Inter-Subject Comparability: An International Review

ISC Working Paper 4

December 2015

Ofqual/15/5799

Contents

Suggested citation:

Ofqual (2015d) *Inter-Subject Comparability: An International Review: ISC Working Paper 4*. Coventry, the Office of Qualifications and Examinations Regulation.

This report was written by Dennis Opposs (Standards Chair).

1. Introduction

We are investigating whether some GCSE or A level subjects can be considered harder or easier than others. If they can, we want to decide whether it would be beneficial to produce a new alignment and how that might be done. This working paper is one of a set examining technical, practical and policy issues in relation to inter-subject comparability. These papers are intended to throw light on the present position, stimulate informed debate and help us decide what to do.

The purpose of this working paper is to provide a broad overview of the methodologies used in high-stakes assessments in a variety of jurisdictions to align the results of those assessments in different subjects. The information on the assessments has been gathered using desk-based research of publicly available sources, so it should not be taken as definitive. This is very much an exploratory study. Chapter 2 describes more fully the methodologies.

Chapter 3 provides brief descriptions of the statistical methods that are used internationally to investigate and take action in relation to inter-subject comparability. Chapter 4 covers eight assessments where, in different ways, actions are taken using statistical methods to address inter-subject comparability. The text includes descriptions in each case of how the relevant statistical method is applied. Where we have managed to find relevant information, there is also a description of public perceptions of the statistical adjustments to assessment results.

To provide a contrast, chapter 5 describes a variety of international assessments where, as far as can be judged from the public sources available, such adjustments are not made. The final chapter, 6, is about lessons we might draw from this international experience for GCSEs and A levels in England. In those qualifications, we do aim at the design stage to ensure that, as far as possible, the demands in the breadth and range of content of different subjects are comparable, but we make no statistical adjustments across subjects before grades are issued.

We are publishing this working paper at this time and sharing the information more widely. We are seeking, and would very much welcome, feedback from in-country experts, those who are much closer to the operation of the individual assessments referenced in this working paper. That should allow us in due course to publish a fuller paper in which we can have more confidence.

2. Methodologies

Assessments

The review focuses on inter-subject comparability in high-stakes assessments – gateway assessments that enable students to access the next stage of education or employment. Most of the assessments reviewed are mainly or wholly concerned with university entrance, as most jurisdictions do not have assessments that are high stakes for younger students.

In some jurisdictions, entrance into almost all higher education institutions at undergraduate level requires students to take one particular assessment. An example of this is the *Gāokǎo* (National Higher Education Entrance Examination) used in China.

In other jurisdictions, there may be more choice. So, for example, in New Zealand, students in some schools are prepared for the International Baccalaureate, students in others are prepared for international A level exams, whilst students from most schools in the country take the national exam – the National Certificate of Educational Achievement (NCEA). All of these assessments would then be used directly in the university entrance process.

Another way in which choice happens is exemplified by the United States of America. Here, each university judges high school students on the basis of its own criteria. These might include ACT scores or SAT scores. Students wanting to enter a university might or might not take one of those assessments, or neither. However, some states, such as New York and Massachusetts, have their own state-wide exams, which are critical for students wanting to enter university.

Some of the assessments covered in this review are based closely on the curriculum that the students study in senior secondary school, for example the Leaving Certificate exams in Ireland. Others are designed more as reasoning or aptitude tests, for example the Psychometric Entrance Test used in Israel for university entrance.

In some jurisdictions, the assessment described here appears to be the sole criterion used for university entrance (for example, China). In addition, in some there are university-based tests as well (for example, Japan). In others, national exams or school-based assessments also contribute (for example, Israel).

In this study, the rationale for selecting the first set of assessments focused on the jurisdictions within which they were taken. We were guided by the following criteria:

- jurisdictions that were identified as high performing in international benchmarking studies – PISA (2012), TIMSS (2011) and PIRLS (2011);

- jurisdictions that were known to undertake specific methods of addressing inter-subject comparability;
- jurisdictions that had the greatest similarities in terms of assessment structure to England's system.

Other jurisdictions were then added to give greater geographical coverage and to ensure the inclusion of assessments that were better described as university entrance aptitude tests rather than achievement tests related to a taught curriculum.

Thirty assessments were reviewed in total:

- Australia: New South Wales Higher School Certificate
- Australia: Tasmanian Certificate of Education[1]
- Brazil: High School National Exam (ENEM)
- Canada: Alberta High School Diploma
- China: *Gāokǎo* (National Higher Education Entrance Examination)
- Cyprus: Pan Cyprian Exam
- Fiji: Fiji School Leaving Certificate
- Finland: *Ylioppilaskirjoitukset / Studentexamen* (Matriculation Examination)
- France: *Baccalauréat général*
- Germany: *Abitur*
- Ghana, Liberia, Nigeria, Sierra Leone and The Gambia: West African Senior School Certificate Examination
- Greece: Pan-Hellenic Exam
- Hong Kong: Diploma of Secondary Education
- International Baccalaureate Diploma

[1] All the Australian states and territories use scaling procedures to convert their end-of-school assessment outcomes into a score that is used as the main criterion for entry into most undergraduate courses in the country. The two states chosen here are, therefore, illustrative of the more general approach in Australia to inter-subject comparability adjustments.

