrator. However, as the correlation approaches zero the adjusted teacher assessed marks approach their mean, reducing their weighting when added into marks on other examination components.

4.4 Good and Cresswell (1988) examined the use of linear scaling in the context of their work on differentiated examination papers and pointed out that a straight line may not correctly represent the relationship between two sets of marks, although their work was influenced by floor and ceiling effects in differentiated papers that may not occur with the statistical moderation of teacher assessments. Despite this they regard linear scaling as a better option than not scaling at all, although they do place considerable value on judgemental scaling processes. These could be particularly effective if backed up by routine presentation of the consequences of different judgements so that examiners could immediately see the effects of the decisions that they make (an approach originally developed by French and others in their Decision Analytic Aids To Examining project).

4.5 Apart from the uncertainties about the suitability of the process that may follow low correlations between the external calibrator marks and the teacher-assessed scores, there may be problems associated with skewed mark distributions, particularly where the skew of the marks on the calibrator is opposite to the skew on the marks on the teacher assessment. This suggests that there is a non-linear relationship between the two sets of marks that is quite commonly seen (left-skew is often associated with generous marking, pushing marks towards the top of the scale, as reported, for example, by the State of Victoria; Robinson-Pope, 2004) and that that may come from centres with small candidate numbers. Awarding bodies have often adopted special procedures for dealing with small groups (which could be so commonplace in some subjects as to make statistical moderation a practical impossibility) and it is theoretically possible to apply a statistical transformation to the marks that will force a linear relationship, but this might be very hard to justify publicly (but see the discussion of practices in New South Wales in Section 5 below).

4.6 Of course, the skewing of teacher-assessed marks towards the top of a scale may not be a process that is equally applied to all students. Spear (1989) showed that the propensity of teachers to over-mark students' work increased as the student's ability dropped, leading to a bunching of marks towards the upper-middle range of the distribution. This is not, however, a sufficient basis for developing a model that would allow a systematic transformation to be applied to all teacher assessment in order to enable a linear model to be used, making en empirical application of a best fit to each set of scores (as in New South Wales) the only practicable option. Whether this is acceptable to stakeholders may vary from setting to setting.

4.7 Linear scaling assumes that the teacher assessment and the external calibrator have equal reliability. Good (1988) discussed the use of what she called structural linear scaling (actually structural regression) that modelled errors on both measures, based on a range of assumptions, and sought to choose the best available regression line. Unfortunately there is no clear-cut best choice since different candidates are disadvantaged by different decisions. In particular, it was candidates at the extremes of the distributions whose marks were most markedly affected by the choice of assumption (such as whether there was error on one or both variables and how it should be modelled). Structural regression does not appear to have been used operationally.

4.8 We can regard multiple regression methods as an extension of simple regression methods, the most obvious of which use more than one component in the external calibrator (or independent variable). The effect of improved correlations is to reduce the number of candidates whose scores fall a long way from the regression line but there is no guarantee that some will not continue to be disadvantaged. The Nuttall and Armitage research is an example of this approach but there are few other cases where it has been used though modern computing power makes it more possible than ever before. In an unpublished report of a piece of development work in the late 1970s Wilmut (1977) discussed a regression model that used centre identity as a variable and showed with simulated data that this was capable of delivering better alerts or moderation than other available methods. However the computing power required and the lack of transparency in the method put it beyond the available resources and demonstrated once again that the potential difficulties with statistical moderation methods are less to do with technical power than with credibility.

4.9 Linear scaling that assumes a correlation of unity remains the most common method of statistical moderation, chiefly because of its transparency. It is probably more justifiable in this application than in the differentiated papers application because there is no skewing of distributions due to floor and ceiling effects, although skewed distributions do commonly occur. It is the method that Good and Creswell settled on in their work on differentiated paper equating.

Equipercentile scaling

4.10 This is an example of *curvilinear regression scaling*. There are many possible ways of representing non-linear relationships between sets of marks, and of scaling to create a linear function, but a common example is where the marks or scores on the teacher assessment and the external calibrator are said to be comparable if they are reached by the same proportion of a given group of candidates. Points representing pairs of scores or marks are plotted and a smoothed curve is drawn between them.

4.11 In the version implemented by the State of Victoria an 'external score' is created by regression of the set of scores from the external assessments on the raw teacher assessment. Equipercentile scaling is then applied by matching (pinning) at the maximum, the 75th percentile, median, 25th percentile and at the minimum, where a 'true' minimum is established by 'cleaning up' the data in an appropriate fashion. Details of the procedures used in Victoria are in Section 5. Data cleaning is a process that involves the removal of outliers, normally at the extremes of a distribution, since these have a disproportionate effect on any scaling methods, particularly in small data sets. These are sometimes described as 'flop scores' (see below) and their removal or adjustment is justified on the grounds that one or other of the performances of the candidate in question is untypical. Some criteria will be needed to judge which are outliers, and the process clearly carries some risk.

4.12 Vernon, in comparing the scaling methods for use in secondary school selection reckoned that equipercentile scaling was to be preferred (Vernon, 1957, cited in Wood, 1991). It was, he thought, as technically sound as any other, and easier to implement. It is fairer, too, in that scaled scores more nearly reflect the intervals between candidates, which is the objection to simple rank conversion. And, of course, it is not necessary to assume that a straight line describes the relationship between scores on the components. However, he appears to have neglected the disproportionate effect of single data outliers within small candidate groups.

4.13 Good and Cresswell (1988) demonstrate (in the context of a study of differentiated paper scaling) that there is a risk that achievement in one paper cannot be placed appropriately on the distribution of marks for the other because of the compression of mark scales at one end of the distribution or the other ('floor' and 'ceiling' effects). They believed that linear scaling was safer in these applications since equipercentile scaling could bunch the scaled marks at one end of a distribution. They also preferred linear to non-linear scaling methods because they argued that non-linearity was due to the same 'floor' or 'ceiling' effects generated in a differentiated paper that was either of an inappropriate standard or that had been taken in error by some candidates, the underlying abilities of the different groups of candidates actually being linearly related.

How far this conclusion may be applied to the moderation of teacher assessments is unclear; similar bunching and non-linearity are frequently evident, but there is no reason to suppose that the two mark distributions should have similar shapes.

Rank order scaling

4.14 Smith (1978) describes the technique of mapping based on rank orders, applied to the moderation of teacher assessments, but the method also appears in some of the work being done on the use of differentiated examinations, where there is a need to scale the results from two or more alternative papers where the marks are to be combined with a common paper taken by all candidates. Backhouse (1976) discussed this latter problem as part of the work arising from the feasibility studies into a common examination that would replace CSE and GCE O level examinations.

4.15 When this method is used for moderation the means and standard deviations are automatically equated because the distribution of teacher-assessed marks is deliberately made the same as the distribution of those for the external calibrator. The rank ordered teacher-assessed marks are replaced by scores on the external calibrator distribution, the highest ranking student receives the highest score, the second highest ranked the second highest score, and so on (Smith provides a detailed explanation and diagram of the method). Backhouse acknowledged that the method was vulnerable to the 'company you keep' factor where the performance of some candidates might affect the outcomes for others because the ranks did not provide information about the mark intervals in the distribution. The effect is most marked for candidates at the distribution extremes and is likely to have the greatest effect when the group size is small, and Wood was strongly critical of the method for this reason (Wood, 1978); see also Section 7 below.

4.16 Wood also emphasises the care that has to be taken at the bottom of the distribution where unexpectedly low teacher-assessed scores, what have been called 'flop' scores (McIntosh, Walker and Mackay, 1962, cited in Wood, 1991; Walker, 1979b), will be allocated to the lowest ranked students unless something is done to stop this happening. The difficulty is that it can never be known with certainty which are the 'flop' scores and these may, in fact, also occur at the top end of the distribution where candidates exhibit unexpectedly high teacher-assessed scores, and have similar effects. In this respect the term 'flop' is unfortunate since a performance which is seen to be anomalous may occur anywhere is a distribution and may be uncharacteristically high as well as uncharacteristically low. It may also occur for a perfectly legitimate reason such as a situation where a teacher has properly rewarded the high coursework achievement of a student whose examination performance is low.

4.17 The problem is obviously most acute with small groups but there is clearly some concern with a method that 'throws away' the information about the intervals between candidate scores. Much more significantly, Vassiloglou & French (1982) discussed the applicability of Arrow's Impossibility Theorem (first proposed in 1951 in the context of social choice theory) to examination assessments where the marks obtained on an examination are derived solely from the candidates' rankings on its component parts.

This theorem presents six properties or axioms that all appear to be essential, but states that there is no function that will satisfy them all at the same time[3].

4.18 Vassiloglou & French examined each of the axioms in relation to examination assessment, particularly using a differentiated paper model, and reworking data previously analysed by Backhouse (1976) and Wood and Wilson (1980). They took account of a wide range of earlier research that had sought and failed to disprove Arrow's Theorem or to show that a weaker set of properties would be sufficient to enable results based entirely on rank orders to be generated and tested each of the axioms in an examinations context. They concluded

> that it is impossible to assess candidates consistently by only taking rank orders into consideration

and they asserted that, because Arrow's Theorem could be shown to be applicable, this process was bound to fail with differentiated papers. This means that, in practice, every scheme of combination of ranks that can be developed has the potential to be unfair to some candidates. There seems to be no reason to suppose that it will work any better with statistical moderation of teacher assessment and Kingdon *et al* (1983) commented that

> Arrow's Theorem applies to any method based on rank orders

and went on to demonstrate (using data similar to that used by Backhouse and by Wood & Wilson) the inferiority of rank order scaling when compared to linear scaling and other parametric methods; on the basis of this work Good and Cresswell (1988) regarded the method as unsatisfactory and did not use it in their work on grading differentiated papers in GCSE.

4.19 Other methods of mapping that are directly based on relating mark distributions are possible but appear not to have been used in operational examinations. Kingdon (undated) describes a method that he calls *mapping ranges* where the range of teacher-assessed marks is mapped onto the range of external calibrator (examination paper) marks so as to preserve the ranks and relative mark intervals. This appears to have been tried out on an experimental basis by the London Board; however, on the basis of trials with simulated data, it was found to offer no advantages over linear scaling (in that it did not reduce the number of cases where candidates' grading was adversely affected by the moderation process) and was vulnerable to the effects of non-typical extreme scores.

Some research evidence on teacher assessment for summative purposes

4.20 Wilmut (1999; 2004) has reviewed research evidence relating to the conduct of teacher assessment. Reviews of this type do provide essential context for a discussion of statistical moderation, whose potential for use in public examinations cannot be determined solely (or even mainly) on technical grounds. It is difficult to identify those aspects of general research on teacher assessment that should be included in the present

[3] Details of the 6 axioms are complex and need to be reviewed in the original paper. In relation to examinations they relate to non-triviality of the case (at least 3 candidates and 2 components), weak ordering of ranks (transitivity), universality of domains (an overall ranking is possible whatever the component rankings), independence of irrelevant alternatives (such as the presence of other candidate groups), the Pareto principle (a ranking that appears on all components will appear in the overall ranking) and non-dictatorship (all components play a role in the ranking of the students).

discussion and those that should not; what follows is a small selection of issues thought to be immediately relevant but which should be seen in the wider context of all work in this area.

The EPPI review of teacher assessment for summative purposes

4.21 A recent EPPI review of the evidence of reliability and validity of assessment by teachers used for summative purposes was designed to address the main question:

> What is the research evidence of the reliability and validity of assessment by teachers for the purposes of summative assessment?

and the subsidiary questions:

> What conditions affect the reliability and validity of teachers' summative assessment?
>
> What are the implications of the findings for policy and practice in summative assessment?
>
> (Harlen, 2004)

4.22 The review methodology[4] involves the systematic searching for, appraisal and selection of evidence sources related (in this case) to assessment by teachers for summative purposes. The findings of the review are based only on those sources that meet a range of criteria relating to relevance and methodology. Given the topic of this particular review the sources relate to research of teacher assessment undertaken in a range of different contexts (countries, student ages, occasions, purposes etc) so that generalisation is difficult.

4.23 Harlen characterises evidence in relation to the reliability and validity of teacher assessment in different subjects as 'mixed'. Differences between subjects in how teacher assessment compares with standard tasks or examinations results have been found but there is no consistent pattern suggesting that assessment in one subject is more or less reliable than in another.

4.24 These conclusions are in line with what we would expect. There is no reason to think that correlations will come up the same in the US and the UK, in national curriculum assessment and in GCSE, or when comparing 1990 in one place with 2000 in another, and so on. Context effects must surely play a part. In any case correlations have a distribution commensurate with sample size, so a range would always be expected even with correlations calculated at the same time and place and involving the same subjects.

4.25 Harlen comments:

> It is important for teachers to follow agreed procedures if teacher assessment is to be sufficiently dependable to serve summative purposes.
>
> The training required for teachers to improve the reliability of their assessment should involve teachers as far as possible in the process of identifying criteria so as to develop ownership of them and understanding of the language used.

[4] Described on the website of the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) http://eppi.ioe.ac.uk/EPPIWeb/home.aspx

4.26 This of course makes a lot of sense - the more support mechanisms there are the better. It is why any ongoing interventions where teachers are involved, like the QCA project on monitoring English at key stage 3, are so important. Section 6 tells something of what the UK awarding bodies have already put in place in terms of support mechanisms and here is an example from Australia. The Western Australia Curriculum Council offers teachers a package which includes: a school moderation visit, consensus moderation, assessment tasks which teachers can use to practise on, district seminars where teachers can compare notes. In addition, there are education officers designated 'accreditation and moderation' who are available to teachers. It is only when teachers have utilised as much of this support as they wish to take up that statistical moderation is implemented.

The reliability of teacher assessments and external calibrators

4.27 Evidently it is pointless just looking at the technical properties of teacher assessment when in a statistical moderation environment the reliability and validity of the external calibrators matter equally (if not more so because there is more you can do about them when developing the instrument). Unreliability in tests used as external calibrators certainly introduces error into the scaled scores. In a Monte Carlo study based on the Australian Scholastic Aptitude Test Sadler (1992) estimated the seriousness of the errors introduced into scaled scores for various student group sizes and student abilities. He showed that these errors, arising from the unreliability of the calibrator, were serious for small groups and for students at the upper extremes of the score scales. This is a potential problem that was highlighted by in the earlier Queensland Grade 12 Study by McGaw (1977) although he felt that alternative approaches would result in less rather than more fairness. These and other issues relating to small groups, anomalous scores and the 'company you keep' factor are discussed further in Section 7.

4.28 The successor to the ASAT in Queensland, the Core Skills Test, had, in 2003, an estimated value of Cronbach alpha of 0.88 across all four subtests. In the subtests it was 0.90 for multiple choice and 0.83 for short response; these numbers are reported to be typical for the test (Nott, 2004) and are respectable.

