Guideline for calculating General subject results

December 2020

Queensland Government

Queensland Curriculum & Assessment Authority

For all Queensland schools

Queensland Curriculum and Assessment Authority
PO Box 307 Spring Hill QLD 4004 Australia
Phone: +61 7 3864 0299
Email: office@qcaa.qld.edu.au
Website: www.qcaa.qld.edu.au

Contents

Introduction

This guideline describes how subject results are calculated for students completing studies in General, General (Extension), and General (Senior External Examination) subject syllabuses in the Queensland Certificate of Education (QCE) system from 2020.

Marking internal and external assessments

Teachers make judgments about *internal assessments* (also known as school-based assessments) through the use of Instrument Specific Marking Guides (ISMGs). ISMGs, which are part of the General and General (Extension) subject syllabuses, describe qualities of student work for each internal assessment. As the name suggests, separate ISMGs are defined for every internal assessment, in every subject.

ISMGs identify criteria that relate to the relevant assessment, and describe levels of performance in each of these criteria. These levels of performance are awarded integer scores (marks), and a total score on an assessment is the sum of these individual component scores. The criterion marks also represent the relative contribution of each internal assessment to a student's overall subject result. An assessment worth 35% is marked out of 35 while an assessment worth 25% is marked out of 25. As ISMGs are specified in the syllabus, and remain relevant to assessment instruments across the life of a syllabus, they are comparable across years and between schools. The QCAA's endorsement and confirmation processes ensures this comparability.

Each question in an *external assessment* is marked according to an external assessment marking guide (EAMG) written specifically for that question. Obviously, as questions are different each year, the EAMGs vary each year. The difficulty of the assessment also varies to some extent from year to year, and not always in a uniform way across the ability range[1]. Differences in the external assessment from year to year must be accounted for when calculating an external assessment result that is comparable between years, and that is comparable to the internal assessment marking. This is necessary because the final subject result is constructed from the addition of the internal assessment results and the external assessment result. These two results must be linked, and equated, and equitable.

Creating a subject result using internal assessments for each General and General (Extension) subject

The process outlined in this guideline is based on an analysis method known as the Rasch model (Rasch, 1960; Masters, 1982). This model can be used to produce a measure (or scale) in a General and General (Extension) subject.

The Rasch model is used in a number of Australian and international contexts and is highly regarded and widely cited. Examples of where it is used includes:

- the National Assessment Program — Literacy and Numeracy (NAPLAN)
- the Programme for International Student Assessment (PISA)
- the Trends in International Mathematics and Science Study (TIMSS)

[1] As a simple example, all questions could be of similar difficulty between one year and the next, but the few hardest questions could be slightly more difficult in the second year. This will affect the difficulty of the whole assessment.

- aspects of the past Queensland Core Skills (QCS) Test
- construction of subject scores for Year 12 students in other jurisdictions
- the National Assessment Program — Science Literacy (NAP-SL).

It is one of the most highly respected methods for analysing educational assessment data worldwide, and is used in a huge number of research studies, as well as analysis of survey data and in other contexts.

The logit scale

The fundamental measure (or scale) that is produced from the application of the Rasch model is known as the *logit scale*. The logit scale is interval in nature. This scale allows comparison of student results in assessment instruments in a fair manner.

Common assessment instruments between students allow their *ability* (as it is known in the Rasch model) to be compared. This does not mean that all students need one or more assessment instruments in common, simply that if, for example students A and B have some assessment instruments in common and B and C have some in common, it also allows comparison of results for students A and C through the results for student B. This might be the case, for example, where a student has not undertaken one of the internal assessments (IAs).

The Rasch model assumes that there is a single attribute, an 'ability' which is being tested. The ability of students represents their knowledge and skills in the subject. The same scale can be used to represent an assessment instrument, that is made up of questions (referred to as items) and how difficult or easy that they are to answer known as 'difficulty'. In the Rasch model, the probability that a particular student answers an item correctly depends on two factors — their ability, and the difficulty of the item being assessed. Both difficulty and ability are positioned on the logit scale.

While logits can be rescaled to any arbitrary location, it is commonly described with an average of 0 and a standard deviation of 1 relative to either the performance of the students or the difficulty of the items. This means that in terms of performance, zero is around the middle, and positive and negative values are above and below that respectively, and most students will be in the range of about -3 to 3 logits in many cases.