- Ireland: Leaving Certificate
- Israel: Psychometric Entrance Test
- Japan: National Centre Test
- Kazakhstan: Unified National Test
- Netherlands: *Voorbereidend wetenschappelijk onderwijs* (VWO)
- New Zealand: (NCEA)
- Poland: *Matura* (High School Examination)
- Russia: Unified State Examination
- Singapore: PSLE)
- South Africa: National Senior Certificate
- Switzerland: Federal Maturity Certificate
- Taiwan: The Basic Competency Test
- Thailand: General Aptitude Test (GAT) and Professional Aptitude Test (PAT)
- UK: Scotland Standard Grade, Intermediate 1 and 2, Higher and Advanced Higher
- USA: SAT I and SAT II
- USA: ACT.

The assessments included in this study are all end-of-upper-secondary assessments that enable students to access higher education or employment, apart from two end-of-primary assessments, which facilitate access to selective secondary schools.

In reviewing the jurisdictions' approaches to inter-subject comparability, it is important to consider the context of the assessments. Integral to a jurisdiction's approach is the educational framework that determines the structure of the assessment. We have, therefore, categorised the assessment structures into three groups:

- Free choice: Students can select subjects of their choice to study and be assessed in (within this option there may be one or more compulsory subjects, but elective subjects are chosen from a broad menu).
- Restricted framework: Students can select subjects from pre-defined, limited subject groups in which to be assessed.

- Uniform subjects: Students are all assessed in the same subjects.

Approach

The review was undertaken as desk research, focusing on publicly available information. The majority of the information was drawn from official education ministry and assessment agency websites. Where information could not be found on these sites (particularly information regarding public perceptions), news websites were included in the review. Other sources referred to in this report include published research reports and journal articles.

Caveats

There are limitations to this approach and the caveats below should be considered when reading this report.

- The findings are not definitive and have not been validated by the jurisdictions involved.
- In the majority of cases, the sources publish limited detail on the methodologies of the approach taken and no detail on the rationale of selecting the approach or the impact it has had.
- It was not always clear whether the detail available was current.

We are publishing this working paper at this time and sharing the information more widely, for example through a paper at the 41st annual conference of the International Association for Educational Assessment in Kansas in October 2015. We are seeking, and would very much welcome, feedback from in-country experts, those who are much closer to the operation of the individual assessments referenced in this working paper. That should allow us in due course to publish a fuller paper in which we can have more confidence.

3. Inter-subject comparability methods

This section is about the various methods that can be and are used to investigate and address inter-subject comparability.

Different methods are available to make adjustments to the comparability of standards between subjects. They can be broadly categorised into two groups:

- judgemental methods, which rely on the analysis and judgement of subject experts to assess the subject or assessment demand of assessment materials;
- statistical methods, which employ a range of modelling techniques to assess and adjust for subject difficulty.

These methods are each described and evaluated in section 4 of *Inter-Subject Comparability: A Review of the Technical Literature: ISC Working Paper 2* (Ofqual, 2015b).

The majority of the assessments reviewed in this report do not appear to make such adjustments. This may be because the jurisdictions judge any benefits to be outweighed by wider implications. It may be because they do not accept that there is any problem with the comparability of their assessments. Gaining an understanding of why decisions have been taken is difficult when using only publicly available information.

Where they do make adjustments, systems appear to be most likely to apply statistical methods to place students taking a range of subjects onto a common scale. Jurisdictions that do not appear to apply statistical methods may address inter-subject comparability in other ways, for example as part of the assessment design phase, although this was not evident from the public sources we reviewed. There was evidence in many of the jurisdictions we reviewed that subjects were given different weightings within assessments. Typically though, this was based on such factors as the number of hours of teaching rather than being related to judgements about relative subject difficulty.

Statistical models make adjustments to the exam results of different subjects based on the performance of the same cohort of students in those subjects, that is the relative difficulty of the subject is assessed based on the collective performance of students in that subject and other common subjects.

Subject pairs analysis calculates the difference in the grade achieved by students who have taken the same two subjects. The mean of the differences across all the students in the analysis is the measure of difficulty in grade units for those two subjects. This process is repeated for all pairs of subjects until a list of relative subject difficulties is generated.

Kelly's method compares a student's grade in one subject with his or her average grade in all the other subjects he or she has taken to estimate the difficulty. This is repeated for all subjects and the difficulty estimates are used to apply a correction factor. The process is repeated with the difficulty corrected, and the process continues to be repeated until the corrections shrink to zero and the estimates of difficulty for each subject converge.

Average marks scaling is based on the notion that when the same group of students takes a set of the same subjects, then the average performance of the group on each subject should be roughly the same. The results of each group of students in every possible pair of subjects are compared. Scaling then adjusts the raw scores in all subjects so that the scaled scores in the different subjects will be comparable.

Other scaling methods include Z-scores and T-scores, which position a particular mark in relation to the mean mark of all the students who have taken that assessment, measured in standard deviations. T-scores have a mean of 50 and a standard deviation of 10. Z-scores have a mean of zero and a standard deviation of one. The scores simply describe the location of a mark within a distribution.

Percentile analysis is used as the basis of some inter-subject comparability methods. The group ability index used in the Hong Kong Diploma of Secondary Education is the method for aligning outcomes in elective subjects. It can be thought of as a set of suggested percentages for each reporting level. It is calculated for an elective subject using the candidature's results in the four core subjects and the correlations between the subjects.