4.29 Clearly the nature of the calibrator is going to be important; given what we know about examination reliability an open-response essay test used as a calibrator may give more trouble than a more structured or closed-response form (for discussions of examination reliability see Wood, 1991 and Wilmut, Wood & Murphy, 1996). It would be impossible for awarding bodies to develop a full analysis of the characteristics of every examination in their stables, but there has been a wide range of studies, on a selected basis, as part of the routine monitoring of examination performance. The inviting possibilities of generalisability analysis, which enables meaningful statements to be made about the dependability of results over varying conditions, have never been picked up (Wood, 1976; Johnson and Cohen, 1984) and no further work appears to have been done following Good's (1988) analysis of structural regression in moderating teacher assessment.

Validity: the relationship between teacher assessment and the external calibrator

4.30 The way into the validity and validation argument is typically through a consideration of overlap between measures. Conventional wisdom is that there should not be too much of a relationship between what is assessed by the teacher and the examination, but not too little either; Smith put it rather elegantly:

> …the degree of overlap or correspondence is a contentious point: too little and the moderating instrument is unsuitable, too much and doubt is cast on the advisability of having both components as part of the same examination process.
>
> (Smith, 1978)

The point is indeed contentious. Usually too little is favoured at the expense of too much. The argument goes as follows: if those objectives which cannot be assessed in an examination situation are assessed at school or college then it should not necessarily follow that teacher assessment will bear any close correlation to examination performance. As an extension of this, it is sometimes argued that only those assessment tasks which are congruent with those assessed in the written examination should be moderated by the examination (see, for example, New South Wales Board of Studies, 1998, p.42). Strong though these arguments are the consequences are problematic for statistical moderation, in relation to both calibration accuracy and fairness. Of course, there is an asymmetry between teacher assessment and what is assessed in a written examination; just about anything that can be assessed in an examination can be assessed internally by the teacher but the reverse is far from true.

4.31 A common mistake when checking out teacher assessment - external calibrator correlations is to look only at within-centre correlations and not also at levels of scores. You have to look at within-centre correlations to see what ‘bouncing’ there is (that is, variations in coefficients between centres and between occasions) but you also need to look at the levels of performance in the centres since, where these differ considerably they affect overall correlations, sometimes considerably. The across-centres correlation, which is formed by correlating average scores for schools or colleges, will expose uneven performances on teacher assessment in relation to the examination, and this parallels Cronbach’s observation that a subject in which one competence or skill or set of skills was developed at the expense of the other could go unnoticed since one or more schools or colleges could on average be high on one competence and low on the other without this showing up in the correlation between scores (Cronbach, 1970).

4.32 It follows from the asymmetry between teacher assessment and the external calibrator that the validity of the assessment process rests first and foremost on what is done with the teacher assessment. For the most part teacher assessment can do what the external calibrator can do (except for the pressure of performing against the clock) but the external calibrator cannot do what teacher assessment can do in terms of the domains being assessed. Logically we would design assessment to operate in the classroom and add in examinations where they might fill a particular role, but the tendency has always been to do it the other way round, in pursuit of what is always called ‘rigorous’ assessment. Where we do develop a complementary structure that combines teacher assessment and examinations we need to do so within a structured arrangement that clearly identifies the purposes of both and the relationships between objectives or assessment criteria allocated to each.

5 Examples of experience with statistical moderation

5.1 In this section we want to highlight selected experience of the statistical moderation of teacher assessment (past and present) and take account of some reviews and initiatives. As far as possible we discuss the various experiences of statistical moderation in relation to

- the type of teacher assessment and the kind of teacher-assessed result that is adjusted
- the explanation of the logic of the adjustment process
- the instruments and statistical procedures that are used to link standards
- technical problems that have been encountered and how they have been addressed
- political, social and educational problems associated with the approach and how they have been addressed

Australia

5.2 In some respects Australia is currently the home of statistical moderation methods. The most sophisticated (and elaborate) methods are to be found there and there is the greatest range of analysis of the method available in the literature. The concern with the approach arises from the need to place students on a common scale for tertiary entrance, leading to a rank order that can be used to fill quotas, using a relatively automatic process. However, the construction of tertiary entrance scores has been problematic and controversial in Australia and the methods used vary from state to state. McCurry (1995) divides these methods into two broad groups.

- There are syllabus-based examination systems which examine a centrally determined curriculum with subject-specific examinations that have varying amounts of school-based assessment within them. The examination results are the basis for the construction of tertiary entrance scores.
- Elsewhere there are school-based assessment systems that are moderated by the use of a common test of skills or abilities that are not curriculum-specific.

In this second category the Australian Scholastic Aptitude Test (ASAT) was developed, not to generate scores of individuals, but to provide a basis for scaling groups by school and by subject in order to provide placements on the tertiary entrance scale. Criticisms of this test related to the basis for its construction (seen to be divorced from the curriculum), perceptions that it favoured some groups or subjects over others and that students did not take it seriously because the results influenced group rather than individual placement.

Queensland

5.3 Queensland is a state that falls into the second of McCurry's groups. It has 100% school-based assessment that is moderated for certification purposes by review panels that are strongly teacher-based. This is a system that originated as a result of the extraordinarily radical Radford report in 1972 that recommended the abolition of external examinations and the establishment of a school-based assessment under the management of a state-wide assessment authority; with some development and modification over the last 30 years this provision remains largely unchanged as the basis for individual certification of achievement.

5.4 The construction and delivery of tertiary entrance scores for Year 12 students aims to be conducted "without damaging the principles and values of externally-moderated school-based assessment." (Pitman, 1999). It was originally based around the use of the ASAT but this approach was modified following an extensive review in 1991 that led to the use of a curriculum-based test. This is the Queensland Core Skills test (which is an achievement test based on the 49 Common Elements in the Queensland Senior Curriculum) which form the basis of a complicated statistical apparatus to determine eligibility for university entrance. Pitman describes this as

- teachers making decisions about each student's relative achievement in each subject in each school; these are based on task-specific criteria developed in advance for each curriculum-embedded task and expressed on an interval scale (this describing rank order and spacing)
- these judgements are scaled using linear scaling on mean and standard deviation of the group results from the core skills test to establish (i) equivalence across subjects within the school and (ii) equivalence across schools
- generating for each student (i) a position of overall achievement in a scale of 25 bands and (ii) positions of achievement in fields of study each reported in 10 bands; these two pieces of information are available to the tertiary institution (both as rank orders) alongside the student's profile of results on the core skills test and any additional information outside this assessment structure.

5.5 Tertiary institutions are thus able to do a staged decision process that takes account, in order, of

- subject prerequisites and levels of achievement
- rank order indications of overall achievement
- subject rank orders of achievement
- results of the cross-curriculum test
- additional information supplied by the candidate or school

5.6 The statistical moderation element is actually quite limited and does not lead to any merging of scores other than those required to generate positions of achievement in the fields of study. It is, however, constrained within many of the same considerations already discussed in earlier sections. We have already noted general concerns about the earlier use of ASAT for this purpose, and some of the correlations between the ASAT scores and teacher assessments were modest; for example, McGaw (1977) reports a pooled within schools correlation between teacher assessments aggregated across subjects and ASAT scores of 0.51 (1970 data for 7595 students across 178 schools) and a weighted within schools correlation of 0.19. The latter reflects, of course, the need for scaling of school results to a common metric but McGaw comments

> Whether it is worthwhile to adjust between-school differences in teacher assessments to a variable which correlates only 0.51 with the assessments within schools is another question … it is probably unrealistic to expect any higher correlation than about 0.7.

He then reports later (1973) coefficients that rise to 0.64 and 0.45, following the introduction of consensus moderation methods designed to reduce between school variations in standards.

5.7 What empirical evidence do we actually have on correlations for the core skills test that is now used in place of ASAT? The Queensland Board uses two measures of teacher judgements - the 'within school measure' and 'polyscore'. Within school measures are based on judgements by teachers and result in a continuous measure which correlates at about 0.73 with the test scaling score. Polyscore is based solely on levels of achievement, but is also a continuous measure and correlates slightly higher at about 0.75 to 0.77. These two measures are based on different types of teacher judgement but are sufficiently similar to give a good idea of the relationship to the test. These correlations are similar to the correlation of the Queensland Core Skills test and the tertiary entrance score that students receive (0.73-0.75). The polyscore and the tertiary entrance scores correlate at about 0.93 (Nott, 2004).

5.8 Writing just after the introduction of the core skills test as the external calibrator Maxwell (1994) cited evidence suggesting that, although there were points of public confusion about the operation of the process of scaling for the generation of tertiary entrance scores, the process had worked well and had wide public acceptance, partly because the fundamental value of assessment that was school-based was widely approved. He claimed that the core skills test had already gained wide acceptance and was seen to be successful, not least because it was reported in its own right on each student's Senior Certificate. It had thus attained a value in its own right, rather than being an add-on that was required only for scaling purposes.

5.9 Wolf (1996) questions whether the core skills test can be a fully valid predictor in all subjects; it is clearly unlikely that it will be equally valid for all subjects but may, of course, be sufficiently valid for group equating purposes. She also questions whether the effect of scaling doesn't disadvantage high achievement by one particular class group in an otherwise average school; this is an anxiety that is not particular to Queensland. Subject performances by sub-groups within a school are likely to correlate differently with the calibrator and may be at different standards, so that what appears to be a single regression line may actually be composed of many smaller and perhaps non-parallel regressions. However, the issue is whether the Queensland test operates better than most to secure dependable tertiary entrance; Wolf suggests that a single test will always have limitations but Pitman *et al* (1999) reckon that the broad structure of the test, its high quality and the precision of its relationship with the curriculum makes it entirely suitable for this purpose and we do not know of any evidence that suggests that this is not the case. This does suggest that a curriculum-embedded test that is not subject specific may be a suitable external calibrator in a range of subjects in the UK although it would be essential to test its suitability and power in relation to each subject that is to be moderated.

5.10 The present requirements in the UK do not require statistical moderation that equates subjects although reference tests were used at one time in subject comparability (see Bardell, Forrest & Shoesmith, 1978 and the earlier comments on the use of NFER Test 100) and there is a tacit assumption in the generation of UCAS scores that subjects of equal type are of equal standing. Any concerns about schemes such as that in Queensland - that it does not treat subjects equally and that the interests of some subjects (or those pursuing them) are poorly served - would not apply if the calibrator were used solely to equate schools within a subject but not subjects within an aggregate such as a tertiary entrance score. In this situation a subject-specific test or examination probably has the considerable advantage of greater power in moderating teacher-assessed marks in that subject, although specially-designed subject tests have the disadvantage of additional costs and greater administrative demands. However, if (as with the core skills test in Queensland) the calibrator has some credibility in its own right, and its results are certificated independently, its reduced power is offset, at least to some extent.

5.11 It might be a mistake to assume that an equivalent test to be used for moderating purposes could immediately be generated on the basis of current UK experience with key and basic skills. A feature of the Queensland model is the integration of the core competencies into the curriculum design, within which the separate subjects are developed. Key skills, on the other hand, have the appearance of being superimposed on the current range of GCSE and A level provision and certainly do not command the support required if a Queensland-style moderating instrument were to be introduced, even if performances on it were to be reported in their own right.

Victoria

5.12 Like most Australian states Victoria has moved in stages from a traditional school examination structure to a position where there is a substantial contribution from teacher assessment. Unlike Queensland, however, in Victoria the balance between external examination and school-based assessment is now generally 50:50. In the course of this evolution a verifications system, introduced around 1980, that involved standardisation and expert review procedures was judged to be insufficiently reliable and increasingly costly (Brown and Ball, 1992). Whilst simple statistical moderation was seen to be cheap and objective it was thought to be weak in situations where objectives covered in the school assessment are not covered in the test and where entries are small. It was also said to be open to manipulation in schools where, in order to maximise students' positions on a tertiary entrance score, teachers concentrate on maximising written paper results, so as to drag up scores on school-based assessment.

5.13 A key issue in the debate was seen to be the need to provide continuing assessment support to teachers without requiring teachers to meet in large numbers across the state (Hill, Brown & Masters, 1993); the issue of the cost and practicality of conducting consensus activities across a large geographical area is more pressing in Australia than in the UK. Hill, Brown, Rowe & Turner (1997) describe a new Victoria procedure that used a mixture of teacher-marked Common Assessment Tasks undertaken over a period of time alongside school-based tasks assessed according to centrally-developed criteria and supported by school-based systems of support and quality assurance.

5.14 The Common Assessment Tasks thus represented a standard element that operated across all schools, but embedded within the curriculum as ‘coursework’ components, rather then being operated on an externally-determined timetable. There were three or four tasks in each area of study and not all of them were tests. Between them, they assess all of the key learning outcomes for the study, including a range of generic skills, and they are specified within the study design. Some were marked in the schools, using supplied marking schemes or criteria while the marking of others (such as more extended tasks or performances) were verified, usually through sample external marking. Hill, Brown and Masters report mean correlations amongst these Tasks within a range of subjects; these typically vary from 0.5 to 0.7 although there are some lower values. They also offer estimates of the reliabilities of school-based assessments, most of which are in the range 0.6 to 0.9.

5.15 Alongside these tasks teachers developed their own tasks upon which aspects of their assessment were based but the whole structure (unsurprisingly) has been reported as becoming increasingly prescriptive and tending to control the whole content of the curriculum. The upshot was that the large Common Assessment Tasks were recently replaced by requirements for a larger number of much smaller tasks that sample the range of required skills and knowledge rather than ensuring that everything is covered. Teachers set these tasks within the criteria provided and mark and grade them. These marks are aggregated and are subject to moderation.

5.16 As part of their earlier review of the existing provision in Victoria Hill, Brown & Masters (1993) had recommended a system of statistical monitoring to operate within the above provision that would identify where a school's assessments moved outside certain prescribed limits in relation to its examination scores; if that happened an inspection process would be initiated. They saw this as part of a process of moving away from quality control and towards quality assurance and specifically rejected the use of statistical moderation. They did, however, suggest that there should be routine monitoring of the internal consistency of teachers' use of the assessment criteria, of the reliability of the marking of both internal and external tests and of the correlations between internal and external grades. If the latter fell outside the range 0.5 to 0.75 this, too, should trigger an investigation.

5.17 Victoria also has its external written examinations to which are now added a General Achievement Test (that lasts 3 hours and is in two parts that focus on problem solving in a mathematical, scientific and technological context and on abilities in a social, artistic and cultural context) that is administered before the finalisation of school-based assessments and the examination. All of these elements are reported separately for those students who have indicated that they wish to be considered for tertiary entrance. Victoria then uses a process of statistical moderation of the teacher assessments. It is justified (on the Victoria Curriculum and Assessment Authority website) in terms of comparability and fairness when schools are using different sets of assessment activities that may be of a different standard or marked to different standards. The method used is one of linear scaling of means and standard deviations using the external examination as the calibrator. The General Achievement Test scores are also used alongside the examination scores

> … in studies where in doing so a better match with schools' assessments throughout the state is achieved. In all such cases, the examination scores will always be the major influence.