The logit scale is essential to understanding the method. It represents the measure of how well students have achieved in the General and General (Extension) subject, based on the IAs that they undertook. The higher up the logit scale, the better the estimated achievement of the student. Also on this logit scale, we can place the location of the items. In the case of an item which is either right or wrong, they are located so that a student at that same location (that is, with the same logit value) has a 50% probability of getting that item right. If a student is higher up the logit scale than the location of the item on the logit scale, the probability of getting that item right is higher than 50%, and conversely a student who is positioned lower than the item on the logit scale has a probability of correctly answering the question of less than 50%.

The partial credit model

Many assessments, including those of the IAs, are marked using a polytomous scale (for example, A to E, or scored from 0 to 5) rather than dichotomously (right or wrong). A variant of the Rasch model known as the *partial credit model* (Masters, 1982) can deal with polytomous items. In this case, each score or grade on an item is placed on the logit scale in a similar way — that is, the probability of achieving a score of 3 on an item, for example, is located on the scale.

Each criterion on each ISMG might have a different number of marks against which it is graded. For example, an IA might have three criteria with 5, 3 and 7 possible marks respectively, marked from 0 to 5, 0 to 3, and 0 to 7. The difficulty of achieving a mark on each item will be different. It might be harder to achieve a mark of 4 on an item marked from 0 to 5 than it is to achieve a mark of 5 on an item marked from 0 to 7.

For simplicity of language in the following discussion, the term 'item' will refer to any one of the criteria by which any of the IAs are marked. For example, a General subject with three internal assessments, marked on three criteria each, is described as having nine items. In this example:

- the first item would be the mark on the first criterion on the first IA
- the second item would be the mark on the second criterion on the first IA
- the seventh item would be the mark on the first criterion on the third IA

and so on. This is represented in Table 1.

Table 1: Example of how items might map to IA and criteria

Item	IA	Criterion
1	1	1
2	1	2
3	1	3
4	2	1
5	2	2
6	2	3
7	3	1
8	3	2
9	3	3

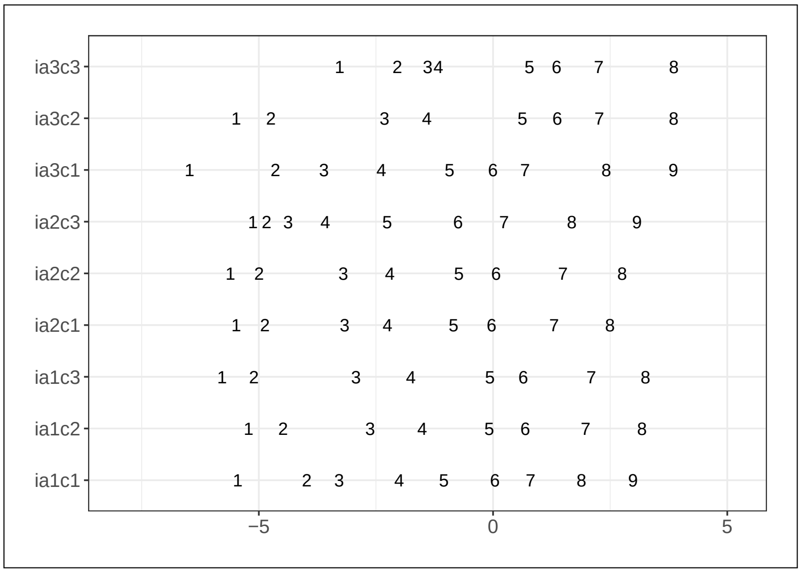
Linking scores and standards

Graphs that follow in this guideline will be drawn from an analysis of the General subject English in 2020. All Rasch analyses are performed using the TAM analysis package (Robitzsch, Kiefer and Wu, 2020) in the statistical package R (R Core Team, 2018).

The item map

Figure 1 shows the relative locations of scores on items.

Figure 1: An example of a map of item locations



In Figure 1, each row on the graph represents one of the criteria assessed in an internal assessment. In this case, ia1c1 represents the marks in criterion 1 (c1) on internal assessment 1 (ia1). Marks that were the easiest to achieve are on the left of the map, and marks that were the hardest to achieve are on the right. The difficulty of achieving each mark in each item is shown by the location of the numbers. For example, the relative difficulty of achieving a mark of 8 on ia2c1 and a mark of 8 on ia3c1 was about the same, even though the maximum mark on ia2c1 was 8 and the maximum mark on ia3c1 was 9.