Item response theory and the Rasch model are examples of latent trait models. Item response theory is a probabilistic model that predicts outcomes based on the difficulty of items and the abilities of students. Similarly, the Rasch model measures the difficulty of items and the abilities of students on the same scale, with the student's probability of success on a particular item determined by the difference between these two measures, related by the logit function (the difference being equal to the log of the odds). When item response theory or Rasch is used in the context of subject comparability, each subject is treated as an item. That allows the difficulty value of each subject to be compared directly to the difficulty values of the other subjects.

A reference test approach uses a common test, usually a general ability test, against which performance in different subjects can be compared. A regression model is often used, with results in the different subjects being regressed against the reference test scores. This technique has the advantage that students do not have to take the same subject exams in order for those subjects to be compared.

4. Assessments where there is evidence that inter-subject comparability is addressed

Methods to address inter-subject comparability were implemented in some of the jurisdictions we reviewed. Statistical modelling techniques are commonly applied in these jurisdictions, taking into account the relative difficulty of subjects when assessing the results of each student. In most cases, this is in order to support access to higher education, though in the Singapore and Taiwan examples it is to support access to selective options within secondary education. A summary of the findings is provided in table 1 below.

Table 1: Overview of assessments that address inter-subject comparability.

Jurisdiction	Assessment	No. subjects	Assessed subjects	Description	Method
Cyprus	Pan Cyprian Exam	4	Free choice	Z-score scaling is applied to the results to rank students regardless of subject choice.	Z-score scaling
Fiji	Fiji School Leaving Certificate	4	Free choice	Average marks scaling is applied to the results to rank all students regardless of subject choice.	Average marks scaling
Hong Kong	Hong Kong Diploma of Secondary Education	4	Restricted framework	Standards-referenced reporting and a group ability index are conducted with subject choice within a restricted framework.	Group ability index/ Rasch model
New South Wales, Australia	New South Wales Higher School Certificate	Varies	Free choice	Standards-referenced assessment that uses average marks	Average marks scaling

				scaling. The Australian Tertiary Admission Rank is scaled.	
Scotland, UK	Scotland Standard Grade, Intermediate 1 and 2, Higher and Advanced Higher	Varies	Free choice	Kelly's method is applied to produce a national rating for each subject.	Kelly's method
Singapore	PSLE	4	Uniform subjects	T-score scaling is applied to rank students in order of attainment.	T-score
Taiwan	The Basic Competency Test	5	Uniform subjects	Item response theory is applied to rank students by attainment.	Item response theory
Tasmania, Australia	Tasmanian Certificate of Education	4	Uniform subjects	Rasch analysis is used to produce scaled scores using the relative difficulty of each subject.	Rasch model

From the 30 assessments we reviewed in total, eight jurisdictions explicitly address inter-subject comparability. Other jurisdictions may address inter-subject comparability, although that was not clear from the information we reviewed.

The assessment structure varied for each of the eight jurisdictions we identified as explicitly addressing inter-subject comparability.

- One of the eight jurisdictions limited the subject choice within a restricted framework (Hong Kong).
- Three of the jurisdictions had uniform subject choices for assessments (Singapore, Taiwan and Tasmania), including the two jurisdictions where the

assessment reviewed was to access secondary education (Singapore and Taiwan).

- Four of the jurisdictions allowed free subject choice (Cyprus, Fiji, New South Wales and Scotland).
- Five of the jurisdictions examined four subjects (Cyprus, Fiji, Hong Kong, Singapore and Tasmania); one included five subjects (Taiwan); and two of the jurisdictions varied the number of subjects based on student choices (New South Wales and Scotland).

The following statistical methods have been used to address inter-subject comparability in assessments identified in the review:

- Latent trait models:
  - Rasch model: Tasmania
  - Item response theory: Taiwan.
- Common examinee linear models:
  - Kelly's method: Scotland
  - Average marks scaling: New South Wales and Fiji
  - Scaling using T-scores: Singapore
  - Scaling using Z-scores: Cyprus
  - Percentile analysis (group ability index): Hong Kong.

The Hong Kong Diploma of Secondary Education exam uses multiple methods that are designed to address inter-subject comparability and maintain standards over time. Subjects are categorised into three groups: core, elective and applied. In setting standards, judgemental methods, through the inspection of scripts and reference to level descriptors, and statistical methods are applied. Different statistical methods are applied to different subject groups to produce a set of recommended cut scores. To address inter-subject comparability in specific elective subjects and to assist in grading applied subjects, a group ability index is calculated for each level, based on the candidature's results in the four core subjects and the correlations between the subjects. Results are then adjusted accordingly. The four core subjects are also monitored annually with a representative group of selected schools. A latent trait model is applied to the monitoring test data and live exam data, to standardise all items in the different exams and generate the suggested cut scores.