5.18 The moderation process is described in three steps

- The first step in moderating schools' assessments in each subject is to identify the moderation group. Where there are several classes in one school they are treated as a group and where a group in a school is small it may combine with another school to form one moderation group (see the note on this procedure in Section 7).

- The second step is to form an external score for each student doing the study, based on their examination scores for the study and, for a number of studies, using their General Achievement Test scores as well. These external scores are used as the common standard for all schools teaching that study.
- The third step is to use the external scores of the moderation group to adjust the school coursework scores for the group. To do this, the moderation procedure ensures that the highest moderated score is made equal to the highest external score and that the median and quartiles of the moderated scores are made equal to the median and quartiles of the external scores.

5.19 According to the website this equipercentile scaling procedure aims to make the mean (average) of the moderated scores as close as possible to the mean of the external scores. The scores for students with anomalously low external performance or students who did very poorly on the school assessment but very well on the external assessment are removed from the distributions and treated separately, thus avoiding the problems associated with flop scores, discussed elsewhere in this review. The whole process does not change the rank order of students, as determined by the teacher assessment scores.

5.20 This is clearly not a particularly transparent system and it has been observed that, given the complexity of the method, teachers are not able to predict the outcomes and therefore cannot play the system. That does not seem to represent much of a commitment to teachers' professional development in assessment although it is clear that many teachers in Victoria do participate in the moderation process and are therefore presumably quite well informed. Reports suggest that the procedures for teacher assessment and its management have reduced costs and workloads, concentrating resources in schools where moderation was needed, and is generally effective.

5.21 Victoria also uses a method of equating based on the scores of all candidates in the core subject of English, taken by everyone and used as an anchor (it appears that mathematics may also be used for this purpose, although not always). This aligning of subjects is a necessary requirement for the generation of a tertiary entrance rank and uses a method of linear scaling of means but not standard deviations. It is a process that has created considerable alarm amongst students and their parents, particularly when it has resulted in scaling down, and it has also lead to some distortion in subject choices. This is a phenomenon that was being reported in Australia in the 1970s and it is somewhat surprising that some states continue to use methods that are clearly obscure to stakeholders and that come between the performance in the classroom or examination room and the final mark or grade that is reported. The tertiary entrance rank is actually calculated by the Victoria Tertiary Admissions Centre using subject scores received from the Victorian Curriculum and Assessment Authority by adding a subject score in English, English Language, Literature or ESL, the next best three subject scores permissible and 10% of the fifth and/or sixth permissible subject score that is available[5].

5.22 Thus, once again, the operation of the certification process (analogous to GCSE or A level in the UK) does not use statistical moderation methods for teacher assessment but does use a linear scaling method to align subjects, leading to the generation of tertiary entrance scores.

[5] Described on http://www.vtac.edu.au/common/enter.html

New South Wales

5.23 In 1977 New South Wales adopted an assessment structure for Higher School Certificate that gave equal weight to moderated school-based assessments and external examinations. This evolved in 1986 into a provision where the two components were reported separately on the grounds that this enhanced the validity of the information provided. At the same time the Board of Studies issued a series of support publications that aimed to underpin and explain the whole assessment process in which there appeared to be considerable confidence. The basis for the teacher assessment was, for each school, an assessment programme for each of their courses, drawn up in conformity with Board requirements that included objectives to be assessed. Schools were expected to adopt and implement assessment policies to support the delivery of this process. A review in 1997 further supported this process and the adoption of increased levels of information and support for teachers particularly seeking to ensure that the range of objectives that was assessed in schools did not simply mimic what was assessed in the external examination.

5.24 The Board had used a linear scaling of means and standard deviations for its moderation of teacher assessments. However there was a view that this resulted, in some cases, in students who gained the highest examination mark in a group receiving a moderated internal mark that was lower; it seems that this was felt to be publicly unacceptable (and may have had an effect on the computation of the University Admissions Index - see below). One of the more unusual aspects of the New South Wales Board's procedures was the adoption of a method for dealing with situations where the relationship between teacher-assessed and examination marks is non-linear; this is usually evident in small class groups and the distributions of the two sets of scores differ markedly. MacCann (1995) developed a method for fitting a quadratic polynomial function that would show the curved relationship between these two sets of scores, as a basis for moderation. He supported its use in situations where, because of small group size, the use of equipercentile methods is unsatisfactory because of the undue influence exercised by what are called 'flop' scores (the issues relating to flop scores in equipercentile scaling have already been discussed in Section 4). For many larger groups where the distributions are similar the method reduces, in effect, to a linear scaling as before.

5.25 The method hinges on the use of a second-degree polynomial function whose use results in the retention of the rank order of scores but not the relative intervals between them (except approximately, over a very small score range). In order to make the computations possible three constraints are imposed on the curve fitting: these are that the mean of the moderated marks is set equal to the mean of the examination marks (as in linear scaling) and that the maximum and minimum values of the moderated marks are fixed. These last two constraints were imposed by the Board that required that the candidates who gained the highest and lowest examination marks should gain moderated marks that were the same as their examination marks.

5.26 McCann modified the last of these constraints in two general cases.

- Where the minimum examination mark is atypically low the candidate is removed from the analysis which is then based on the reduced group. The excluded candidate is then moderated by a linear extrapolation procedure (note that, where a candidate is excluded because of absence a non-linear interpolation is used, based on the fitted curve for all other candidates).

- It is possible that, at the foot of a mark distribution, the curve slope is near-horizontal. In this case scores are adjusted to the minimum value necessary to ensure curve-generation.

He provides procedures for dealing with these cases and a full description of the curve fitting methodology.

5.27 In a review of the assessment programme (NSW Board of Studies, 1998) the authors report that, with this moderation procedure, the mean correlation between teacher assessments and examination over 93 subjects (we assume in 1997) rose from 0.75 (unmoderated) to 0.88 (moderated). In addition to ensuring justice for the best candidates the New South Wales Board of Studies says that it takes similar care to avoid injustice for those at the bottom end: this does not just respect the teacher's ranking of his or her students, but also how he or she distributes the students on the mark scale:

> If there are two students quite close and then a large gap and then the rest of the class, say, we need to make sure the moderated assessment marks reflect this distribution (even if the exam marks don't actually turn out in a similar pattern). If you used the equipercentile method, only the teacher's ranking of students (and not the distribution) would be preserved.
>
> (Pickering, 2004)

5.28 The calculation of the University Admissions Index is conducted separately from the Board of Studies conduct of the Higher School Certificate. It employs a scaling process that seeks to overcome the perceived lack of comparability between different courses and the fact that students take different course combinations. The method uses an algorithm that estimates what an individual’s scores would have been had all students taken all subjects. This preserves the rank order of a student within the mark scale for each of the courses that he or she has completed. The detail of the process is complex (NSW Vice Chancellors Conference, 2002) and beyond a simple summary here, particularly since the need to develop a similar rank is not a priority in the UK. However, it is important to note that the process is statistically self-contained - that is, no external calibrator is used and the manipulation of the mark scales rests on a series of assumptions about their inter-relationships that it is difficult to critique without a very detailed insight into the education system in New South Wales.

Western Australia

5.29 Western Australia has an assessment regime that uses both external examination and school-based assessment, generally on a 50:50 basis. The teacher assessments are curriculum-based so that

> Schools enjoy the freedom to design their assessments to meet the needs of their students. This implies that a raw School Mark does not have the same meaning as the same mark at another school. Statistical moderation makes the marks comparable between schools.

Statistical moderation is built into the assessment system as a basis for adjusting teacher-assessed scores between schools. External examinations in the same subject (the Year 12 Tertiary Entrance Examination (TEE)) are used as the external calibrator. The moderation of the raw school marks is by linear transformation.

5.30 The moderation population is usually about 90% of the whole group and consists of students who are expected to have performed in the Tertiary Entrance Examination at about their normal level. Students who might have performed at a level significantly different from usual (defined as those whose standardised marks are substantially different in the exam from in the school-managed assessments[6]) are excluded so that the moderation parameters can be determined more accurately.

5.31 However, the whole process is clearly geared into tertiary entrance. Moderation, standardisation and scaling are three separate processes applied to students' scores. Following the moderation process the TEE scores and the moderated TA scores are standardised to put them on the same linear scale and distribution and to make scores in different subjects comparable to allow for aggregation. Students are then ranked to produce the Tertiary Entrance Rank.

South Australia

5.32 South Australia has a 100% teacher-assessment regime for about half its subject awards accompanied by support mechanisms for teachers and their schools or colleges and a moderation system that is not statistical. The other subjects are assessed on the basis of 50% public examination and 50% moderated school-assessed coursework stipulated by the South Australian Board of Secondary School Studies, with an examination component as the moderating instrument for the coursework.

5.33 The method of moderation prior to certification is by linear scaling: the mean and standard deviation of each class's school assessment marks are adjusted to bring them into line with the distribution of the examination marks. Teacher mark rank order is preserved but there are some differences between rankings on the examination and rankings on the moderated teacher assessments which are reported separately. There appears to be some systematic attempt to ensure that the scaling process is not unduly affected by untypical scores that result in outlying scatterplot points; in a private communication (Benger, 2004) says that

> .. the LOWESS procedure that we use in the moderation process does its best to retain the relativity of the positions in the ranking of the school assessment marks as well as the ranking itself.

The LOWESS procedure referred to is an acronym for Locally Weighted Scatterplot Smoothing. It is a weighted least squares method for producing a smooth set of values from, for example, a scatterplot with a 'noisy' relationship between the two variables. It involves creating a linear fit between specified limits, but there is an option of fitting to a quadratic line. Descriptions can be found on several websites including http://www.itl.nist.gov/div898/software/dataplot/refman1/ch3/lowess_s.pdf and http://www.math.yorku.ca/SCS/sssg/lowess.html (the former provides examples of its use). It will be in the empirical choice of parameters that this procedure can be operated so as to reduce the effects of scores that are thought to be anomalous without violating the teachers' rank orders.

[6] The critical value is calculated afresh each year. In 2002 it was a difference of about 16 on score scales of 100.

5.34 To scale between subjects for the generation of a tertiary entrance rank, the equal achievement principle is used, as in New South Wales, so that marks are adjusted on the assumption of equal ability between subject groups.

Commenting on statistical moderation in Australia

5.35 The ongoing debate about tertiary entrance in Australia has resulted in some volume of literature on the use of external calibrators and the operation of scaling and equating procedures. Much of this related to the use of the Australian Scholastic Aptitude Test that was seen to have some defects; some that relate to its relevance to the curriculum have already been mentioned and others are concerned with matters such as gender bias. Daley (1985) showed that gender bias was likely (to one extent or another) in the use of ASAT in several states and commented that

> …the validity of producing comparisons of achievements between students at different colleges solely by means of moderating internal achievement measures by an external multiple-choice aptitude or ability test must remain questionable.

5.36 Similarly, as Australia has developed as a multi-cultural society there are increasing concerns about the ethnic and cultural biases that might arise in the use of external calibrators. This does appear to be part of a wider debate about the influences of standardised testing on education, opening something of a divide between the teaching profession and psychometric professionals that includes anxieties about the ways in which students manipulate the system in order to ensure the right tertiary entrance score and thus distort learning (Doecke *et al*, 2000). This concern is not new and was one aspect of the review of the use of ASAT in Queensland (Viviani, 1990).

5.37 It is not obvious which of the Australian experiences offers the most appropriate evidence for the UK situation. In the context of the present discussion it is easy to forget that, below the senior secondary level and apart from the process of deriving tertiary entrance scores, Australian states operate school assessments that incorporate large amounts of teacher assessment that are either un-moderated or moderated by an inspection or consensus, rather than a statistical, process. At the more senior level, however, statistical moderation methods are widespread but show considerable diversity in their approaches to moderation and in the generation of tertiary entrance scores. It is not clear whether the most sophisticated statistical adjustments used for certification (which may be beyond the understanding of lay people) make an appreciable difference over more rudimentary methods, and there is some evidence that the derivation of tertiary entrance scores continues to cause anxiety and controversy. An Australian statistician, writing about the Australian tertiary education scene twenty years ago, observed that

> … there has been no demonstration to my knowledge that under Australian conditions our tertiary institutions, representing a major consumer of the students emerging from the end of the secondary school years, have been better served by the sometimes radical revisions in methods of constructing aggregate scores for tertiary admission procedures.
>
> (Daley, 1985)

and comments about practices in Victoria suggest that it continues to be very difficult to convince stakeholders that methods that involve scaling are fair.

United States

5.38 There is no experience known to us that pushes the USA forward as a comparator. An enquiry through Educational Testing Service in Europe (Utrecht) elicited a blank, likewise an enquiry to ETS in Princeton. A paper reviewing the UK scene 10 years ago written by two senior American educational measurement figures noted the "disenchantment" with the "shortcomings" of statistical moderation procedures and argued that reliance on an external examination to make adjustments in the results of internal assessments is likely to undermine the goals of the internal assessment and distort instruction. Furthermore, they say, relying solely on moderation by inspection would probably have credibility problems in a country like the USA where technological and statistical solutions are more the norm. Strict statistical moderation, on the other hand, would undermine the goals of the assessment systems that are currently under consideration. Burton and Linn concluded that

> …it seems more likely that some sort of hybrid system will be required that relies on a combination of an external assessment and statistical comparisons to identify places where more detailed information, the use of audits, or the use of moderation by inspection is needed.
>
> (Burton and Linn, 1994)

5.39 Earlier we reported the general findings from the EPPI review of the use of teacher assessment for summative purposes. Some of the evidence for that review was drawn from work in the USA on large-scale portfolio assessment by teachers for summative purposes, and reviewed by Koretz (1998). This does not deal with statistical moderation issues - indeed, it doesn't mention moderation at all - but the interest is in the extent to which the instigators of those programmes (in Vermont, Pittsburgh and Kentucky) saw them as stand-alone provisions that would be capable of delivering valid and reliable assessment unsupported by external controls. In fact, Koretz illustrates the difficulties that the schemes have encountered, not only in terms of limited validity and poor marking reliability but also in relation to teacher workloads and uncertainties about the relationship with the rest of the curriculum.

5.40 It is possible that the solution would be some form of moderation, and Koretz does suggest that there is a need for more support and information for teachers, some of which could come from networking and forms of agreement trialling. He also highlights the difficulties in using a provision such as portfolio assessment to serve both learning and summative assessment purposes, and this is perhaps where some of the difficulties with teacher assessment really lie.

Sweden

5.41 Noah & Eckstein (1989) reported that

> In the mid-1970s, Sweden discarded a limited but usable final secondary school examination system in order to reduce the strain on pupils, produce more valid and reliable predictors of university success, and (it was hoped) correct socio-educational inequities in assessment. In place of the final examination, the Swedes installed a combination of marks gained during regular classroom and homework and in nationally set tests administered at intervals during the school career.

5.42 Abandonment of final examinations was also motivated by the desire to improve the diagnostic and predictive value of tests of individual student achievement and to give teachers national benchmarks against which to set their own pedagogical efforts. The provision was heavily norm-referenced and the nationally set tests provided the basis for standardisation of marks across schools (but did not contribute to the certification of the individual student). Test results were collected from a nationally representative sample and cut scores were decided according to a pre-determined distribution. These cut scores were distributed back to each teacher who then knew how to mark their own pupils on the tests in line with national norms.