Comparing ISMGs to syllabus reporting standards

This information is not only of use to subject experts in a number of ways, but also assists in the process of setting standards in the subject, where the overall final subject result in a subject must be translated to the ‘A’ to ‘E’ reporting standards in the syllabus. The method for mapping assessment results to syllabus standards is from a group of methods known in the literature as *item-descriptor matching methods*[2]. One summary of this group of methods is given by Cizek and Bunch (2007).

The ‘abilities’ of students, as estimated by their marks in all criteria, are also placed on this same scale as an aspect of the Rasch analysis. The higher up the scale, the better the estimated achievement of the student. Items are located so that a student at that same location (i.e. with the same logit value) has a 50%[3] probability of achieving that mark on that item. If a student is to the right of the position of the item, the probability of attaining that level of performance on that item is higher than 50%, and conversely a student who is positioned to the left of the item has a probability of attainment of lower than 50%.

The advantage, now that this scale describing achievement in the subject is constructed, is that we can also map syllabus reporting standards, in this case A, B etc., onto this same scale. Each item can also be matched by subject experts to the syllabus reporting standards. In this example, subject

[2] Nuanced variations of the method are sometimes referred to by different names in the literature, such as *construct mapping*, *item-mapping*, and *map marking*. This method is also discussed as a variation on the *bookmark method* (Karantonis & Sireci, 2006).

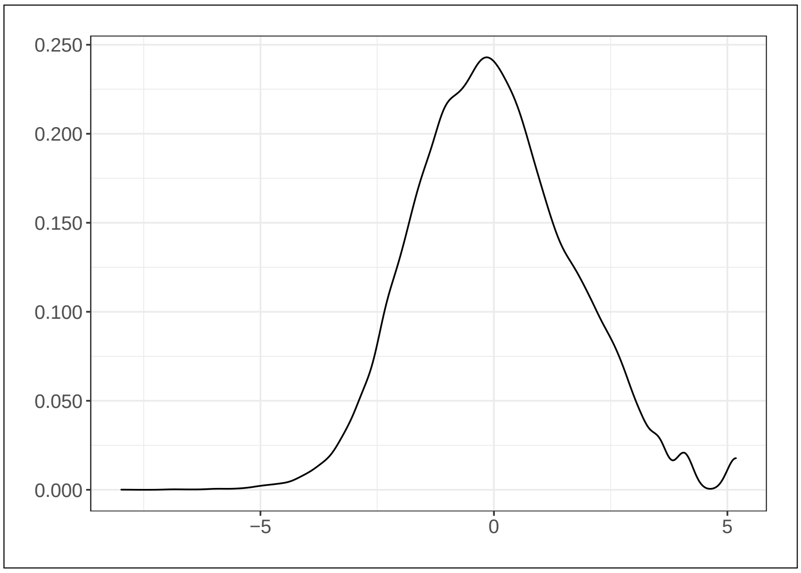
[3] There is discussion in the academic literature regarding the probability location of the items, with values between 50% and 80% being used. There is a suggestion that using values greater than 50% may lead to better outcomes — 67% is commonly used as a result, but 50% is often easier for people to understand and is used in this case. It ultimately should not affect the outcome.

experts can examine the syllabus reporting standard for, say, the ‘A’ standard and the level of performance required to attain 7 marks in ia1c1, and discuss the agreement between them. The idea is to identify the marks on each criterion on each assessment that corresponds to each standard. After this is done, the position on the scale that represents the boundary between two consecutive standards can be identified[4].

Comparing student work to standards

As described earlier, the logit values for students, as estimated by their marks in all criteria, are also placed on this same scale as an aspect of the Rasch analysis. A graph of the distribution of all abilities for students in English in 2020 is shown in Figure 2.