In Tasmania, the Rasch model is used to scale subject scores in the Tasmanian Certificate of Education in order to generate a tertiary entrance score for each student to enable him or her to access higher education. In order to make comparisons between subjects, Tasmanian authorities assume that all subjects are underpinned by a common construct of 'general academic ability' or 'merit to enter university'. Rasch analysis of whole subject assessments, rather than items, is undertaken for every subject, and each subject is equated onto a common scale at three award points (satisfactory achievement, high achievement and outstanding achievement). The model takes into account all the subjects undertaken by the students, and the award threshold positions are adjusted on the scale according to the relative difficulty of the subjects. Once the analysis is complete and the scaled thresholds for each subject have been finalised, the scores in between the threshold positions are filled in and a combined score on the common scale is produced for each student. This ensures that the scaled subject scores are directly comparable.

In New South Wales, average marks scaling is applied to Higher School Certificate results and adjustments made to generate an Australian Tertiary Admission Rank score for each student to enable him or her to access higher education. English is the only compulsory subject, and students can choose from over 100 courses to complete their Higher School Certificate. The scaling approach is based on the principle that when a common candidature takes two or more of the same subjects, then the average performance of the group should be roughly the same. The results of each group of students (common candidature) in every possible pair of subjects are compared and the raw scores are then scaled so that the results in different subjects are adjusted to take into account the difficulty of the subject. This combined score forms the Australian Tertiary Admission Rank.

In Scotland, the Scottish Qualifications Authority annually produces national ratings based on a similar approach to that used in New South Wales. It employs Kelly's method to compare grades achieved by students in one subject with how the students performed in all other subjects to estimate the difficulty of that subject – the national rating. The national ratings indicate how many grades higher or lower the student group achieved in a subject than they achieved on average in their other subjects. Although no longer published, the ratings are still considered during the development of assessments and are discussed at the meetings where grade thresholds are determined.

In Cyprus, a Z-score scaling method is applied to convert the raw scores of subjects in pre-university exams. A standard deviation of 3 and a mean of 10 are applied to rescale all the scores, and an aggregate score is calculated for selection purposes.

Similarly, in Fiji, average marks scaling is used, and standardised scores are reported for subjects. Raw scores are converted onto a scale with a mean of 50 and

a standard deviation of 17. The mean and standard deviation are set centrally by the Ministry of Education based on the performance of previous cohorts.

The Basic Competency Test, the exam taken at the end of primary school in Taiwan, uses item response theory models to convert raw scores in each subject to scaled scores. The multiple-choice tests, taken over two days, comprise six subjects: Chinese, English, mathematics, science, social science, and writing, and each has a scaled score ranging from 1 to 60 points. There are two opportunities each year to sit the test, with students given a reported score out of 300 and a percentage ranking (1 to 99).

In Singapore, a scaling method, which ranks all students according to their performance, is applied to the PSLE subjects. Students' proficiency in English language, one language selected from a prescribed range of 'mother tongue' languages (Chinese, Malay and Tamil), mathematics and science is nationally examined. Generally, students are able to take subjects at either foundation or higher level. In each mother tongue language there are three levels of exams: standard, foundational and 'Higher Mother Tongue'. Because of the varying raw marks between assessments, scores are converted to T-scores, where the mean is 50 and the standard deviation is 10 in each subject. An aggregate score is then produced to assist in secondary school selection.

4.1 Public perceptions of addressing inter-subject comparability through statistical methods

It is important that the conduct of an exam system is perceived as fair and acceptable to the public in order to be trusted and promote confidence in the results. Despite the use of statistical methods by some of the jurisdictions we reviewed, they still experience critical comments from public, professional and academic sources. These highlight a risk that using statistical methods which are unintelligible to most audiences to align subjects can result in mistrust and a lack of confidence in the system.

Scaling of scores can also influence student choice. Students and teachers can try to devise methods to 'work' the system. For example, they might identify 'easy' subjects – those that are expected to be scaled down. They could then devise specific (favourable) combinations of subjects in an attempt to avoid their results being scaled down.

Some examples of these issues are given below.

There is widespread concern in Australia about the reduction in the number of students studying calculus-based mathematics courses in the final year of secondary education. Recent research indicates that, on average, those in New South Wales

who study general mathematics as part of their Higher School Certificate achieve materially higher scaled scores than those who undertake the calculus-based course.

> The current scaling mechanism provides a strong incentive to take HSC general mathematics for a very large group of students. At a time when many are deeply concerned about the reducing numbers of students studying higher level mathematics in the final year of secondary education, it is useful to consider the evidence presented here which supports one of the possible explanations for this drop in numbers. (Pitt, 2015, p. 80)

Of more than 1,000 mathematics teachers surveyed, half believed that some students in their school were selecting senior mathematics courses below their capability. A desire to optimise Higher School Certificate and Australian Tertiary Admission Rank results was the most common reason given for these selections, and it was cited over 200 times by these teachers (Pitt, 2015).

A newspaper article (Sydney Morning Herald, 2015) based on the research generated many online comments, some of which indicated that the implications of scaling go much wider than mathematics. For example, “I went to an HSC information night.. . . Parents and students were very concerned about scaling both within and between subjects. Some students spoke of the difficulty of deciding whether to do one level of a subject or another, trying to take into account their own ability and the perceived scaling of that level of the subject. . . every part mark counts when you are trying to enter courses.”