5.43 Teachers were required to maintain the mean of the school assessment marks that they awarded within 0.2 of the mean on the national test (that had a distribution with a mean of 3 and a standard deviation of 1). Elley and Livingstone (1972) discuss this reference test approach in some detail (pp80-87) and make some specific recommendations concerning the conduct of the process and the scales that should be used. It does seem that asking teachers to keep within tight limits like this could result in a manipulation of marks even where student performances suggested achievements that were not covered by the test and individuals with unusually high performances may suffer from the ‘company you keep’ factor that we discuss elsewhere. It is also unclear how teachers coped with unexpectedly good or poor performances of individual students or what the effects were on assessments done in small class groups.

5.44 Noah & Eckstein say that the Swedes were willing to incur rather heavy costs to achieve their goals. The system required time-consuming collaboration among teachers in a given school, and across schools in a region and exceptionally detailed record-keeping is required. Although Sweden may have abandoned its final secondary school examinations, there had been no abandonment of tests and examinations in general. However, in 1994, after several years of debate, Sweden abandoned the standardisation of assessment via the imposition of external norms in favour of a criterion referenced (or rather a standard referenced) marking system. This represented a developmental path that reflected dissatisfaction with norm-referencing and the use of tests for establishing standards rather than a dissatisfaction with statistical moderation as such, although some of the more problematic constraints of the standardisation process will undoubtedly have contributed to the pressure to move towards a standard referenced system in which all students are equally referenced to an external set of criteria.

5.45 However teachers were ‘recommended’ to use tests devised by the national agency in the interests of consistency and in 2000/2001 the use of the tests in Swedish, English and maths was made compulsory. In other subjects teacher assessment is still supported by the availability of test materials from an item bank but there does not appear to be any system of moderation that directly uses data from these tests.

South Africa

Independent Examinations Board

5.46 For some years this Johannesburg-based Board has provided examinations for a range of mainly non-racial private schools in South Africa and some adjoining countries; see Lubisi & Murphy (2002) for information about the development of the Senior Certificate (Matric) Examination in South Africa. It uses a straightforward system of linear scaling to marks on school-based assessment (SBA), but only if these are outside certain limits (Long, 2004). The possible adjustments for any candidate group are exemplified as follows (the limits are clearly arbitrary and empirical and the marks on the SBA and examination are on the same scale):

- If their SBA mean is more than 15% above their examination mean: the SBA mean is adjusted to be 5% more than their examination mean
- If their SBA mean is from 10-15% above their examination mean: the SBA mean is adjusted as follows:
  - for 14% above adjust SBA mean to exam mean + 6%
  - for 13% above adjust SBA mean to exam mean + 7%
  - for 12% above adjust SBA mean to exam mean + 8%
  - for 11% above adjust SBA mean to exam mean + 9%
- If their SBA mean is from 5-10% above their examination mean: no adjustment is applied
- If their SBA mean is less than 5% above their examination mean: the SBA mean is adjusted to be 5% more than their examination mean

5.47 This appears to be a system that progressively penalises high performances on the school-based assessments and forces the means on school-based assessments to be somewhere in the region of 5-10% higher than examination means. The effect on teacher assessment is not known. The process is, in fact, in 2 stages: this adjustment of the mean, where this is needed, followed by the scaling of teacher assessed scores (seen now to be on a common scale) to the same standard deviation as the examination marks as a whole. There is no evidence that teachers participate in this process or that there is any attempt to discover reasons for larger adjustments. Given the dispersion of IEB schools that may not be practicable.

The Gauteng Senior Certificate

5.48 In contrast to IEB the Gauteng Senior Certificate - which is a national qualification operated provincially, and which has been of pivotal importance in South Africa for many years (Talbot, 1995) - is taken by well over 100 000 candidates in over 700 centres in a small densely populated region of South Africa, and is evolving rapidly in the post-apartheid era of educational reform (Govender, 2002). This evolution has included the incorporation of teacher assessed components in what had previously been an entirely external examination that now was seen to be inappropriate; this was as a result of a national policy decision that was rushed through with inadequate preparation and that in some provinces saw continuous assessment simply as the incorporation of tests in classroom activity (Lubisi & Murphy, 2002).

5.49 In Gauteng this development was accompanied by the design of a moderation system that sought to align assessments at the school, district and provincial levels largely through consensus (cluster) processes supported by moderators. This process has suffered difficulties because of a lack of expert support and difficulties in supporting the work of teachers in the clusters. Most significantly for this review Govender points to two particular problems:

- teachers lack that expertise and information to relate their assessments to national norms, and the cluster system is not yet enabling this

- statistical adjustments to provincial results are made at the national level in order to achieve desired distribution patterns; Govender does not specify the nature of these adjustments (but they are almost certainly through linear scaling, probably only of means) but does observe that the process is not understood by teachers and has

> … accustomed Senior Certificate teachers to having their assessment decisions overruled by invisible statisticians and this has thereby tended to undermine their confidence in their own assessment competence

New Zealand

5.50 Elley and Livingstone's work, already extensively cited, was part of a process of reform to New Zealand's School Certificate, and focused heavily on the introduction of some elements of teacher assessment and the use of specified learning outcomes as a basis for assessment. Their extensive discussion of statistical methods of moderation reflected past practices where the examinations were normative in nature and where inter-subject statistical scaling was used to ensure fair treatment of candidates taking different subjects or subject combinations (Crooks, 2002).

5.51 During the 1980s New Zealand incorporated internally assessed components in the School and Higher Certificates and later made more extensive use of descriptions of learning outcomes, with all secondary qualifications moving under the control of a single authority. Although some elements of the earlier dual system of qualifications remain one of the tasks of the New Zealand Qualifications Authority has been to define the basis for moderation of assessments in the qualifications under its control, and particularly to emphasise the role of moderation as an activity that helps to ensure the consistent interpretation and application of standards rather than one that ensures that results on any particular occasion are consistent and of the correct standard. Thus Strachan (1995) emphasises the noticeable shift away from after-the-event statistical adjustments that characterised past moderation systems. Once again, it is not that statistical moderation itself has failed but that it is seen to be incompatible with new requirements for a qualifications system.

5.52 In this respect New Zealand practices reflect a shift that has taken place in many countries as the boundaries between assessment practices in school-based, vocational and occupational qualifications have become blurred and as assessment paradigms associated with authentic assessment and paying more attention to formative processes and the professional development of teachers in assessment methods have played a more prominent role. Whilst there remains a wish to ensure that quality control processes ensure the greatest dependability for individual certification there is a greater interest in the processes of internal assessment that lead up to the teacher assessed mark or grade and whether it is possible to develop structures, regulations and systems that will ensure the quality of the assessments, rather than having to rely on repeated *post hoc* adjustment procedures. Within this spectrum of possibilities, statistical moderation has sometimes been viewed as the least appropriate solution since it offers little or no constructive interaction with the assessment process.

5.53 In this context, quality assurance systems based on the provision of clear methods and criteria for assessment, the operation and verification of suitable internal systems for the conduct of assessment, agreement trialling and consensus moderation, a clearer role for awarding bodies in the provision of support and feedback, and mechanisms that will identify and remedy situations where the assessment process is failing all appear more progressive and more constructive.

West African Examinations Council

5.54 For his PhD Amedahe F K (1998) conducted some experimental work on statistical moderation done by the West African Examinations Council, chosen because of the heavy financial requirements and the need for high levels of subject matter and educational measurement expertise to implement the either inspection or consensus moderation procedures. The assessment by the teachers is continuous and curriculum-based and involves frequent and systematic criterion-referenced assessment over the period of the course, averaged to give a single score. This is regarded as a very valid measure (but not particularly reliable across schools) while the external examination is seen as less valid in relation to the whole curriculum, but reliable across schools. In the statistical moderation, the aggregated continuous assessment scores are made comparable across schools through a linear scaling process that is referred to as standardisation. It appears that WAEC also uses a mapping procedure when the number of candidates in a school is small although Amedahe does not give details of this.

5.55 The process of linear scaling adjusts the teacher assessment scores in each school to the mean and standard deviation of the school's external examination scores and the moderated and examination scores are then combined as a simple weighted average to give a final score in a subject. In his research Amehade tried out a different statistical moderation procedure (with mathematics data) scaling each school's teacher assessed scores to the mean and standard deviation of the teacher assessment in a reference school. He also tried out different approaches to combining the teacher-assessed and examination scores (including the WAEC procedure, a square-mean-root sum with nominal weights and a square-mean-root sum with achieved weights). Unfortunately, there appear to be errors in his sums and we must therefore doubt his conclusions that point towards a slight advantage in the use of a reference school for the standardisation (scaling) exercise but that, otherwise, there were minimal differences in the outcomes of the methods in the grades assigned to students.

5.56 However, he also provides some general conclusions that do not depend on the specifics of his calculations; these are that

- guidance should be given to schools about the number of tasks that are to be given to students as part of their continuous assessments
- teachers need advice on how to compute the continuous assessment scores
- item banks should be developed from which teachers can draw materials for assessing their students and to serve as benchmarks for standards in schools and as a guide to teachers in developing their items
- constant in-service training should be given to teachers on assessing their students so that in the long run there can be a move towards group moderation and monitoring if this combination of continuous assessment and external examination is to be continued.

This last conclusion suggests to us that statistical moderation may be seen as an interim process that may be replaced as expertise and resources develop. It may be the case that there is a progression in many countries from examinations with no internal assessment, through some degree of internal assessment, but unmoderated, to moderation by some simple method of scaling, eventually leading to inspection or consensus forms of moderation that may be able to support high levels of internal assessment and which are not dependent on the existence of an external examination. The constraints are resources and expertise rather than the technical qualities of statistical moderation as a process in itself.

5.57 On the specific issue of correlations Amedahe also cites a study by Adeyegbe (1993) of the relationship between continuous assessment and external examination scores in Nigeria (one country where the West African Examinations Council operates) which yielded correlation coefficients ranging from 0.24 to 0.86 for different schools and subjects. In English language, for example, the correlations ranged between 0.35 and 0.77 with 60% of schools sampled having correlations below 0.5. In mathematics, the correlation coefficients ranged between 0.44 and 0.86. Another study by the Council found that in all sampled subjects the majority of schools demonstrated positive and significant correlations, although these varied widely from school to school and from subject to subject. Some schools with high correlations exhibited differences in standards of performance and therefore a disparity between the two modes of assessment. Teachers tended to award higher scores in their school-based assessments than their students were able to achieve in the external examination. The assessment instruments were found to be unreliable and lacked content validity in all schools. Moreover, it was suggested that combining the two sets of scores was made more problematic because continuous assessment is carried out by teachers in a criterion-referenced framework while the external examinations are basically norm-referenced. As a result the procedures for moderating the teacher assessed scores and combining them with the external examination scores were reviewed and (perhaps as a concession to caution) the weighting ratio was changed from 40 : 60 (teacher assessment: examination) to 30 : 70. The three approaches to moderation were considered – moderation by inspection, moderation by consortium or consensus, and statistical moderation (Ademola, 1992) and, as noted above, statistical moderation was chosen.

Hong Kong

5.58 Hong Kong has an examination system that traditionally operates in an extremely high stakes environment and that is subject to increasing pressures following the incorporation of the territory into China. School examinations operate on a model that strongly reflects the British influence on Hong Kong but there have been many concerns that the examination system does not serve the whole school population well, that achievement standards are too low, and that the curriculum is not appropriate for the 21st century.

5.59 The Hong Kong Examinations Authority commissioned a study in 1998 to examine a range of issues in connection with schools examinations at the school leaving and advanced levels; one of the principal parts of the remit was to make recommendations concerning the development and conduct of school-based assessment which, at that time, was a very small part of the public examinations system. Much of the report deals with the curricular and other implications of adopting more school-based assessment and its impact on wider issues of certification, public confidence, access to tertiary education and the like. However, the review group did support the expansion of school-based assessment (though rather conservatively, given the pressures in Hong Kong) and statistical moderation to support it (HKEA, 1998).

5.60 The recommendation to adopt statistical moderation is, in some ways, quite surprising since Hong King has a well-developed education system with teachers who are generally well qualified (though more used to traditional rote-learning testing methods than to the management of school-based assessment) and the territory is not large. The reasons for using statistical methods cited by the West African Examinations Council do not apply in Hong Kong, and the enthusiasm for statistical methods reflected to some extent the Australian influence on the review group. The model proposed was by linear scaling of means and standard deviations with special provisions for small class groups, situations where the teacher assessments have almost no dispersion, occasions when the distributions of teacher-assessed marks and examination marks are oppositely skewed, where there is a need to take special account of students at the extremes of the score scales and where there is missing data.

5.61 In relation to all of these technical issues the review group recommended following the practices in New South Wales (discussed above) and also offered the possibility of using the method of quadratic polynomial transformation developed by MacCann (1995) for the New South Wales Board of Studies. This is designed to fit a curved line of adjustment when equipercentile methods are not appropriate (as when a class group is small) and score distributions on teacher assessment and examination are markedly different and was devised to address some of the special issues mentioned in the previous paragraph. It is also described in the earlier discussion of statistical moderation in New South Wales. As in almost all situations where statistical moderation is used some method of manual intervention was seen to be necessary when class groups fall below about 20 students (see the discussion in Section 7 below).

5.62 The Hong Kong review report also provided a little evidence in support of its claim that school-based assessment would work in Hong Kong though the authors did recommend limited weightings for this component, at least initially. The evidence was in the form of a set of correlations between tasks in a few subjects assessed independently by teachers in a small selection of schools and results on the wholly externally assessed school certificate examination. The tasks were of various types but were mainly teacher-set tests and examinations (perhaps a pointer to the types of internal assessment that might take place if school-based assessment were introduced!). The correlations were respectable: all above 0.86 in mathematics, between 0.35 and 0.75 in Chinese language and between 0.74 and 0.95 in English language.

6 Current UK practice in relation to statistical moderation

6.1 We approached awarding bodies in England, Wales, Scotland and Northern Ireland to ask about their current moderation practices and, in particular, any use of statistical methods. None of them currently make use of an external component for moderating internal marks but some interesting practice did emerge and the following gives examples of where awarding body thinking may be headed. Note that this summary is not intended to be a comprehensive account of moderating practice but simply a selection of examples to illustrate a particular class of approaches.

AQA

6.2 Moderation of teacher assessment is by inspection with a statistical layer on top. A regression equation is derived for a sample of centre and moderator marks for a centre and adjustments made on the basis of the regressed marks. The external component of the examination is not used.

6.3 The old JMB system described by Smith (1978) was replaced by moderation by inspection with subject officers deciding on adjustments and then this was replaced by the current system using regression. If the regression equation is seen to have an adverse effect on candidates then it is not used and all of the centre's work is assessed by the moderator and the moderator's mark is used. The procedure for deciding the goodness of fit of the regression equation is based on finding the centre marks to be 'demonstrably inconsistent'.