Figure 2: An example of the distribution of logit (ability) values for IAs in a subject



Once provisional cut-offs have been established, it is possible to find a student with the corresponding ability on the scale and see if their performance matches the standards expected. For example, a student with a logit value (from their profile of internal assessments) at the cut-off between an ‘A’ and a ‘B’ is identified, and their actual responses compared with the standards to verify that they match the expected response of a borderline A/B student. Other evidence and evidence from previous years are also used as a check — for example, the distribution of final grades are checked for reasonableness.

Updating the standards each year

This standards-setting exercise is done as soon as most of the internal assessment results are available, which is before the external assessment is marked. A full standards-setting exercise must be completed whenever a syllabus is used for the first time. In subsequent years, a confirmatory process of the mapping from previous years might be possible, rather than a full exercise. Monitoring of any changes in standards from one year to the next and confirmation of the standards together with the evidence from the external assessment is crucial and takes place each year. The model allows the relative difficulties, for example, between those marks on each criterion on each internal assessment, to be monitored.

[4] For example, the boundary between the ‘A’ performance standard and the ‘B’ performance standard on items.

Translating the cut-off to a raw score

Having established the location of standards relative to the logit scale, one advantage of a Rasch model analysis is that it also allows these locations to be mapped onto the raw score scale (in most cases). For example, in the example shown, if the location of the cut-off between ‘A’ and ‘B’ is at 2.5 on the logit scale, this might correspond to a raw score location of 67. The raw score, in this instance, is the numeric sum of the marks attained by the student in the internal assessments. It does not matter how a student got to that raw score — for example, a student with scores of 14, 4, 14, 4, 14, 14 and 3 in seven criteria for a total of 67 will receive the same score as a student with a profile of 4, 14, 14, 14, 4, 10 and 7 in those same criteria. The translation of this raw score to the logit scale takes into account the non-interval properties of these scores. For example, the difference between achieving a raw score of 2 and a raw score of 4 on the logit scale will not be the same difference as that of achieving a raw score of 20 and a raw score of 22.

Queensland’s history of standards setting

A critical aspect of this method of standards setting is the fact that it reflects a historical commitment in Queensland to the way that standards are set — evidence of student achievement and judgments relative to that achievement are still at the heart of the process. During standard setting, there are multiple ways in which qualities of student work, items and syllabus reporting standards are compared:

- We can pick up an item and relate it to a syllabus reporting standard — how does achievement in this criterion compare to the reporting standard descriptor?
- We can pick up a student’s response and relate it to a syllabus reporting standard — how does a student with this raw score compare with what we expect from a student at this syllabus reporting standard?
- We can check the distribution of results on different internal and external assessments to ensure that they relate the way we expect — how does the distribution of results on each criterion within each assessment compare?
- We can compare the difficulty of criteria with each other to ensure that they have performed as we expected — are there criteria that seem much easier or harder (relative to other criteria) than we expected?

Constructing external assessment results

Anchoring the scale

The analysis of the IA results has established a scale, on which (in logits) both student ability and the difficulty of achieving each mark on each criterion/IA is located. A next step that is often undertaken in analysing results using Rasch is to ‘anchor’ the location of the items on this scale. In this case, this means that, in further analyses, the location of marks in each criterion/IA (in logits) does not vary. Rather, the location of these marks keeps the scale itself in place.

An analysis is now conducted using both the internal and the external assessment results for each student, but with only the internal assessment results ‘anchored’. This allows the results in external assessment questions to find their own location on the already established logit scale[5]. After this

[5] The difference between ‘scaling’ (where student marks are moved using the results of other students), and ‘equating’ (which involves making all students’ marks comparable on the same scale) is critical, not just in terms of perception, but also relative to the purpose of the analysis. If about 20% of students get an ‘A’ on the internal assessment, this will translate to roughly the same number of students getting an ‘A’ on the external assessment. No student marks will be rescaled.

second analysis, the following values are located on this scale:

- internal assessment item difficulties
- external assessment question difficulties
- student ability estimates (based on their performance across both the internal and external assessments)
- standard or grade cut-offs.

Any values in this list can now be compared with each other as they are all on the *same scale*. Another benefit of having a single scale is more obvious when results for students who have missing assessments are considered.

Mapping the external assessment raw results to a 25- or 50-point scale

The challenge now is to map the *variable* raw external assessment marks onto a scale that is the same from one year to the next, and is therefore able to be numerically added to the internal result to construct a total final subject result. This involves running a Rasch analysis with *only* the external assessment marks on questions, but anchored to their already established locations on the common, logit scale. This third analysis produces a lookup table between external raw scores and the existing logit scale.