In 2013, in his speech at the National Day Rally, the Prime Minister of Singapore spoke about the T-score system used in the PSLE:

> The PSLE, everybody thinks it matters, heaven and earth. I do not know what my PSLE grade is. . . But today, it is different. . . Not just everybody knows his T-score, everybody knows his friends’ T-score and his friends’ sons or daughters’ T-score. . . One-point difference in the PSLE scores, 230 versus 231, may make all the difference in your secondary school posting. But at the age of 12, one examination, four papers and you want to measure the child to so many decimal points and say well, this one got one point better than that child? It is a distinction which is meaningless and too fine to make. Who is going to grow up abler, more committed, more capable, a better contributor to society? At the age of 12, you can guess, you cannot tell. Certainly, you cannot tell based on one point difference and I do not think we should decide secondary school postings based on such fine distinctions.
>
> So we will score PSLE differently. We will use wider bands for grades, ‘O’ levels are like that. . . A1 to 9. . . I think if we have a system of grades like

> that rather than precise scores, it will reduce the excessive competition to chase that last point. If you get an A* that is an A*, it does not matter where it is 91 A* or 99 A*. It is an A* and that is good enough. (Prime Minister's Office, 2013)

Although the issue raised in the speech is not specifically about inter-subject comparability, the purpose of the scaling to produce the T-score is to align subject scores so that they can be aggregated.

In Cyprus, since 2006, the raw results of the upper secondary school graduation exam (the Pan Cyprian Exam) in 'easy' subjects have been scaled down and those in 'difficult' subjects scaled up to provide comparable access scores for university entrance purposes. One consequence is that students try to avoid subjects that historically have been scaled down. For example, entries for chemistry dropped by 70 per cent following the introduction of the scaling system (Lamprianou, 2007).

The media became interested because, to the public, some of the stories about particular students' scores appeared inexplicable. Some students with the same raw scores received very different scaled scores because they had taken different subjects. There were stories about students who had taken the same subjects and had the same average raw score but different raw scores per subject. Depending on the statistical difficulty of each subject, some of these students then ended up with very different scaled scores.

Parents and students questioned the fairness of the system and even the accuracy of the calculations. Although the purpose of scaling is to adjust the raw scores because of the differential difficulty of the exam subjects, the end result puzzled parents, students and the press (Lamprianou, 2012).

In Fiji, there have been recurring reports of distrust in the scaling of marks. In 2006, students and parents complained to the Fiji Human Rights Commission that the scaling of exam marks was unfair and not transparent. In 2008, the Fijian Teachers Association requested a review of the scaling policy as it felt it was confusing for students and seemed to scale down able students whilst less able students were being scaled up.

In 2010, the Fijian Teachers Association stated that it did not support the scaling of exam marks and felt it was a government exercise rather than in the best interests of the students. However, the Fiji Principals Association was in favour of scaling, believing that it put all students on a level field and enabled comparability. The Ministry of Education defended the system, saying that it was based on sound educational assessment principles and was used internationally.

In the Fijian Parliament, in February 2015, the education minister outlined his case for the removal of scaling, saying the practice has, "caused substantial damage to the education system and graduates in the market." Raw mark evidence from 2009 to 2014 showed that mean marks in Years 12 and 13 had steadily declined. "Madam speaker, there was no other alternative but to remove or discontinue the scaling of marks. . ." (Fiji Times, 2015).

Prior to the introduction of the NCEA a decade ago, the New Zealand system adjusted students' results to attempt to improve inter-subject comparability using percentile analysis. Subjects' standard scores were adjusted so that the performances of their groups of students were comparable to that of the groups in their other subjects. The inter-subject scaling of marks was a percentile analysis process based on the 95th, 90th, 75th, 50th, 25th, 10th and 5th percentiles and was applied to the national distribution of the marks for a subject based on students who had entered three or more subjects. However, this approach also meant that only a certain number of students could pass the exam, and a fixed number of students would receive a fail grade. The system was felt to be unfair on students, as their success was relative to the performance of others, and it was feared that the focus on inter-subject comparability masked overall changes in student performance over time.

These concerns led to the introduction of the NCEA, which is standards-related and credit-based. It allows students the flexibility to choose the subjects they want to study to gain credits towards their final certificate. When the NCEA was first implemented, the proportion of results awarded at each achievement level (Achievement, Merit and Excellence) varied from standard to standard and within a particular subject. Such variation was not considered problematic by central authorities – it was simply accepted that some standards were harder to achieve than others. However, schools, teachers and parents were concerned with the variability. In addition, students appeared to be adopting strategic approaches to collecting credits, on the basis of those that were easier to obtain and those which would allow demonstration of higher levels of achievement. To improve inter-subject comparability, standards were more tightly defined, where necessary, and levels of achievement were closely monitored (Jones, Phillips and van Krieken, 2005).

4.2 Summary

The jurisdictions we identified utilised a range of statistical approaches in attempting to address inter-subject comparability. The jurisdictions are diverse in the composition of their assessment systems, with variance in structure, number of subjects, exam approaches and marking. However, whilst acknowledging the individual nature of each system and the caveats outlined in the methodologies section about the limitations of this review, some observations can be made.

From the eight jurisdictions we identified as implementing attempts to address inter-subject comparability, all used statistical methods. This may be because of the necessity of publishing the calculation and approach to ensure transparency and confidence in the education system. Other jurisdictions may attempt to address inter-subject comparability, for example during assessment design through the use of judgements, but we have not been able to find such information through publicly available websites.

