FAQ 2: Why is there no statistical moderation in the redesigned system?

1. What the report said

On pages 62−63, 82 of Volume 1

> We see no requirement for − and indeed recommend against − the statistical scaling of teachers' assessments against the external assessment activity.
>
> Teachers' assessments should not be statistically scaled against the external assessment (Recommendation 5)
>
> However, the school assessment would not be statistically moderated against the external assessment.

2. .What you need to know about before you can understand the answer to FAQ2

- Subject Result in the redesigned system
- The notion of comparability
- The notion of combining scores from different assessments
- Nature and purpose of statistical moderation

3. Disclaimer

In providing answers to FAQs we are deliberately not using technical language. Our audience is not the educational measurement community and we trust that purists will not be critical of any over-simplification in our explanations.

4. Subject Result

A student's result in a particular subject, "Subject Result", is the simple sum of marks on four prescribed assessments − three school assessments and one external assessment.

SR = SA1 + SA2 + SA3 + EA

SR maximum = 10 + 10 + 10 + 30 = 60

The answer to FAQ1 explains the thinking behind a 60-point scale for reporting Subject Results. Whether there are 60 points or 80 points or some other number of points on the reporting scale the system has to be able to say that a "50" in a particular subject at your school is the same as a "50" in that subject at my school.

5. Comparability – in general

Meaningful comparability of students' achievements can be obtained only if the achievement is assessed reliability and validly. An assessment is valid to the extent that it reflects performance on the criteria on which teachers intended students to be judged – presumably set down in the subject syllabus. Reliable assessment comes about when marking schemes are applied consistently − across students and across judges (teacher-assessors).

Comparability can be accomplished only on the basis of a common measure, which may take the form of *teacher consensus* (in school-based assessment) or *common items/tests* (in external assessment). In each case − school assessment and external assessment − marking criteria are established during the design process.

6. Comparability – new school assessments and an external assessment

For the external assessment, comparability is assured through common assessments – same assessment for all students in a particular subject, at the same time, under the same conditions, and marked according to the same marking scheme, with double marking and marker monitoring optional for open-ended assessments, and with computer marking according to a verified key (list of correct options) for multiple-choice tests.

For each of the three school assessments, comparability is assured through two rigorous processes: first, a school's three assessments must be endorsed by QCAA before students can do them; and, second, a school's marking of each of the assessments must be checked by QCAA after the assessment and before the marks can be confirmed for counting towards subject results, with schools having to re-mark the work of all their students on that assessment should a problem be identified.

Given that comparability for each of the four assessments is assured on an assessment-by-assessment basis, there is no need for any further intervention before marks are summed to produce a Subject Result.

7. Statistical moderation

Statistical moderation is used to adjust assessment results from different sources to make them "comparable". The two "sources" of assessment results relevant to the present discussion are school assessment and external assessment. Why do they need to be comparable in the first place? Because they are to be added together to give a subject result and you can't add things together that are not on the same scale.

Statistical moderation is sometimes referred to as scaling or anchoring. It is probably not a good idea to refer to it as scaling in the context of this review because there are at least three ways in which the term "scaling" is used in the review report – all accurate in context but likely to be confusing here.

In many parts of Australia, statistical moderation is used to adjust the distribution of school-based assessments (school scores) so that the average and spread of the school scores match the average and spread of the school's distribution of results on the external examination (exam scores) for the subject. The adjusted school scores are then on the same scale as the exam scores. The adjusted school score (called "scaled score") for an individual student might turn out to be higher or lower than her original school score depending on the average and spread of the school scores for students in her subject-group.

Each student's adjusted school score can now be added to her exam score to obtain her Subject Result. (Sometimes a weighting is applied before scores are combined − if, for example, a system requires the school assessment to count for 70% of the total score rather than 50%).

In technical terms statistical moderation uses a *linear* transformation. A linear transformation does not change the rank order of students on the school assessments in a particular subject or the relative differences between students. Nor does it force the school assessments under a bell-shaped curve.

In the simplest of terms, statistical moderation constrains average performance on school-based assessment to be no better or worse than students' performances on an external examination.

Reasons for rejecting the "statistical" moderation of school assessments

The first of the four reasons given below is alone sufficient for rejecting statistical moderation of school assessments.

1. Statistical moderation is simply not needed as comparability is assured at the level of each of the four assessments: external moderation for the three school assessments one by one, and a common measure for the external assessment.

2. Even if statistical moderation were deemed necessary, the cost to the system in terms of effort and resources would be prohibitive because there would have to be a "full-on" external examination in every subject, even in those subjects that do not lend themselves to traditional external exams. Furthermore, if statistical moderation were deemed necessary and the government willing to pay for it and the school sector able to accept traditional external exams, the new system would suffer from the same demonstrated lack of understanding of complex scaling procedures as does the existing system (e.g. the QCS scaling that enables the OP calculation).

3. Some outcomes of the review – like introducing an ATAR (not elaborated on here) – can be rationalised in terms of the advantages of having a national approach to university selection. The statistical moderation referred to in the review, however, is part and parcel of the curiously Australian practice of scaling and combining results from different assessments to devise a rank order based on overall achievement. The argument that some other states have statistical moderation and so therefore should Queensland does not hold up because the time has come for all states to question the relevance of ranking university applicants from the Year 12 completer population in terms of their overall achievement in a collection of typically five subjects with no restrictions on subject combinations.

4. There are problems with the assumptions underlying statistical moderation. The scaling process assumes that one set of scores (from school-based assessment) mirrors the other set of scores (from an external examination). But the set of scores on the school assessment and the set of scores on the external examination refer to different student performances. It is possible that these performances are not being assessed against the same criteria. Even if they were, the standard of the performances could be quite different. That the school assessment and the external assessment in the redesigned model for Queensland should be equivalent in content and form – an assumption upon which statistical moderation is based – would seem to defeat one of the purposes of the external assessment which is to gather even more (and different) information about student learning and gathering it in different ways.

Equation used to scale school-based assessments

$$\text{MSA}_{\text{student 1}} = \{[(\text{SBA}_{\text{student 1}} - \text{Mean SBA}_{\text{school}}) / \text{SD SBA}_{\text{school}}] \text{ x SD E}_{\text{school}}\} + \text{Mean E}_{\text{school}}$$

Where $\text{MSA}_{\text{student 1}}$ is the moderated school-based assessment (scaled score) for student 1 in the school for a subject;

$\text{SBA}_{\text{student 1}}$ is the school-based assessment for student 1 in the school for a subject;

$\text{Mean SBA}_{\text{school}}$ is the mean or average of the school-based assessments for the school in the subject;

$\text{SD SBA}_{\text{school}}$ is the standard deviation of the school-based assessments for the school in the subject;

$\text{SD E}_{\text{school}}$ is the standard deviation of the examination marks for the school in the subject; and,

$\text{Mean E}_{\text{school}}$ is the mean or average of the examination marks for the school in the subject.

Notes

1. The terms, *score*, *mark* and r*esult* are used interchangeably.
2. Standard deviation is a measure of the spread of the scores. The mean is the average of the scores.
3. This FAQ answer is not the place to respond to the assertion that the purpose of statistical moderation is to check on teachers.