The three analyses needed in this process are summarised in Table 2.

Table 2: Rasch model runs needed for each subject in the QCE system

Run #	Items	Anchored items	Details
1	Internal	None	Used for standard setting and establishing the logit scale
2	Internal, external	Internal	Places marks on questions on the external assessment onto the logit scale
3	External	External	Provides lookup from external raw score to year-to-year comparable external score

Because it is desirable to have the internal and external assessments to be of approximately the same difficulty in terms of achievement, we can now equate raw scores on the external assessment to raw scores on the internal assessment by linking through the logit scale. For example, suppose that in a subject where internal and external assessments are worth 50%:

- a raw score of 147 on the external assessment has a logit value of 0.6
- an internal result of 35 on the sum of internal assessments has a logit value of 0.58
- an internal result of 36 on the sum of internal assessments has a logit value of 0.61.

Therefore, the external assessment raw score for a raw score of 147 on the external assessment would translate to somewhere between 35 and 36, and is comparable from year-to-year.

This mapping is performed for all external assessment raw scores. In subjects where internal and external assessments are not weighted equally, the internal result is appropriately transformed; for example, in a subject with 75% weighting on internal assessment and 25% external assessment, the location of the internal assessment result divided by three is used.

Figure 3 shows three example item maps (one for each analysis) to further illustrate the process.

Obviously, this is a simplified diagram that does not reflect an actual General or General (Extension) subject.