In summary:

- The purpose of taking six of the assessments we reviewed was to access university; the purpose of taking the remaining two assessments was to access selective secondary education options.
- Of those where the purpose of taking the assessment was to access university:
    - three directly applied the inter-subject comparability corrections to student results (Cyprus, Fiji and Hong Kong);
    - two applied the inter-subject comparability corrections to generate a separate national university entrance score, supplying students with both an assessment result (unscaled) and an entrance score (scaled) (New South Wales and Tasmania);
    - one applied inter-subject comparability scaling to generate ratings for internal use by the exam board (Scotland).
- The majority of the jurisdictions applying inter-subject comparability approaches shared a similar structure, with most students studying four subjects from a free-choice menu.

5. Assessments with limited or no evidence that inter-subject comparability is addressed

There was little or no evidence that the other jurisdictions we reviewed implemented approaches to improve inter-subject comparability. However, because of the limitations of this review, it may be that the jurisdictions do attempt to improve inter-subject comparability but that the information is not publicly available. There is some evidence that jurisdictions which apply weightings to particular subjects are using judgements to decide these weightings. However, it is unclear from the evidence available whether the weightings are due to subject difficulty, demand or a variety of other factors (such as teaching hours). In all cases, the assessments we reviewed were used at the end of upper secondary school to gain access to university.

From the 30 assessments we reviewed, 22 assessments did not appear to address inter-subject comparability, based on the evidence available. The findings are summarised in table 2 below.

Table 2: Overview of assessments that do not address inter-subject comparability.

Jurisdiction	Assessment	No. subjects	Assessed Subjects	Description
Alberta, Canada	High School Diploma	Varies	Free choice	Students must achieve 100 credits made up of six mandatory and some elective subjects. Most courses are each five credits.
Brazil	ENEM	5	Uniform subjects	Increasingly used in Brazil to gain university entrance. Comprises 180 multiple-choice questions in five main areas: natural sciences, mathematics, human sciences, Portuguese and a foreign language. Candidates are also required to write an essay. The exam is scored out of 1,000 points.

China	*Gāokǎo* (National Higher Education Entrance Examination)	4	Restricted framework	In most provinces, students take Chinese, mathematics and a foreign language (generally English) and either the humanities suite or the science suite. The mandatory and elective subjects are given different predefined points values. Total score out of 750.
Finland	Matriculation Examination	4	Restricted framework	Subject choice is limited within a framework. Each subject is graded from 1 to 7.
France	*Baccalauréat général*	6	Restricted framework	Students select one of three series, within which subject choices are weighted differently depending on the series selected. Each subject is marked out of 20 with 10 being the minimum pass.
Germany	*Abitur*	10	Restricted framework	Subjects are divided into three areas, which are single, double or triple weighted in the final score. The *Abitur* uses a 15-point grading scale using numbers.
Ghana, Liberia, Nigeria, Sierra Leone, The Gambia	West African Senior School Certificate Examination	8 to 9	Restricted framework	Multiple-choice plus essays. Subject choice is limited within a framework. There is a nine-point grading system from A1 (excellent) through C4 to C6 (credit/minimum

				acceptable pass) to F9 (fail).
Greece	Pan-Hellenic Exam	16	Restricted framework	Students select from one of four predefined pathways. Each subject is marked out of 20, with 10 being a pass.
International Baccalaureate	Diploma programme	6	Restricted framework	Students select subjects from six subject groups, which can be taken at standard or higher level. Subject grades range from 1 to 7. A student's final score is made up of the combined scores for each subject. The diploma is awarded to students who gain at least 24 points.
Ireland	Leaving Certificate	~7	Free choice	With the exception of Irish, students are able to choose which subjects they study, although English and mathematics are effectively compulsory, and the majority of students take a third language. There are 13 grades, from A1 to F.
Israel	Psychometric Entrance Test	3	Uniform subjects	This test covers three areas: quantitative reasoning, verbal reasoning and the English language. One writing task plus 124 multiple-choice questions. The scoring scale ranges from 200 to 800 points.

Japan	National Centre Test	5	Uniform subjects	There are a total of 29 multiple-choice tests in six subjects. Students take the subjects specified by their university. Most subjects are scored out of 100 points.
Kazakhstan	Unified National Test	5	Restricted framework	One hundred and twenty five multiple-choice questions. The exam covers five subjects: Kazakh language, Russian language, mathematics, Kazakh history and an option – normally biology, physics or geography. The test is scored from 0 to 100; this is then converted to a grade of 2 to 5.
Netherlands	VWO	9	Restricted framework	Students select from one of four predefined pathways, in addition to mandatory general education subjects. Each subject is graded from 1 to 10, with an average final grade of 6 being the lowest pass.
New Zealand	NCEA	Varies	Free choice	When students achieve the standards in a subject, they gain credits, and once they have enough credits they get an NCEA certificate. A single achievement standard generally attracts 3 to 4 credits, and a single subject usually has 5 to