6.4 A new system allowing moderators to use a website program to key in the two sets of marks and get an immediate response about whether or not the regression is acceptable is being introduced in summer 2004 to help speed up the process. Moderators can then get back to centres for the remaining coursework in reasonable time. This process is being evaluated.

OCR

6.5 A distinction is made between internal assessment as the 'standard' assessment carried out by a teacher of candidates' work using awarding body prescribed mark schemes (and, presumably, awarding body set or suggested tasks) and teacher assessment that allows teachers to do what they want rather than what the awarding body wants. In teacher assessment there may be no awarding body prescribed mark schemes; it depends on the purpose of the teacher assessment.

6.6 Moderation of internal assessment is by inspection and then professional judgement about adjustments; alternatively, centres are asked to re-mark all their work if the marks are outside the tolerance zone.

6.7 OCR currently has a pilot project in Geography GCSE which includes both internal assessment and teacher assessment; there is a short course with assessment at the end of year 10 (50%) and a full course of seven optional units all of which are internally assessed and one or two of which may be teacher assessed. OCR, together with QCA, is considering two possible approaches to moderation:

- statistical moderation using other parts of the examination or something else such as Key Stage 3 data

- an accreditation scheme for teachers.

It has not yet been decided which of these will be used although the second option seems to be an approach that is becoming increasingly favoured.

Edexcel

6.8 A new system of moderating GNVQ and VCE portfolios was piloted last year and is now being implemented. This is similar to SQA's proposed approach that is based on helping teachers to mark to National Standards during the course rather than check and possibly adjust their marks at the end.

> *"Research has shown that assessment decisions made during a course are indicative of summative decisions. ... between 80% and 85% of GNVQ programmes are assessed to National Standards."*
> (Description of new GNVQ/VCE moderation procedures, Edexcel, November 2003)

6.9 Moderators inspect and give feedback on a centre's marking twice during the course, and decide if further moderation is needed at the end of the course. For the first exercise staff in centres assess exemplar work provided by Edexcel and in the second exercise they assess a sample of their own live candidate work taken from across the range of marks. Note that the exemplar marking is used in Edexcel's GNVQ and VCE portfolios moderation; it may not be more widespread in Edexcel.

6.10 For the BTEC Nationals, the approach is similar to verification of vocational qualifications. An external verifier checks that a centre is operating to National Standards by inspecting the centre's assessment briefs and decisions (decision can be pass, merit, distinction, fail). If the course is a two-year course (e.g. the 18 Unit Diploma) the external verifier samples work in each of the two years. If the verifier finds that the centre is not assessing to National Standards (applying the grading criteria appropriately to candidates' work) then the centre is given feedback and the verifier's report is sent to a quality development assessment manager for checking. The centre is asked to produce further samples properly assessed. Note that the external verifier does not change the centre's assessment decisions and the centre knows that it must produce the correct decisions before the candidates can be awarded a qualification.

SQA

6.11 SQA moderation of teacher assessment is currently by inspection and no statistical manipulation is involved. A recent evaluation of moderation by SQA has found that its current system of external moderation has failed to provide appropriate levels of quality assurance for the full range of SQA provision. There were problems because the process is end-loaded and nearly always resulted in double marking. Hence it was difficult to get results out in time and it was heavy on resources. Specifically, the evaluation notes:

> The underlying design flaw of the current system was the failure to recognise that there are intrinsic differences between qualification blocks (National Qualifications, Higher Nationals and Vocational Qualifications) and the type of centres which offer them. The varying experience and expertise of those centres was also not recognised.
>
> The system works with difficulty for NQ. Moderation of NQ provision is end-loaded and takes place within a very small time-frame before the external assessments thereby placing heavy pressures on centre staff, SQA staff, moderators and most importantly leaving very little time to address any quality assurance problems without jeopardising certification.

> The current system cannot support the year long activity in HN and SVQ and large areas of the SQA catalogue are untouched by moderation

6.12 As part of this review, SQA is intending to harmonise the moderation process across its various awards (National Qualifications, Higher Nationals and Vocational Qualifications) and is currently piloting a new system focused on school provision. The new system is based on training for prevention rather than cure – if teachers share a common understanding of national standards then they'll get the assessment right in the first place. This approach is very heavy on school resources and SQA is hoping to tie it in with LEA and school commitments to teachers' professional development. It also involves a formal risk analysis of each centre so that moderation activity can be targeted where it is most needed.

Specifically, the proposed new model is described as follows.

> This new model for SQA external moderation has evolved as a result of the evaluation of the current policies and procedures for moderation; the costing of the various modes of moderation; the outcomes of the centre questionnaire; the analysis of moderator reports; scrutiny of other systems of moderation and the views expressed by colleagues on the Project and Steering Groups and appointees.
>
> The new model is intended to meet the requirements of the quality assurance principles and elements and to provide a flexible, practical and cost-effective system for the wide range of SQA provision and the different types of centre. It is intended to deliver robust and rigorous moderation.
>
> The new model has been built on the following principles:
>
> - the use of interim as well as complete evidence
> - a rolling programme of moderation which is no longer solely dependent on entry information
> - targeting activity and resources on the basis of need and risk
> - supporting centres in maintaining national standards by the wider involvement of their staff
> - openness of process
> - effective management information supported fully by appropriate IT
>
> In order to target activity on the basis of risk and need moderation will be planned on the basis of a number of factors:
>
> - Centre type
> - Centre history/track record
> - Subject/award type
> - Subject/award history e.g. a new subject/award or one in which problems have been recognised in the understanding/maintenance of national standards
>
> The application of these factors will recognise the intrinsic differences in provision and centre type but will allow moderation to be planned over all blocks.
>
> Moderation will no longer be dependent on entry information or completion dates.

This proposed model is currently being piloted with 169 schools. The teacher assessment involved is that for the Units used in Highers and Intermediate Levels 1 and 2. Teachers can devise their own assessment instruments but these need to be 'pre-moderated' in advance of their use and most teachers prefer to use the National Assessment banks that are available. Teachers mark the tests, apply the pass/fail cut-off point supplied by SQA and submit the result pass or fail to the SQA (the mark is not submitted).

6.13 The SQA review included evaluation of moderation systems elsewhere and within the UK. It found that

> specific information on moderation models adopted by awarding bodies was not readily available.

The review did, however, make extensive use of Edexcel's quality assurance system for vocational qualifications, and SSABSA's quality assurance strategies.

WJEC

6.14 No statistical moderation is undertaken. Coursework assessment is moderated by inspection – a comparison of the centre and moderator's marks for a sample can lead to a centre's marks being accepted or adjusted. The adjustment is made by 'eye' rather than by statistical formula. The only statistical procedure in use is in the identification of 'rogue' centres, post-hoc, in assessment of spoken English. The information is used for the next series of assessment, rather than the current one.

6.15 There is an early warning system for identification of centres at risk, similar to that in Edexcel. A video of five candidates talking is sent to centres during the course for assessment by teachers. The centre's assessments are then checked by moderators and if necessary, the centre will be visited.

CCEA

6.16 None of the examinations for which the CCEA is responsible makes use of statistical moderation. The process in CCEA is that all centres are sampled and moderated by inspection; the sample size is determined in relation to the number of candidates in the centre. An initial sub-sample is inspected and if the moderator is in agreement with all of the centre's marks, within a tolerance limit of +/- 6% of the available marks, the inspection process stops there. If not the whole sample is inspected. The initial moderator then makes a recommendation which is adjudicated upon by the team of senior moderators. A set of rules governs this process, relating to the maintenance of the centre rank order and the proportion of the centre's sample that must be outside tolerance before there is any adjustment. There are also rules about the tapering of adjustments and how that process should relate to the tolerance limit.

International Baccalaureate

6.17 Moderation of teacher assessment follows a similar procedure to that for aligning markers of external components but the criteria for acceptance are less stringent. Moderators are asked to judge whether or not a centre's marking seems to be appropriate rather than re-mark the sample of work; centres' marks are altered only if the moderator decides they are inappropriate and then regression is used, based on the moderator's marks for a sample. The criteria for passing through moderation are based on the correlation coefficient and regression equation. The correlation between the teacher's marks and the moderator's marks needs to be greater than 0.85 and the slope of the regression line between 0.5 and 1.5. There is no limit on the difference between the sample means. Adjustments are made using the regression equation with 'tailing' at the lower end so that work which deserves zero or very few marks does not get awarded more marks. 'Tailing' is applied to the lower 20% of the marks with the 'tail' linking the regression line to the coordinates of the minimum available marks (0, 0).

7 Some underpinning issues in the conduct of statistical moderation

7.1 We are using this section to review a number of key issues that have emerged from the discussion, in advance of considering what models might be appropriate for future use.

The effect of what the teacher assesses

7.2 There is a wide range of approaches to what we have called ‘teacher assessment’; for example, it could be based on

- a small number of large, curriculum based tasks
- a large number of small, curriculum based tasks
- some form of overarching judgement of the student's achievement on the course.

In each case the tasks may be externally specified or curriculum-derived by the teacher. Whilst it is not part of this review to discuss the general conduct of internal assessment it is appropriate to ask whether we may have different expectations of statistical moderation based on each of these approaches.

7.3 If we take the simple view that teacher assessment can be treated in the same way as a test then an application of the original Spearman-Brown formula (which says that the greater the number of items in a test the higher the reliability will be, other things being equal) would suggest that the highest reliability would be when there is a large number of small, curriculum based tasks and lowest when an over-arching judgement is made. However, we need to be careful about this. The test analogy assumes that the items are all drawn from the same domain, so that test lengthening is a matter of adding more items that cover the same curriculum as those that are already there. Teacher assessment that consists of a lot of curriculum tasks may not be like that; Newton (2003) argues (in a discussion of whether a continuous assessment model of teacher assessment in the national curriculum could be used for high stakes purposes) that

> ..continuous teacher assessment is not an exercise in replication. Psychometrically speaking, replication requires that both the assessed construct, and the state of the student in respect of the construct, be identical across each testing occasion. Neither of these are true in relation to the proposed continuous assessment model: the evidence-base is likely to be multi-factorial, describing micro-developments within different sub-domains from one assessment occasion to the next; and students would also be expected to develop from one occasion to the next.

However, in comparing continuous teacher assessment with tests Wiliam argues that

> …if we adopt the conceptual framework provided by generalizability theory … then it is a logical necessity that the degree of unreliability contributed by student-task interactions will be lower than for a traditional test because there are more tasks, provided, of course, teachers can apply the correct standards, and base their levels on the latest and best evidence.

Wiliam (2003: 133)

7.4 It seems likely that we cannot be confident that the use of many small tasks is the best option from the point of view of the validity and reliability of the assessment unless we can be sure of

- the sampling of the domains - that this is done appropriately
- the quality and standard of the assessment decisions made by teachers
- the effects of later assessments compared with early assessments.

It is possible that a single judgment could be arrived at, not by plucking a figure out of the air, but by consideration of a run of marks obtained on assignments. In such a case, and where there may be a process of training and support for teacher, there may be little or no difference between the three approaches in 7.2.

7.5 There are some issues other than reliability to consider in comparing these approaches.

- The use of a large number of small tasks may place additional workloads on teachers though these may fit rather better with curriculum delivery. A casual conflation of formative and summative purposes may, however, do more harm than good in relation to learning quality.
- Any systematic bias that a teacher operates will affect the assessment outcome but it is not clear whether we will be better or worse off with a lot of tasks or one task. It may depend on the quality of the assessment criteria and whether they are well matched either to each task that is being assessed or to a group of tasks considered as a whole. Under circumstances where the criteria for a single assessment embody or lead to a bias, multiple assessments that each use their own criteria might mitigate that bias; but we cannot be sure that this will happen. It might be difficult to match rather general assessment criteria to large tasks making it possible that a teacher could apply consistent misjudgements across a whole range of tasks and thus reinforce and amplify bias.
- It is possible that a single judgement could be more valid than multiple judgements if it required the expert and balanced application of several criteria together rather than separately. This may boil down to the difference between an *actuarial* judgement, based on an aggregation of marks that may conceal bias, and a *clinical* judgement based on an expert weighing up all the assessable evidence, with an awareness of bias. However, the implications for the amount of professional development and training required could be alarming.
- Finally, there is the issue of what teachers are actually assessing. They might substitute estimates of ability or IQ for achievement and this could be entirely concealed in an apparently acceptable rank ordering (Murphy, 1974). Distortions of this kind might be more easily concealed in a single over-arching judgement.

The transparency of statistical moderation and issues of accountability

7.6 There is justifiable concern about ‘black box’ systems where the students, parents, teachers and many of the players in the assessment process cannot understand what is being done to raw teacher assessment. We should not lump together all statistical methods however. It is not difficult to explain a scaling that is designed to compensate for the leniency or severity of an individual teacher, although there will be more anxiety about scaling down than about scaling up. Manipulating mark distributions more extensively will be harder since there will apparently be some in a class who gain and some who lose from these changes. The use of more complex data transformations, state-of-the-art scaling procedures, multiple regression and other methods could only be justified if they produced demonstrably fairer outcomes for students than what went before. The difficulty will be in deciding what criteria we will use for establishing fairness and who will decide whether these criteria have been met in any particular circumstance.

7.7 A specific concern about transparency is the effect of statistical moderation on teachers and students. Black box systems are apt to induce passivity and fatalism in those who are in the loop. This can translate into a feeling of it doesn't matter what I do (that is, what marks I submit) because whatever goes on in the black box will produce something I won't necessarily recognise. That being so there is little for teachers to learn other than methods of 'playing the system'. Murphy identified a long-held belief that statistical moderation techniques

> …add to the 'mystique' that already is associated with the marking and grading of public examinations. These techniques do not do anything to develop the professional status of teachers or their role in the assessment of their own pupils. On the contrary they can greatly exaggerate the already existing 'them and us' divide between teachers and examining boards.
>
> (Murphy, 1981)

7.8 Evidently centres have a responsibility to assess candidates' work properly whatever the assessment environment but we must assume that they will be most likely to do so if they have been given a clear structure of assessment objectives or criteria to work with, clear relationships between the tasks set and the assessment criteria used, an understanding of the expectations and standards required and experience in conducting assessment of this kind. For teachers the difficulty with statistical moderation is that, however diligent and experienced they may be, they are not ultimately responsible for the moderated teacher assessment mark, nor do they have complete information that would enable them to defend it. Other methods of moderation may provide the feedback that allows this (though this is not inevitably the case) and a teacher's position is strengthened if there is a strong system of internal moderation before marks are submitted to an awarding body and strengthened further if explicit approval for his or her judgements has been given by an awarding body representative (as is the case in a verification procedure). Statistical moderation effectively puts control outside the centre and creates an ambiguous accountability. It will fall to the awarding body to provide a public justification for the moderated teacher assessed marks as a contribution to the final grades.