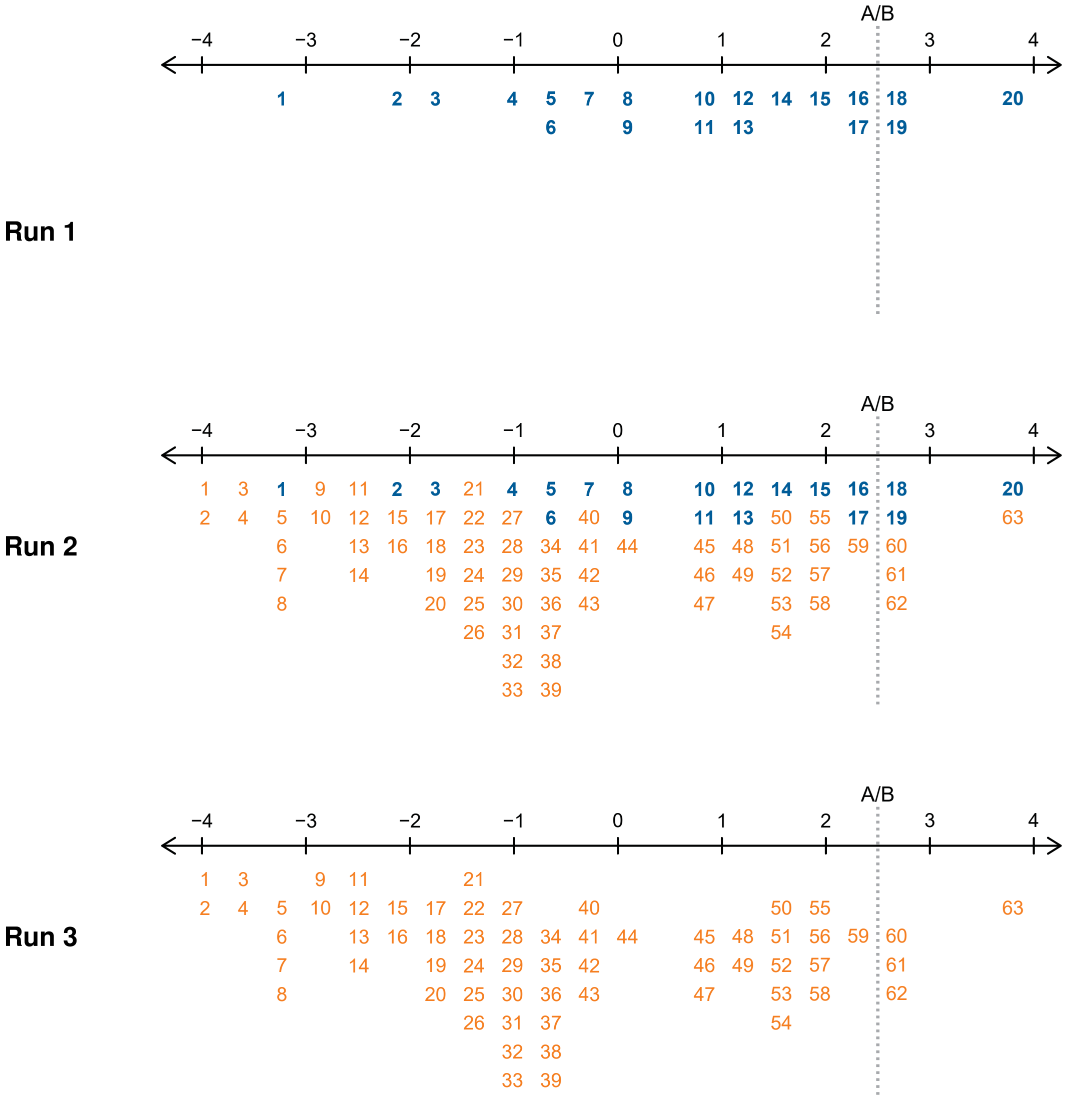


Figure 3: Example item maps for run 1 (top), run 2 (middle) and run 3 (bottom), showing the items involved in each run and their location on the logit scale. The example internal assessment marks are shown in blue and are numbered 1 to 20. The 63 external assessment marks on questions in this example are shown in orange. The location of a proposed A/B cut-off from a standards-setting exercise is also shown as an example.

This process accounts for variation between years on the external assessment. For example, in one particular year, 67 marks out of a total possible 135 marks on the external assessment could translate to 25 on a 50-point external scale but in the following year 84 out of 148 marks may map to that same point. In this example, students would receive a result on the external assessment that was out of 50 each year. This result would be comparable from one year to the next, regardless of:

- changes in the difficulty of the external assessment
- the fact that each year’s external assessment contains different items
- the fact that the maximum possible raw score may change over time.

Standards validation

Following the calculation of results in EAs, a second round of the standards-setting exercise takes place. This involves finding students with EA results that have a logit value corresponding to each of the cutoffs for standards set during the first standards-setting exercise, and confirming that their EA result also corresponds to performance on the corresponding boundary between each of the standards. If the performance does not correspond, the boundary is revisited by the subject experts using both the IA and EA results, and a new cut-off is determined.

Combining the internal and external results

An overall final subject result in a subject is constructed by simply adding the individual results in the internal and external assessments. The cut-off for the reporting standards are described on the overall subject result scale. The cut-off for an A, for example, might be a raw overall subject result of 73, and a student can calculate their own overall subject result by adding their individual assessment results.

The last step that is needed, then, is to find these cut-offs on the overall subject result scale, because the cut-offs for reporting standards are originally located on the logit scale from the item-descriptor matching procedure. For every student who completes all internal and external assessments, they will have both:

- an overall final subject result
- a logit score from the combined internal and external Rasch analysis.

A regression can be performed between the overall subject result and logit scales, and this will allow the logit scale to be mapped to the overall subject results scale. It is now also possible to map the location of the logit value cut-offs to the overall subject result scale. If a student has a logit score based on an incomplete assessment profile, this regression would also allow mapping of this logit score to a total subject score.

In summary, results on internal and external assessments are now validly and simply added together mathematically to construct a final subject result so that:

- results on internal assessments are the sum of the marks on individual criteria, as specified in the instrument specific marking guides
- the result on the external assessment is the year-to-year comparable external assessment score, which is reported to students (out of 25 or 50). The variable raw score on each year's external assessment is mapped to this scale before it is reported to students.

Cut-offs relative to the overall final subject results are published for each syllabus reporting standard. If the internal assessments and their ISMGs remain the same from one year to the next, these cut-offs should remain fairly stable, although they are reviewed for validity each year. If the reporting standards are modified in the syllabus, the standard-setting exercise is performed again. In the case of a more minor modification to the syllabus taking place, an analysis that anchors in place all of the unmodified aspects of the syllabus and determines the relative placement of the modified aspects of the syllabus is be used to inform the new final subject results and cut-offs.