				8 such standards. Usually, 18 to 25 credits are needed per course of study. Students do not get an overall grade for a subject.
Poland	*Matura* (High School Examination)	3	Restricted framework	Subject choice within a restricted framework. Percentage and percentile results are reported to compare results on a national scale.
Russia	Unified State Examination	4+	Restricted framework	Russian and mathematics are compulsory. Optional tests in foreign languages, physics, chemistry, biology, geography, literature, history, social sciences and computing science. Multiple-choice plus written answers required.
South Africa	National Senior Certificate	7+	Restricted framework	Seven subjects, including two compulsory official South African languages, either mathematics or mathematical literacy, life orientation and three elective subjects. Grading of subjects is on a seven-point rating scale, where 4 is the minimum acceptable pass.
Thailand	GAT and PAT		Restricted framework	The compulsory GAT covers reading, writing, analytical thinking,

				problem solving and English communication. The PAT has a choice of seven subjects – mathematics, science, engineering, architecture, education, arts and languages. The GAT and each PAT is scored out of 300 points.
Switzerland	Federal Maturity Certificate	9	Restricted framework	Every student studies mandatory subjects with an elective subject of focus and a supplementary subject. Subjects are weighted according to teaching hours. Each subject is graded, 6 being the maximum grade.
USA	SAT I	Varies	Uniform subjects	There are two versions of the SAT test. SAT I tests measure general verbal and quantitative reasoning. They comprise three sections: writing, reading and mathematics, the last two of which are tested primarily by multiple-choice. Possible scores range from 600 to 2,400.
USA	SAT II	Varies	Free choice	SAT II tests – which far fewer students take – are subject-based. Students typically take three subjects, chosen from the 20 available.

				Each test is scored from 200 to 800.
USA	ACT	Varies	Uniform subjects	The subject-based ACT consists of four multiple-choice tests: English, mathematics, reading and science (each scored from 1 to 36), with an optional writing section (scored from 1 to 12).

The assessment structure varies for each of the jurisdictions:

- Most jurisdictions limit subject choice within a restricted framework (examples include Finland, the International Baccalaureate and the West African Senior School Certificate Examination).
- Other jurisdictions have uniform subject choices (including Japan and Brazil).
- A minority of jurisdictions allows free subject choice (including Alberta, New Zealand and Ireland).

The following assessment system structures were evident:

- Subject choice limited within a restricted framework:
    - units/subjects arranged in prescribed subject groups from which students must select a subject per group (for example, Finland, International Baccalaureate and Switzerland);
    - pathway approach that predefines subject combinations (for example, Greece and the Netherlands);
    - application of predefined weightings to different subject areas (for example, France and Germany).
- Uniform subject choices for exams:
    - set national/state exam that all students take (for example, Japan).
- Free subject choice:
    - credit-based system that assigns a prescribed number of credits to each subject (Alberta and New Zealand).

Jurisdictions where no evidence could be found as to whether inter-subject comparability approaches were implemented did not explicitly state why they did not address the issue. In this regard, we could assume, for example, that where subject choice is offered within a restricted framework, subjects are categorised and selection is controlled so that the design of the education system means that students take similar combinations of subjects, making attainment broadly comparable – but it would be just an assumption.

In education systems with restricted frameworks there is the necessity for students to identify their preferred university course and choose the appropriate pathway early on. This ensures that all students applying for specific university courses will have a very similar assessment profile and will, therefore, be comparable within their field. Similarly, those jurisdictions that allow students to select from groups of subjects assume equivalency within or between groups and, therefore, there is an overall balance in the assessment profile of the students. In most cases, there is the opportunity for students to take subjects at different levels or at different weightings within the restricted framework, which differentiates between them. The selection of subjects from a restricted framework can ensure breadth through the necessity of studying subjects from disparate areas and depth by focusing/weighting particular subjects. It is again necessary for students to be aware of the subjects they need to study, and at which level/weighting, for their desired university course.

Education systems that use a credit-based system provide the flexibility for students to gain more credits by selecting subjects which are perceived to be more rigorous and challenging. So do those that apply weightings to subjects to inform the final grade. The process of defining the subject credits or weightings varies by country.

In New Zealand, the credits available per subject for the NCEA are based on curriculum standards that have been defined by subject experts.

In France, the subject weightings for the *Baccalauréat général* depend on the importance of the subject to the pathway, the depth of the syllabus and the teaching hours. Science subjects receive a higher weighting if you select the scientific pathway and a lower weighting if you select the literary pathway, and vice versa for literary subjects. For example, philosophy is a key subject for the literary pathway and, therefore, it has a wide-ranging syllabus, receives eight teaching hours per week and has a high weighting. In the scientific pathway, philosophy is not a key subject and is subsequently taught for 2 to 3 hours per week, covers a limited syllabus and receives a low weighting. This allows students to select the pathway that best suits their strengths and interests.

The International Baccalaureate allows students to study subjects at either higher or standard level, with the higher level options having a more considerable syllabus and increased teaching hours. In Switzerland, weightings are related to teaching hours. In

Germany, the weighting is dependent on the point when assessment is taken, with core subjects double weighted throughout the duration, other subjects single weighted, and final exam subjects triple weighted.

One of the outcomes of inter-subject comparability approaches in the jurisdictions we reviewed was to enable students to be placed on a common scale for selection purposes. This was also found to be evident in those jurisdictions that did not appear to apply methods of inter-subject comparability. Finland and Poland use norm-referencing to assign grades in school leaving exams. Norm-referenced methods aid stakeholders in selecting the highest attaining students for higher education and employment opportunities.