7.9 It is often suggested that better consumer education on the basics of examinations is needed and there are many examples around the world of awarding body publications and publicity materials that attempt to explain the processes that they use. Whether these have a significant impact on making systems more transparent is not clear. It is also not clear whether, even if methods of statistical moderation were understood, they would be seen to be acceptable.

7.10 Awarding bodies may also feel that it is necessary to introduce a process in which stakeholders see that the necessary support systems and checks and balances are present. So, for example, SSABSA in South Australia, which operates a 100% teacher assessed regime in what it calls a high stakes environment (the school-based assessment) has in place what it calls *assessment standards validation* which follows on from a process called *assessment standards support.* This is essentially preventative in nature and aimed at getting teachers to a point where they deliver school-based assessment that does not need moderation (Keightley and Coleman, 2002). Similar aspirations attached to consensus moderation in the 1960s and 70s and attaches to the procedures used for ensuring the acceptability of assessments made in many vocational and occupational qualifications, and to the recently proposed and currently piloted system in Scotland.

7.11 We assume that any system of statistical moderation, however sophisticated, is bound to give trouble sooner or later; there will be some occasion when, without necessarily understanding the reason, a student, parent or teacher perceives that a moderated result is not what was expected. In a system where there is imperfect transparency, what do you do? Unlike a moderation system that relies on judgement you cannot logically re-mark or adjust individual results – you have to either do nothing or do some recalculation that will affect the much wider cohort of candidates of which the complainant is a member.

The ‘company you keep’ factor

7.12 This is easily the most significant implication of using statistical moderation. The inescapable fact is that while performance on an external calibrator is not affected by who else takes the test or examination (assuming, that is, that it is marked objectively and not comparatively), teacher assessment that is moderated by group – as it has to be so as to harmonise teachers’ varying metrics – is inevitably affected by who else is in the group, and how they fare on the external calibrator. This is the ‘company you keep’ factor that has been extensively discussed by Wood (1978, 1991) and others. In his 1978 paper he analysed the scaling by different methods of two groups of candidates who have taken different papers within a differentiated set on the basis of scores on a common paper taken by all and shows that different methods produce different final rank orders for half of the candidates. He demonstrates that this is attributable to the performance of the group that each candidate is in, on the basis of the differentiated paper taken; this situation is close to the membership of different class groups assessed by different teachers.

7.13 The reasons for the problem that Wood encountered are ones that we have seen earlier in this review. In his rank allocation method, for example, the placing of individuals on the final mark scale depended on the relationships between the mark intervals and the ranks; raise or lower an individual’s mark within the rank order and the final rank (derived from mark combination) may be changed. Where he used linear scaling the different bivariate relationships for the two differentiated papers produced different scaling for an individual depending on the group performance. Backhouse, who had discussed the methods in an earlier paper (Backhouse, 1976) commented on Wood’s analysis suggesting that many of these effects would be smoothed out where candidate numbers were large and many of the discrepancies would be ‘lost’ within the grading system. Whilst both arguments may be true in a differentiated examination paper context the first is not true for moderation of teacher assessments which characteristically deals with small groups.

7.14 Theoretical studies confirm the existence of the problem. A study using simulated scores showed that (a) the size of the group, (b) the presence in the group of one or more students whose examination scores are significantly better or worse than their school based assessments and (c) the position of the individual in the group’s rank order, have an important influence for individual outcomes. Decreasing the group size raises the scores of good students and decreases the scores of weak students. The distorting effect of poor examination performers on the group's moderated scores is not necessarily eliminated or reduced by the presence in the group of good performers (MacGregor, 1987). We are not aware of any simulations that have translated a variety of outcomes into grades, in order to examine the variability arising from different group memberships and moderation methods but we must assume that, while the 'company you keep' factor does not affect all candidates subjected to statistical moderation, it may affect some to an extent that would alter a grade outcome. How that number relates, for example, to the number affected by low marking reliability, is unclear.

7.15 Evidently steps should be taken to ensure that, as far as possible, the 'company you keep' factor does not affect some students adversely and critically. On the limited evidence available rank order methods appear to be more vulnerable than scaling methods of moderation, particularly if either steps are taken to ensure near-linear distributions or acceptable transformations can be applied to make them linear, before using scaling. The trouble is that the 'company you keep' is insidious and it is difficult to locate and deal with the 'flop' scores that often produce anomalous results, although procedures such as LOWESS are available and in use, as in South Australia (see Section 5).

Small candidate numbers

7.16 Small numbers within a class or group will always require vigilance and we have already seen that special arrangements have been proposed or implemented in order to deal with them (see the discussions in Section 4 on moderation methods, in Section 5 on practices in Queensland, Victoria, New South Wales, Sweden, Hong King and West Africa in the previous discussion of the 'company you keep' factor. Thus, the smaller the number the more likely a misleading score distribution will result and the greater the chance of affecting adversely and critically some students' moderated teacher-assessed scores chiefly because of the risk that a non-linear relationship will develop between the teacher-assessed and calibrator scores. Equally there does seem to be evidence and experience that methods such as rank order scaling may deliver injustices where groups are small and should therefore be considered unsuitable for widespread use in statistical moderation of teacher assessed scores. The problem is closely linked to the possible presence of 'flop' scores.

7.17 We may ask: what is a small group? Some boards take the view that groups of less than 20 may have to be moderated or checked by inspection but this is just an arbitrary number that seems to have been quite widely used. Walker (1979b) goes back to the problems arising from flop scores in the context of the development of the Scottish Standard Grade examinations in the wake of the Dunning Report of 1976 and in relation to linear scaling. He comments that

> If a member of the group makes an unexpectedly low score in the external examination, the group average is depressed and hence the average of the scaled assessment is also depressed. The point is that it is not only the unfortunate low-scoring pupil who suffers but also his or her fellows in the group. If the number is ten or more, the average effect on each is likely to be small, but with a number as low as five the effect is correspondingly greater.

He then investigated teacher assessment data from 74 small school groups (between 5 and 9 in number) across 9 subjects and observed that most of them did not appear to contain any flop scores. Although a few correlations were low this was often because the marks of the group were very similar; most correlations were respectable. He suggested that low correlations should not be a sole basis for rejecting teacher assessments from a particular centre, felt that flop scores would not be a major problem and that linear scaling could cope with groups of this size.

7.18 In its wish to avoid the effects of an anomalous performance from one member of a small group, Western Australia requires that each group of 10 or fewer enter into a partnership arrangement with a parallel group in another school; preferably this should be a larger group but sometimes three small groups form a partnership. These arrangements are entered into at an early stage in the programme and the Curriculum Council recommends that larger groups (perhaps up to 15) form partnerships as a hedge against students dropping out; the partnership schools are then required to "combine (or 'match') their school assessments" by specifying in advance a moderation process that will be approved as a method for generating a single score scale that can be entered into statistical moderation. It appears to be the schools' responsibility to specify, manage and conduct this process, but this is assisted by Curriculum Council officers. It will normally involve some form of internal moderation by inspection or consensus but each school can opt out of the moderation of the partnership and be moderated alone or have its results adjusted through manual intervention by officers.

7.19 Victoria operates a similar partnership system with similar formal requirements for its operation. The closest parallels in the UK are the QCA and awarding body requirements that internal moderation generates whole-centre teacher assessed scores that can be treated as whole groups for moderation purposes. This would not entirely solve the small-group problem in statistical moderation since some subjects may always have one small group in each school. Moreover, whether this internal moderation is actually conducted or is effective is unclear and the requirements may not be sufficiently rigorous to address the 'company you keep' factor.

7.20 The problem of small group numbers has led to many subjects' exclusion from statistical moderation procedures since the advantages of speed and low cost are offset by the large number of interventions required. This also raises the question of consistency of treatment across the candidate entry: if some are being moderated statistically and some by inspection can we be sure that there is fairness in the treatment?

What to do when teacher assessment is manifestly 'wrong'

7.21 The QCA Code of Practice that all awarding bodies must follow for GCSE, GCE, VCE and GNVQ is quite clear that the awarding body

> …must require centres to standardise assessments across assessors and teaching groups
>
> (para 84)

and

> …should ensure that adjustments do not change the centre's rank order, unless the centre marks are demonstrably inconsistent.
>
> (para 89)

7.22 For teacher assessment to be manifestly 'wrong' either the teacher's rank order would have to be 'wrong' or the spaces the teacher wants to see between the ranks would have to be 'wrong'. The first is by far the more serious because it has been axiomatic that the teacher's rank order is always correct and all the procedures that we are discussing in this paper are based on that assumption. However, if you further believe (as, for example, the New South Wales Board of Studies says it does) that the spacings must be preserved then you are in effect endorsing the teachers' judgement in everything but allocating marks on a metric that coincides with the state-wide standard.

7.23 Evidence of a problem with teacher assessment scores that are being subjected to statistical moderation would emerge as unusual mark distributions (perhaps extremely skewed, with many candidates located at the same mark or very different from previous years) or as correlations that are either unusually low or unusually high (we could expect awarding bodies to develop a range of expectations for correlations, based on experience). These would suggest that either the teacher had violated the rank order assumption or that the teacher had represented the marks along the mark scale in an inappropriate way. In either case the remedial action would be the same - undertaking an inspection moderation process - and the implication would be that the centre or teachers would have the opportunity to improve their marking to bring it nearer to the required standard. With that comes a valuable learning implication that you just do not get with statistical moderation on its own.

7.24 However, correctional processes such as this do not necessarily establish whether the assessments were actually 'wrong' or just unusual. The process does not generally establish the reasons or special circumstances that may have led to the unusual score distributions or correlations and, unless there is some sort of inspection of teacher assessment procedures or some judgemental scrutiny of assessment outcomes on a more long-term basis, it is not clear how you would become aware that teacher assessment is actually 'wrong'. It follows that systems that flag centres as 'at risk', and then put in place procedures to ensure the teachers are given appropriate training and support to get it right next time, must be on the right lines. Such an approach distinguishes the immediate 'fix' that corrects apparent anomalies that have emerged from statistical moderation from the longer-term processes that attempt to develop an understanding of why the anomalies have occurred and, where necessary, take action to develop better practices in future. If the teacher assessments were 'wrong' a statistical moderation process is unlikely to put them right and if the teacher assessments were simply unusual a statistical moderation process is likely to do more harm than good. Either way, statistical moderation on its own cannot deal with these circumstances.

Other pressures on teacher assessment

7.25 Whilst it is not a key technical issue it may be important to consider briefly how far pressures and influences that arise from the wider conduct of the education and assessment systems might impact on the feasibility or validity of statistical moderation in the UK in, say, the conduct of GCSE. We have already suggested that other moderation methods will offer more interaction with the process of making teacher assessments than does statistical moderation and that the relative complexity and obscurity of statistical processes carry the risk of concealing injustices that may be difficult to fix for individuals.

7.26 We have also noted (as a footnote to the brief report of procedures in New Zealand in Section 5) that statistical moderation is at one end of a spectrum of approaches that has widened over recent years, and that it has the major defect of not constructively interacting with the processes of assessment by teachers. That is to say, although it does generate a final score that can be reported to a teacher or a student it is incapable of offering an explanation for the increase or reduction in the mark that relates to any specific aspect of the student's achievement or the teacher's decisions about it. In an environment where increased attention has been paid to developing local expertise in assessment, providing teachers with tools that will enhance their assessment expertise, and linking summative assessment to the formative processes of the classroom, this is a very serious criticism.

7.27 In a high stakes assessment environment in which there are strong lines of accountability for the quality of educational provision students and teachers are very careful to identify the strategies that are most likely to lead to qualifications success. There is increasing evidence from the post-16 sector of a much more pragmatic approach that is particularly focused on achievement in curriculum-embedded tasks that contribute to a qualification (Ecclestone & Pryor, 2003; Weeden & Winter, 1999) and participants will question procedures and decisions that affect the grade outcome. Procedures inside a statistical moderation 'black box' operated outside the school or college may be increasingly unacceptable in a climate of individual and organisational accountability.

7.28 The existence of league tables and performance-related pay are specific aspects of professional accountability as is Ofsted inspection and the regulatory role of a body like QCA. All relate to questions about the professional competence of a teacher and the extent to which he or she has been equipped to conduct assessment. Whilst other methods of moderation or systems of verification offer the possibility of contributing to professional development, statistical moderation, when operated on its own, looks like a throw-back to an earlier era. It is intriguing, therefore, to wonder why it has retained its place in Australia when it has been abandoned or come to be seen as a back-up in many other countries.

8 Models for use with notional assessment systems

An appropriate proportion for teacher assessment

8.1 We have been asked to consider what forms of statistical moderation would be appropriate in two notional cases:

- A GCSE in which there is 100% teacher assessment
- A GCSE in which there is 50% teacher assessment and 50% external examination possibly using an additional monitoring test

We have not discussed the question about what is an appropriate weighting for teacher assessment within an assessment scheme although decisions about weighting have historically been taken as much in relation to public perceptions of rigour and dependability of teacher assessments as to technical issues over the objectives for the assessment or the technical problems of moderation. Weightings have varied between 0% and 100%, but we were invited to consider only these two possibilities

8.2 Before discussing preferred approaches to moderation in these two cases we present a brief summary of the main findings that have emerged in this review, recognising that the research evidence is slight and the experience of statistical moderation very limited and fragmentary. We have commented briefly on the relationship between statistical and other methods of moderation since it must be clear from the foregoing discussion that we regard a simple ‘technical’ discussion of the statistical methods, divorced from wider contexts that include the viability of alternative methods, as a rather sterile approach to the problem.

Summarising the findings

8.3 The following tables summarise the evidence that we have concerning the various types of external calibrator that may be used. We have put our interpretations and judgements in italics.