Special procedures

Missing results

One obvious special case that must be dealt with is where a student does not have a complete set of assessment results that can be combined. Examples could be serious illness on the day of an external assessment, or a student who transferred partway through a course of study (whether from another jurisdiction or overseas), or between courses during the senior secondary years.

Application of the Rasch model describes a scale onto which a subset of results, whether missing internal or external assessments, can be mapped and compared with the location of reporting standards. As part of the analysis, results for students who are missing assessments can be estimated — the model lends itself naturally to means of dealing with missing results.

The case where different results are missing for different students can also be accommodated. One student might be missing a mark for only one criterion in one internal assessment, whereas another might be missing one complete internal assessment; another might be missing the complete external assessment, and another might be missing just some part of the external assessment. All of these students could have a logit score (and hence an overall final subject result) derived for them as a routine part of the analysis. Conditions where students are *permitted* to have an overall final subject result estimated when their result/s are missing is and exemption where the school and student need to apply to the QCAA.

It is because there is no scaling of individual results with group results that makes managing missing results in this way, without complex imputations of missing scores, possible.

General (Senior External Examinations (SEEs))

The process for calculating results in SEEs is different for language SEE subjects, and non-language SEE subjects.

Non-language SEEs

In SEE subjects, the final subject result is based solely on the EA. In the case of SEEs other than language subjects, this means that students are required to sit the same external assessment from the General syllabus as well as an additional external assessment that covers syllabus objectives that were assessed by internal assessments completed by students who studied the General syllabus.

This can be managed by adding a fourth step to the previously described process. After the logit values for questions on the EA for the students studying the General subject is performed, these results can be anchored and an analysis run on the entire EA for the students studying the SEE. In exactly the same fashion, following this, a lookup table from logit to a score on the 100 point scale can be performed.

Language SEEs

In the instance of SEEs in language subjects, there are no common assessment questions to use to perform a common analysis. In fact, a very similar situation exists for those language subjects where the external assessment is borrowed from other jurisdictions through the Collaborative Curriculum and Assessment Framework for Languages (CCAFL). In these instances, a score must be constructed out of 100 from the information available.

Having established a 100-point scale, it can be used for the purposes for which it is needed. Standards setting can take place separately on this 100-point scale and QTAC can scale the results separately.

EAs that are separate but with different weightings

In some General subjects students undertake more than one EA. In some instances, these two papers form one paper in the analysis, and so can effectively be treated as the same paper regardless of the EA papers not being sat at the same time. In some subjects, however, regardless of the differences between the papers, their contribution to the total EA score is a *weighted average* of the individual papers[6]. The sensible way to translate the individual scores is to translate the scores in each individual paper onto the (weighted) score for that paper. The combined EA score will be the sum of these weighted scores. If the two results are transformed independently to an interval, Rasch scale, then doing a simplistic, linear, weighted combination of the two can be performed mathematically.

Alternative sequence syllabuses with different EAs

Alternative sequence (AS) syllabuses have the same IAs and ISMGs (based on different content). For this reason, the AS syllabus and the parent General syllabus were treated in the same way when analysing the IAs. That is, the analysis for the General syllabus was used for the AS syllabus as well. However, in the case of a number of AS syllabuses, the EA is not the same as for the parent General syllabus.

In this case, the analysis for the AS syllabus is performed separately, but using the results for the IA in run 2. This means that the results in the EA will be transformed onto the IA scale for the parent syllabus. A potential complication of doing it this way is that there are likely to be very few students doing the AS syllabus, and so the data will not support a separate analysis of their EA results. This will be somewhat mitigated by the fact that the IA analysis will provide a stable basis onto which the EA will be translated. Manual checking of the outcome for each AS syllabus will be needed, and may require variations in individual years. Variations to processes are the role of the Ratification Committee.

Subjects with different IAs, but a common EA

Remembering that the IA analysis sets the logit scale, the situation for these subjects is unusual. In one case, for example, Music Extension in 2020, there are three separate syllabuses — Music Extension (Composition), Music Extension (Musicology) and Music Extension (Performance). All three syllabuses have different IA structures, and therefore have individual IA logit scales. However,

[6] If the distribution of results being combined are not taken into account, the implications can be significant. Take, for example, two assessment tasks that both have a maximum score of 10. Combining these two together numerically might imply that they are equally weighted because they both have the same maximum and minimum score. However, if one assessment task has a distribution of results that effectively only covers the range of 7 to 8 for almost all students, it will have less influence on the discrimination of the final result than one that covers the full range from 1 to 10 for the same students, even though 10 was the maximum score in each case. This effectively introduces a form of unintended weighting that could cause unfair advantage or disadvantage. This effect has been observed and analysed in a number of overseas contexts (for example, Rudner 2001; Wang, 1998; Adams and Murphy, 1982; Elwood 1999).

the external assessment is the same for all three syllabuses, and so will require the EA marks to be translated back to the three individual IA logit scales that were used in standards setting.

In normal circumstances, this should be handled by doing three different analyses, one for each of the three syllabuses. However, this will result in (potentially) the same raw score on the same EA translating to a different mark out of 25 depending on which of the syllabuses the student undertook. For this reason, it is best that all three EAs are translated from the same EA raw score to the same EA score. The most robust way of doing this is to perform the analysis based only on one of the three syllabuses (the most populous and with the most stable data), and using this translation table for all three subjects.

Variations to procedures

Inevitably, variations to procedures are required each year. This may be because of factors such as:

- exceptional circumstances in that year
- exceptional circumstances in individual subjects
- exceptional circumstances at individual schools.

Often, these circumstances are not known until very close to the point where results are due to be certified, and examination of the appropriate variation needs to take place in a timely fashion, and with advice from the sectors and independent experts.

A sub-committee of the QCAA Board, the Ratification Committee, reviews and endorses the processes for each year, including appropriate variations. Membership of the committee comprises:

- a QCAA Board member (chair)
- QCAA officer (executive officer)
- one representative from each of the three schooling sectors
- two independent experts
- Director Assessment, Reporting and ICT Systems
- Assistant Director, Measurement
- Assistant Director, Analysis and Reporting
- QCAA officers with technical and curriculum expertise as required.

Importantly, the committee membership contains:

- a representative of the Board, who is the chair
- a representative of each sector, nominated by the sector
- independent statistical experts

that are all important for the independence and validity of their decisions. Although the remit of the committee is broad, guiding principles include:

- the committee is, where possible, shown data that is de-identified by school, to avoid conflicts of interest that might otherwise arise
- the committee is to make decisions about groups, not individual students, as individuals are taken to the Illness and Misadventure Committee
- things that are specific to the calculation of subject results in the year in question
- trying to ensure that students who are affected by circumstances are neither advantaged nor disadvantaged by these circumstances.

A report on the outcomes of the committee are presented to the Board at the next possible Board meeting, but because of the emerging nature of problems that require consideration by the committee and the number of meetings that it needs, this may happen after the decisions have been implemented for certification in the current year.

References

Adams, RM & Murphy, RJL 1982, ‘The achieved weights of examination components’, *Educational Studies*, vol. 8, no. 1, pp. 15–22.

Cizek, GJ & Bunch, MB 2007, *Standard Setting: A guide to establishing and evaluating performance standards on tests*, Sage Publications Inc., California, ISBN 1-4129-1638-6, pp. 193–205.

Elwood, J 1999, ‘Equity issues in performance assessment: The contribution of teacher-assessed coursework to gender-related differences in examination performance’, *Educational Research and Evaluation*, vol. 5, no. 4, pp. 321–344.

Karantonis, A & Sireci, SG 2006, ‘The Bookmark standard-setting method: A literature review’, *Educational Measurement: Issues and Practice*, vol. 25, no. 1, pp. 4–12.

Masters, GN 1982, ‘A Rasch model for partial credit scoring’, *Psychometrika*, vol. 47, pp. 149–174.

R Core Team 2018, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, www.R-Project.org.

Rasch, G 1960, *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish Institute for Educational Research, Copenhagen.

Robitzsch A, Kiefer T & Wu M 2020, *TAM: Test Analysis Modules. R package version 3.5-19*, https://CRAN.R-project.org/package=TAM.

Rudner, LM 2001, ‘Informed test component weighting’, *Educational Measurement: Issues and Practice*, vol. 20, no. 1, pp. 16–19.

Wang, T 1998, ‘Weights that maximize reliability under a congeneric model’, *Applied Psychological Measurement*, vol. 22, no. 2, pp. 179–187.