Table 1: Types of external calibrator

Requirements from paras 1.9 & 1.10	Aptitude tests	General ability tests	Subject-based tests	Examinations	Curriculum-based tests	Item banks
1 calibrator results should correlate positively with teacher assessed scores in each subject in which it is used	Not tested with teacher assessment but low and variable (between subjects and between occasions) with A level and first year university scores	Not tested with teacher assessment but promisingly high and consistent with GCSE (after low and variable with CSE and GCE)	*Probably; it is assumed that tests with sufficient power can be designed*	High correlations but variable between subjects and between centres (NZ, NSW, HK, WAEC)	Modest (ASAT), high (Queensland CST)	*Probably capable of delivering tests of any type, at a price*

Requirements from paras 1.9 & 1.10	**Aptitude tests**	**General ability tests**	**Subject-based tests**	**Examinations**	**Curriculum-based tests**	**Item banks**
2 calibrator should be such as to minimise undesirable backwash effects on learning and teaching	*Unlikely as teachers may feel it necessary to give practice in such tests. In 50:50 GCSE we could have students practising (& taking mocks in) two types of test (external examination & external calibrator)*	*Unlikely as teachers may feel it necessary to give practice in such tests In 50:50 GCSE we could have students practising (& taking mocks in) two types of test (external examination & external calibrator)*	*Likely but in 50:50 GCSE we could have students practising (& taking mocks in) two types of test (external examination & external calibrator)*	*In 50:50 GCSE the external examination will have some backwash effect, likely positive; in 100% teacher assessment also but the external examination may be resented as unnecessary*	*Should be none as external calibrator is curriculum-based but in 50:50 GCSE we could have students practising (& taking mocks in) two types of test (external examination & external calibrator)*	Possibly but danger of tests being seen to determine the whole of teacher assessment (Victoria). May add to confusion of interface between formative/ summative assessment
3 calibrator should appear to teachers and students to be appropriate in the context of each subject in which it is used	Lack face validity in many subjects	*Unlikely to be seen as equally appropriate across all subjects*	*Likely*	*Likely*	*Should be as external calibrator is curriculum-based.* Seen to be of value that it is reported in its own right in Queensland	*Likely*
4 operate equally across all approaches to teacher assessment that are to be moderated	*Approaches to teacher assessment may differ between subjects? Requires the 'outcome' of the teacher assessment to be a mark, not a grade, pass/fail. May not be justified in 50% teacher assessment alongside an external examination; possible with 100% TA*	*Approaches to teacher assessment may differ between subjects? Requires the 'outcome' of the teacher assessment to be a mark, not a grade, pass/fail. May not be justified in 50% teacher assessment alongside an external examination; possible with 100% TA*	*Requires the 'outcome' of the teacher assessment to be a mark, not a grade, pass/fail. Unlikely to be justified in 50% teacher assessment alongside an external examination; possible with 100% teacher assessment*	*Requires the 'outcome' of the teacher assessment to be a mark, not a grade, pass/fail. Could be seen as sensible in 50% teacher assessment, not in 100% teacher assessment*	*Requires the 'outcome' of the teacher assessment to be a mark, not a grade, pass/fail Unlikely to be justified in 50% teacher assessment alongside an external examination; possible with 100% teacher assessment*	*May determine the approach to teacher assessment*

Requirements from paras 1.9 & 1.10	Aptitude tests	General ability tests	Subject-based tests	Examinations	Curriculum-based tests	Item banks
5 be free of particular bias (such as with respect to gender or ethnicity)		Different correlation patterns for males and females			Some bias found in ASAT	*Banked materials may be pre-tested for suitability*
6 be cost effective in development and administration	*Likely to be very costly to develop as we are not even close currently*	*Costly to continue development and maximise applicability across subjects and control bias (and renew on an annual basis) but then few costs involved in administration*	*Costly to develop and administer for all subjects (even for a few major subjects) (and renew on an annual basis)*	*In 50:50 no additional costs in development and administration; no advantage for 100% teacher assessment GCSEs*	*Cheaper than subject-based tests as there is only one. Would need annual renewal* *We have Key Skills tests currently though these may not provide a sufficient basis*	*Costly to develop; maintenance may be relatively economical. Web-based may cut administration costs but systems of security needed*

8.4 The following table summarises the evidence that we have concerning different methods of statistical moderation that are available.

Table 2: Approaches to statistical moderation

Requirements from para 1.10	Linear scaling (adjust TA mean, adjust TA mean & SD, linear regression, multiple regression)	Curvilinear regression scaling	Mapping/ rank order scaling	Non-statistical methods of moderation	Not moderating at all
7 its effectiveness in making adjustments to teacher assessed scores across the range of centres, candidates and subjects where it will be used	Problems with low correlations, skewed mark distributions, centres with few candidates	May be especially vulnerable in small groups where 'flop' scores occur	Vulnerable to the 'company you keep' factor. Care needed with 'flop' scores. Regarded as less useful than linear scaling and is theoretically suspect	Moderation by inspection (perhaps using statistical methods for alerting) widely used and seen to be dependable	There is already trust of teachers' rank orders but not of their perceptions of standard. *No adjustments involved; standards are front-loaded through information and training*

8 its transparency and acceptability as a method including the demands that its use places on centres	Becomes less transparent as we move from a) simple adjustment of mean to d) multiple regression. No extra demand on centres	Not particularly transparent. No extra demand on centres	No extra demand on centres	*Transparent only if feedback mechanisms are instituted.*	*Responsibility lies with the teacher or centre.*
9 the capacity of moderation to support the development of teachers' professional competence in assessment	None*, may have negative effect*	None*, may have negative effect*	None*, may have negative effect*	*Has the capacity, but may not always happen*	*No intrinsic methods of inspection or feedback; professional development programmes could reinforce or develop good or bad practice*

8.5 We will not discuss non-statistical methods or the use of non-moderated teacher assessment results any further but we do want to recall that, many years ago, Smith made the point that it is

> probably just as unwise to place all one's faith in the moderating instrument and to adjust candidates' internally assessed marks in strict accordance with performance in it as it would be to accept the internally assessed marks without applying any kind of moderating technique.
>
> (Smith, 1978, p.26)

Smith implied that not to moderate would be unwise even if to moderate uncritically might also be unwise. We find no compelling reason to opt for statistical moderation in place of the most immediate alternative of moderation by inspection with statistical checks with follow-up remedial action where risk analysis suggests it is needed. A system of the kind we are sketching out might be less expensive than competing systems, unless there happen to be high quality external calibrators on the shelf that are ready to go. A system based on support, preparation and risk analysis, but without external moderation, might be extremely difficult to contemplate but should be on the table as an option.

8.6 On the basis of the summaries shown in Table 1 we consider that

- aptitude and general ability tests do not have a sufficiently strong track record in statistical moderation to be considered at present; the experience in Victoria appears to be with a test that is closer to a curriculum-embedded test of the Queensland type than to earlier general ability tests and the UCLES test has not yet been shown to be sufficiently powerful for calibration of individual scores for certification
- subject-based tests, though potentially powerful, will be relatively expensive to develop on the required scale, would be hard to justify where an examination was

available to be used as a calibrator and would not add any new dimension to the reporting of students' achievements

- curriculum-based tests of the Queensland type, if developed within a whole-curriculum skills framework, may have sufficient power for moderation purposes across many subjects; they are less costly to develop than subject-based tests and have the added attraction of being able to be reported in their own right
- banks of materials may be used for moderation purposes and could be subject-specific or of a generic skills type; banks will be expensive to develop and great care will be needed to see that the assessment tests or tasks do not constrain learning

8.7 On the basis of the summaries shown in Table 2 we consider that

- the various forms of linear scaling have the considerable advantage of being easily understood, particularly where only the mean or mean and standard deviation are scaled
- equipercentile scaling carries particular risks where group size is small and where 'flop' scores occur but does cope with non-linear score relationships; methods of transformation such as the use of a polynomial, though mathematically sound, might be extremely difficult to explain
- rank order scaling carries risks with small groups and 'flop' scores, has been shown to generate anomalies and is theoretically suspect; we do not think that it should be used

A GCSE in which there is 100% teacher assessment

8.8 We are here discussing a GCSE as a whole where all assessment is by teachers. We assume that this assessment will be embedded in the curriculum.

- The introduction of a subject-based test would damage the credibility of the teacher assessment.
- It would be possible to develop banks of tasks for embedding in the curriculum, to sample the knowledge and skills and act as a benchmark for moderation of the teacher's own assessments by linear scaling of the mean or mean and standard deviation; however, the process is cumbersome, would be costly and would work better with moderation by inspection or consensus.
- The only credible and wholly external calibrator would be a curriculum-based test that could be used for a range of subjects and reported in its own right. A test that was based on generic skills would have the considerable advantage of supporting the development of these skills without specifying the nature of their embedding within the subject in question. This is our preferred approach particularly since the test would be taken once by all students, irrespective of the subjects being taken.

A GCSE with 50% teacher assessment and 50% external examination

8.9 We do not think that this model allows us to use a subject-based test. We assume that not all GCSEs would follow this model and that some might be assessed entirely by teachers. In that case there are three possibilities available.

- External tasks from a bank could be used as in the 100% teacher-assessed model but the justification of this method to enable statistical moderation is weak when an examination is available to be used as calibrator. This option would not be very credible and may not be at all cost-effective (but that does not prevent external tasks being made available as support materials or where other moderation methods are to be used).
- The most obvious option would be to use the examination as the external calibrator, using a method of linear scaling of the mean or mean and standard deviation.
- However, where a generic skills or other curriculum-embedded test were being used elsewhere it would be available for use as an external calibrator. The status of such a test would be enhanced if it were used for all subjects but that might entail some loss of power when compared to the examination.

The effect of unitisation

8.10 If the models that we have been asked to consider were to be applied to units within a GCSE there are some additional considerations such as whether one terminal moderation would be conducted (in which case the considerations are similar to those discussed above) or moderation would be unit-by-unit. Apart from the resource implications of the latter arrangement

- if a GCSE were to be composed of units that had varying amounts of teacher assessment the moderation method would have to be seen to be credible across all units
- it becomes difficult to envisage a method of statistical moderation other than one based on a single common test that can be used across all units although the power of such a test might be low for some units in some subjects.

8.11 If we imagine the unitised model to be organised as a set of modules that comprise a subject with very limited exchanges of modules between subjects and some optional modules or units (albeit tightly controlled) then we will have a system that tends to a large amount of common material within any certification. If we then ask which of the various categories of test has the greatest affinity with the unitised model it is the single curriculum-based calibration test that looks best conditioned for this particular task.

Group certification

8.12 Whilst each subject can be considered in relative isolation from others at the same level or others at other levels it is possible to identify a method of moderation that is most appropriate for that subject and for the type and extent of the teacher assessment within it. That is what happens at the moment, although awarding bodies will, of course, tend to favour systems of moderation that can be applied as widely as possible (perhaps with minor variations) so that they can be economical to operate. This capacity to tailor moderation methods to subjects is one feature of single-subject certification.

8.13 In the UK we do not have group certification within GCSE or A level or their Scottish equivalents. There are, however, implicit though loose assumptions that subjects have equivalent standards and demands that make it, for example, possible to generate UCAS scores, although there are also public perceptions that some subjects are more demanding than others. If these assumptions hardened to the point where group certification become a possibility we would have to ask whether there should be greater uniformity in the use of teacher assessment across subjects, including greater uniformity in the moderation method.

8.14 In such a situation the use of a curriculum-based test as calibrator for all subjects might be seen to be necessary and may be seen as a basis for monitoring subject standards or for aligning subject scores or grades prior to certification. Assuming that we continue to argue that an aptitude or general ability test would be a less effective and less justifiable external calibrator than a test based, say, on core, common or generic skills, we would have moved very close to the position in Queensland although the purpose would be slightly different. The incidental advantage might be that a candidate's result on the calibrator would almost certainly need to be reported in its own right, thus giving strong effect to the wish to embed generic skills across the curriculum.

8.15 At the same time, the publication of results in subjects *and* in a common calibrator exposes the statistical methods to a new type of public scrutiny. This might be seen to be a contribution to transparency but awarding bodies would need to be sure that some of the anomalies (such as those arising in small cohorts or for those candidates with flop scores) were well understood, had been dealt with fairly and that the whole process could be defended. It would also be more difficult to isolate the effects of these anomalies because the company kept by any one candidate could become much wider than his or her immediate cohort. Before embarking on the use of a common calibrator in this way it might be prudent to model a very wide range of potential anomalies and the methods that would be used to correct them.

8.16 In the light of this we might want to note that the use of a common calibrator to align subjects within group certification does not necessarily prevent other moderation methods being used for individual subjects within the group. It is possible that a two-stage process would first moderate the teacher assessments using whatever was seen to be the most appropriate method for that subject and then apply the results of the common calibrator to align the subject results. This would contain moderation anomalies within the subject and perhaps within a centre, making them easier to deal with but the curriculum justification for using the common calibrator is weaker, and it might be said that its contribution to the assessment burden was not balanced by its contribution to the candidates' certification.

Developing a test

8.17 It is 20 years since Forrest and Shoesmith underlined the extreme difficulty and expense involved in finding tests that are sufficiently fair, reliable and relevant to perform adequately as reference tests. It may be argued that the socio-political context of assessment is different now from twenty years ago. Although the stakes for students and teachers are now are higher would any tests that could be devised now be any more powerful than were available then? If Dexter and Massey's calibration test is superior to Test 100, which it certainly appears to be, and if Australian states are using calibrators as widely as they are, would a systematic, curriculum-based calibrator with reasonable correlations be possible? If 'reasonable' is between 0.5 and 0.7 that is definitely doable (although there are gender, ethnicity and subject differences to work on). If 0.6 to 0.8 is wanted a lot may need to be invested. To require all correlations to be routinely over 0.7 may be asking too much.

9 Proposals for further research

9.1 This review is not a sufficient basis to establish the viability of statistical moderation within the proposed model – there are too many unanswered questions and relatively little recent UK experience of the method. The educational, social, political and economic context of examinations has changed a great deal in the last 20 years, demanding a reappraisal of some issues last discussed in the 1970s and 1980s and of whether there now exists a technical capability that would meet some of the challenges associated with the use of statistical moderation.

9.2 On the basis of what we have written in this review we have drawn a number of suggestions for further work that needs to be done, and list these below with a brief indication, in each case, of the scope of the work that we think would be appropriate.

1. <u>Developing a curriculum-based calibrator</u>

   A thorough study is needed of whether some investment in development work would pay off, especially around a fixed calibrator to be used from year to year. A curriculum-based calibration test could be developed drawing on generic skills assessment experience and perhaps incorporating some of its materials.

2. <u>Rank orders and intervals</u>

   Develop a study that investigates more fully the characteristics of teacher mark allocations, to test the typical stability and dependability of rank orders and the stability and dependability of the mark intervals.

3. <u>Company you keep</u>

   Probably with the use of simulated data, explore typical effects of the 'company you keep' syndrome, particularly by simulating the effects of appeals by individual students. This is appropriate to any-sized groups but especially interesting for smaller student groups.

4. <u>Data cleaning</u>

   Developing appropriate algorithms for 'cleaning up' data; these are to be tried out and refined using real data and simulations and their acceptability and stability examined over a range of applications.

5. <u>Variable correlations</u>

   Trials to be made with prototype test versions (such as those proposed in 1 above) to check the variability and stability of correlations and regression equations. Particular attention to be paid to major group differences, such as those for gender, ethnicity and size.

6. <u>Efficacy of different models for statistical moderation</u>

   Initially through in-depth analysis of Australian experience, and subsequently through simulations appropriate to the UK situation, study the efficacy of linear and curvilinear regression for teacher assessment. This work could include studies of fine-tuning fitting procedures and of the effects of using test results to boost teacher assessment/external calibrator regressions where the latter is an external examination. There is also a need to examine the acceptability of these, paying particular attention to issues of transparency.

7. Weighting

   Investigate the effects of using various formulae for aggregating single teacher assessments which deviate from simple addition such as a formula which gives more weight to the most recent marks or judgements.

References and bibliography

Adams RM & Wilmut J (1981) A measure of the weights of examination components and scaling to adjust them**.** *The Statistician*, 30, 263-9.

Ademola A (1992) *The challenges of combining internal and external assessment in certificate examinations: the West African Examinations Council experience.* Paper presented at the annual meeting of the American Evaluation Association (Seattle, WA, November 5-7 1992).

Adeyegbe S O (1993) Correlation between continuous and terminal assessments at the senior school certificate examination (SSCE) level: The reality of a novel approach in the implementation of national policy on education *JORIC*, 7, 163-18

Amedahe F K (1998) *Models of combining continuous assessment scores with external examination scores for selection and certification*. PhD Dissertation, University of Pittsburgh

Backhouse JK (1976) Determination of grades for two groups sharing a common paper. *Educational Research* 18.2, 126-37

Bardell GS, Forrest GM & Shoesmith DJ (1978) *Comparability in GCE: a review of the boards' studies, 1964-1977.* Manchester: Joint Matriculation Board.

Black P (2002) *Tests and Assessments: Purposes and Quality.* Paper prepared for the first seminar of the Royal Society 14-19 science assessment enquiry.

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

Burton E & Linn RL (1994) *Comparability Across Assessments: Lessons From the Use of Moderation Procedures in England***.** UCLA**,** CSE Technical Report 369.

Choppin BHL & Orr L (1976) *Aptitude Testing at Eighteen-Plus*. Windsor: NFER Publishing Co.

Choppin BHL, Orr L, Kurle S, Fara P & James G (1973) *The Prediction of Academic Success*. Windsor: NFER Publishing Co.

Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836.

Cohen L & Deale R (1977) *The Assessment of Teachers in Examinations at 16+* (Schools Council Examinations Bulletin 37) London: Evans/Methuen

Cresswell MJ (1987) A more generally useful measure of the weight of examination components. *British Journal of Mathematical and Statistical Psychology,* 40, 61-79.

Cronbach LJ (1970) *Validation of educational measures*. In: Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

Crooks TJ (2002) Educational Assessment in New Zealand schools. *Assessment in Education*. 9.2. 237-354

Daley DJ & Seneta E (1986) Modelling examination marks. *Australian Journal of Statistics*, 28, 2, 143-153.

Daley DJ (1985) Standardisation by bivariate adjustment of internal assessments: sex bias and other statistical matters. *Australian Journal of Education*, 29, 3, 231-47.

Department of Education and Science/Welsh Office (1988) *National Curriculum Task Group on Assessment and Testing: a report*. London: DES/WO

Daugherty R (1998) Consistency in teachers' assessments: defining the problem, finding the answers. *British Journal of Curriculum and Assessment*. 8.1. 32-38

Dexter T & Massey AJ (2000) *Conceptual issues arising from a comparability study relating IGCSE grading standards with those of GCSE via a reference test using a multilevel model.* Paper prepared for the 22nd Biennial Conference of the Society for Multivariate Analysis in the Behavioural Sciences at the London School of Economics (17-19 July)

Doecke B, Raynolds G & Roberts A (Standardised Testing: What space for professional Judgement. *English in Australia*. 135. 5-8

Ecclestone K & Pryor J (2003) 'Learning Careers' or 'Assessment Careers'? The Impact of Assessment Systems on Learning. *British Educational Research Journal* 29. 4. 471-488

Eckstein MA & Noah HJ (1993) *A Comparative Study of Secondary School Examinations.* Research Working Paper 7. International Centre for Research in Assessment: University of London Institute of Education

Elley WB & Livingstone ID (1972) *External Examinations and Internal Assessments.* New Zealand Council for Educational Research

Fitz-Gibbon CT & Vincent L (1994) *Candidates' Performance in Public Examinations in Mathematics and Science.* A report commissioned by SCAA from the Curriculum, Evaluation and Management Centre, University of Newcastle upon Tyne.

Forrest GM (1981) 'Statistical Moderation' in *Combining Teacher Assessment with Examining Board Assessment*. Aldershot: Associated Examining Board on behalf of the CSE and GCE Boards

Forrest GM & Shoesmith DJ (1985) *A Second Review of GCE Comparability Studies.* Manchester: Joint Matriculation Board on behalf of the GCE Examining Boards.

Good, FJ (1988) A method of moderation of school-based assessments: some statistical considerations. *The Statistician*, 37, 33-49.

Good F & Cresswell M (1988) *Grading the GCSE*. London: Secondary Examinations Council

Govender P (2002) *A Critique of the moderation system of continuous assessment in the Senior Certificate as implemented in the Gauteng Department of Education, South Africa.* Paper given to the IAEA Conference, Rio de Janeiro

Govindarajulu Z (1988) *Alternative methods for combining several test scores*. Educational and Psychological Measurement, 48, 53-60.

Harlen W (2004) *A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes*. London: EPPI-Centre, Social Science Research Unit.

Hill PW, Brown T & Masters GN (1993) *Fair and authentic school assessment: advice to the Board of Studies on verification, scaling and reporting results within the VCE.* Melbourne: Victoria Board of Studies

Hill PW, Brown T, Rowe KJ & Turner R (1997) Establishing comparability of Year 12 school-based assessments. *Australian Journal of Education*. 41.1 27-47

HKEA (1988) *Review of the Public Examination System in Hong Kong*. Hong Kong Examinations Authority

Johnson S & Cohen L (1983) *Investigating Grade Comparability through Cross-moderation.* Schools Council: London.

Johnson S & Cohen L (1984) Cross-moderation: a useful comparative technique. *British Educational Research Journal*, 10**,** 89-97.

Kingdon MJ (undated) *Statistical Moderation - a mirage?* Unpublished typescript

Kingdon JM, French S, Pierce GE and Woodthorpe AJ (1983) Awarding Grades on Differentiated Papers in School Examinations at 16+. *Educational Research*. 25. 220-229

Keightley JV & Coleman MJ (2002) *School based assessment in a high stakes environment*. Adelaide: Senior Secondary Assessment Board of South Australia

Lewin KM (1992) *Science Education in Developing Countries: Issues and Perspectives for Planners.* Paris: International Institute of Educational Planning, UNESCO.

Long C (2004) Statistical Moderation based on the Examination Average. *IEB Newsletter* 11 http://www.ieb.co.za/Formal/Newsletters/issue11_StatisticalModeration.htm

Koretz D (1998) Large-scale portfolio assessments in the US: evidence pertaining to the quality of measurement. *Assessment in Education* 5.3 309-334

Luibisi RC & Murphy RJL (2002) Assessment in South African Schools. A*ssessment in Education* 9.2. 255-268

MacGregor M (1987) *Educational research: scientific or political?* Unpublished papers of the First Joint AARE/NZARE conference, University of Canterbury, Christchurch 3-6 December 1987, 12 pages. Christchurch NZ: Australian Association for Research in Education; New Zealand Association for Research in Education.

Massey AJ, Bramley T, Dexter T & McAlpine M (1998) *Calibration Test: Design, Development and Validation*. Cambridge: Research & Evaluation Division, UCLES.

Massey AJ, Green S, Dexter T & Hamnett L (2003) *Comparability of national tests over time: key stage test standards between 1996 and 2001*. London: Qualifications and Curriculum Authority.

Maxwell G (1994) ‘School-based assessment in Queensland’. Typescript copy of a chapter for a forthcoming book on Australian school curriculum reform edited by Collins C.

MacCann R (1995) Sex differences in the NSW Higher School Certificate after adjustment for the effects of differential selection. *Australian Journal of Education*, 39, 2, 163-188.

MacCann R (1995) *The Moderation of Higher School Certificate Assessments using a Quadratic Polynomial Transformation: a technical paper*. Sydney: New South Wales Board of Studies unpublished internal report.

McCurry D (1994) ‘Common curriculum elements within a scaling test for tertiary entrance in Australia’ in Kellaghan T (ed) *Admission to higher education: issues and practice.* Papers from the IAEA conference, Dublin, 1992 published by the Educational Research Centre, Dublin

McDonald AS, Newton PE, Whetton C & Benefield P (2001) *Aptitude Testing for University Entrance: A Literature Review*. Windsor: NFER.

McGaw B (1977) *The use of rescaled teacher assessments in the admission of students to tertiary study.* Report No 3. Brisbane: Research Branch of the Que4ensland Department of Education

Murphy J (1974) Teacher expectations and working-class underachievement*. British Journal of Sociology*, 25, 326-44.

Murphy RJL (1981) 'Statistical moderation - a critique' in *Combining Teacher Assessment with Examining Board Assessment*. Aldershot: Associated Examining Board on behalf of the CSE and GCE Boards

Murphy RJL, Wilmut J & Wood R (1996) Monitoring A level standards: tests, grades, and other approximations. *Curriculum Journal*, 1996, **7**, 279-291.

New South Wales Board of Studies (1998) *A Review of the HSC Assessment Program*. Sydney: Board of Studies.

New South Wales Vice-Chancellors' Conference (2002). *Technical Committee on Scaling: Universities Admissions Centre (NSW and ACT).* Silverwater: NSW.

Newton P (2003) The defensibility of National Curriculum assessment in England. *Research Papers in Education* 18.2. 101-127

Nuttall DL & Armitage P (1983) *The Moderating Instrument Research Project - summary report*. Open University/Technician Education Council

Nuttall DL & Armitage P (1985) *Moderating Instrument Research Project: A summary report.* London: Business and Technician Education Council.

Nuttall DL & Thomas S (1993) *Monitoring procedures based on centre performance variables.* Sheffield: Employment Department.

Nuttall DL (1971) *The 1968 CSE Monitoring Experiment*. Schools Council Working Paper 34. London, Evans/Methuen Educational.

Nuttall DL (1979) The myth of comparability. *Journal of the National Association of Inspectors and Advisers,* 11, 16-18.

Nuttall DL, Backhouse JK & Willmott AS (1974) *Comparability of Standards between Subjects.* London: Evans/Methuen Educational.

Pickering S (2004) private communication

Pitman JA, Matters G & Nuyen A (1998) *A case for Testing Generic Skills*. Paper presented at the 24th Annual Conference of the International Association for Educational Assessment (IAEA). Barbados, West Indies.

Pitman JA, O'Brien JE & McCollow JE (1999) *High quality assessment: we are what we believe and do.* International Association for Educational Assessment Conference, Bled, Slovenia. Queensland Board of Secondary School Studies: Spring Hill, Queensland.

Reeves DJ, Boyle WF & Christie T (2001) The relationship between teacher assessment and student attainments in standard test/tasks at Key Stage 2, 1996-8. *British Educational Research Journal*, 27, 141-160.

Ross A (1990) *A Preliminary Evaluation of the Enga Pre School Tokples Programme.* Dept of Education, Papua New Guinea.

Rowe KJ, Turner R & Lane K (1999) *A method for estimating the reliability of assessments that involves combinations of school-assessed tasks and external examinations.* Australian Association for Research in Education Conference, Melbourne.

Sadler DR (1992) Scaled school assessments: the effect of measurement errors in the scaling test. *Australian Journal of Education.* 36, 30-37.

Skurnik L & Hall C (1969) *The 1966 CSE Monitoring Experiment*. Schools Council Working Paper 21. London: Evans/Methuen Educational

Smith GA (1978) *JMB experience of the moderation of internal assessments.* Manchester: Joint Matriculation Board.

Spear MG (1989) The relationship between standard of work and mark awarded. *Educational Research*, 31, 69-70.

Stracahan J (1995) 'Moderation of assessments in the National Qualifications Framework' in Ajar D & Kandarakis HM (eds) *New Horizons in Learning Assessment*: proceedings of the 21st IAEA conference, Montreal

Talbot C (1995) 'Fair assessment of learning outcomes in South Africa school leaving examinations' in Ajar D & Kandarakis HM (eds) *New Horizons in Learning Assessment*: proceedings of the 21st IAEA conference, Montreal

Vassiloglou M & French S (1982) Arrow's theorem and examination assessment. *British Journal of Mathematical and Statistical Psychology*. 35. 183-192

Viviani N (1990) *The review of Tertiary Entrance in Queensland: report to the Minister for Education.* Brisbane: Board of Senior Secondary School Studies

Vernon PE (1957) *Secondary school selection* London: Methuen

Walker DA (1979a) 'The Standardisation of School Assessment' in Scottish Education Department *Issues in Educational Assessment*. Edinburgh: Scottish Education Department Occasional Papers

Walker DA (1979b) Scaling small groups at the O-Grade stage. *British Journal of Educational Psychology*, 49, 316-318

Weeden P & Winter J (1999) *Learners' Expectations of Assessment for Learning Nationally*. Report to QCA. University of Bristol Graduate School of Education CLIO Centre for Assessment Studies

Wiliam D (2003) National curriculum assessment: how to make it better. *Research Papers in Education* 18.2.129-136

Willmott AS (1977) *CSE and GCE Grading Standards: the 1973 Comparability Study,* Schools Council Research Study. London: Macmillan Education.

Willmott AS (1980) *Twelve years of Examinations Research: ETRU 1965-1977.* London: Schools Council.

Wilmut J (1977) *A Statistical procedure for the moderation of marks on a teacher assessed component of an examination.* Internal paper (RAC46) for the AEB Research Advisory Committee

Wilmut J, Wood R & Murphy R (1996) *A Review of Research into the Reliability of Examinations*. Discussion paper for the School Curriculum and Assessment Authority

Wilmut J (1999) *The use of internal assessment in qualifications*. A review of research for the Qualifications and Curriculum Authority

Wilmut J (2004) *Experiences of summative teacher assessment in the UK*. Review paper for the QCA Advisory Group on Research into Assessment and Qualifications

Wolf A (1996) *The feasibility of using a reference instrument for monitoring standards in GCE A level examinations*. A report to SCAA. London: SCAA.

Wood R & Naphthali WA (1975) Assessment in the classroom: what do teachers look for? *Educational Studies*, 1, 151-161.

Wood R & Wilson DT (1977) A technique for converting ranks into measures. British *Journal of Psychology*, 68, 321-326.

Wood R & Wilson DT (1980) *Determining a rank order when not all individuals are assessed on the same basis*. In: Van der Kamp LJTh., Langerak WF & De Gruijter DNM (Eds.) Psychometrics for Educational Debates. London: John Wiley & Sons Ltd., 1980.

Wood R (1976) Halo and other effects in teacher assessments. *Durham Research Review*, 7, 1120-1126.

Wood R (1978) Placing candidates who take different papers on the same mark scale. *Educational Research*, 20, 210-215.

Wood R (1991) *Assessment and testing: A survey of research*. Cambridge: Cambridge University Press.

Websites consulted or cited

SSABSA, South Australia: www.ssabsa.sa.edu.au

UCLA, USA: www.cse.ucla.edu/CRESST

ETS, USA: www.ets.org/etseurope

Queensland, Australia: www.qsa.qld.edu.au

Victoria, Australia: www.vcaa.vic.edu.au

UCLES, UK: www.ucles-red.cam.ac.uk

Western Australia: www.curriculum.wa.edu.au

New South Wales, Australia: www.boardofstudies.nsw.edu.au

EPPI Centre: http://eppi.ioe.ac.uk/EPPIWeb/home.aspx

IEB, South Africa www.ieb.co.za