5.1 Public perception of not addressing inter-subject comparability

Very little information could be found regarding public perceptions of inter-subject comparability in jurisdictions where there was no evidence of a statistical inter-subject comparability method being implemented. There was some evidence of debate around subject difficulty and which subjects were perceived to be 'easier', but this was limited. Concern was evident, more generally, around the format or management of exams rather than the comparability of subjects, and some jurisdictions were undergoing reform of their assessments as a result of these concerns. Two examples are worthy of mention though.

The Leaving Certificate marks the end of upper secondary education in Ireland. It is taken by more than 90 per cent of the age cohort. Although designed as a terminal exam for certification, in practice this purpose is overshadowed by the certificate's central role in selection decisions for higher education institutions. In the case of the great majority of applicants to most higher education institution courses in Ireland, it is the sole criterion used in the selection decision. A discussion paper (Hyland, 2011) listed ten key concerns about the Leaving Certificate raised by various stakeholders, within and outside the education system. These concerns included:

- It is easier to get a high grade in some Leaving Certificate subjects than in others. Some students choose subjects because it is easier to get a high grade in them, rather than because of their relevance for the third-level course for which the students are applying.
- Some students choose their course on the basis of their likely points rather than on their interest in the course – they don't want to "waste their points".

Recently announced reforms to the Leaving Certificate[2] do not include any proposals to align subjects.

In China's *Gāokǎo* a maximum of 750 points are available from the exams. Chinese, mathematics and a foreign language are worth up to 150 points each, and there are a further 100 points for each subject (up to three subjects) in the humanities and science combinations. The overall mark received by students is generally a weighted sum of their subject marks. The marks in the separate subjects are raw marks.

The Chinese government has announced that changes will be implemented in 2016. The weighting of English is to reduce from 150 to 100 points, and the weighting of Chinese is being increased from 150 to 180 points. The reason for these changes has not been explicitly stated by the Ministry of Education, but commentators suggest it is likely to be for two reasons. First, to reduce the disadvantage faced by students from low-income backgrounds or rural settings who are likely to have less access to English compared with their higher income, city-based peers. Second, the reduced weighting of English could be to favour other subjects like mathematics (150 points). In both cases, this alteration in weighting is due to factors of subject equality and importance rather than subject difficulty or demand (Sinograduate, 2014).

5.2 Summary

Twenty-two assessments were reviewed where no evidence could be found as to whether the jurisdiction was implementing an approach to address inter-subject comparability. It may be that inter-subject comparability is addressed but that the information is not publicly available.

With that caveat in mind, the following observations can be made regarding the context, purpose and subject choice of each jurisdiction:

- The purpose of the assessments we reviewed was to provide access to university.
- Several of the jurisdictions apply varying weightings or credits to specific subjects within the assessment by:
  - applying subject weightings (including France and China);
  - applying credits to subjects (Alberta and New Zealand);

[2] www.transition.ie/files/Supporting%20a%20Better%20Transition%20from%20Second%20Level%20to%20Higher%20Education%20-%20Implementation%20and%20Next%20Steps_April%202015.pdf

- providing options to take subjects at either standard or higher level (including Ireland and Poland).

- Assessments vary in structure, although the majority share a similar structure in that students select subjects from a restricted framework.
- The number of subjects varies from four up to 16.
- In jurisdictions where statistical inter-subject comparability methods are implemented, students are ranked to enable higher education institutions to identify easily the highest performing candidates. In jurisdictions where inter-subject comparability methods are not evident, to compensate for this, some higher education institutions introduce additional requirements for entry such as interviews, selection tests and the submission of essays and portfolios.
- There appears to be less public reaction to issues of inter-subject comparability in jurisdictions that do not apply particular methods. This may be because the parameters for success are clear to students at the beginning of their courses rather than adjusted after their exams. It may be that where no statistical methods are applied, it is less obvious to the public that there may be a problem that ought to be solved. It may even be that statistical methods are introduced in response to a concern about subject difficulty but succeed in redirecting the concerns to the method employed.

6. Inter-subject comparability messages

We should reflect on the following messages when considering what this review might tell us about how we might address inter-subject comparability in GCSEs and A levels in England.

The purpose of the assessments we reviewed here was primarily to enable students to access higher education. In all the jurisdictions, it was important for students to select the subjects necessary to access the university course they wished to study. The same is true of England.

Systems that use inter-subject comparability statistical methods generally award 'certificate' style products, where it is necessary to have both individual subject awards as well as an overall award. However, this approach is not exclusive to jurisdictions using inter-subject comparability statistical adjustments. Many that do not use inter-subject comparability statistical methods also award certificate-style products.

Those that use statistical methods are also more likely to operate a free-choice structure, whilst those that appear not to use these methods are more likely to operate a restricted framework approach, although both structures are evident in each category. Systems that use statistical methods most commonly include four subjects, whilst those that appear not to use them most commonly include a higher number of subjects, ranging up to 16.

Statistically adjusting results to address inter-subject comparability has the risk that it can lead to perceptions of unfairness, perhaps partly due to the complexity of the calculations used, which makes them unintelligible to most audiences. This is evident from some public perceptions of the systems in Australia, Cyprus and Fiji in particular.

As reported in *Inter-Subject Comparability: A Review of the Technical Literature: ISC Working Paper 2* (Ofqual, 2015b, p. 36), over the years similar rank orders of subjects have been found around the world in research into subject difficulty – in England, Scotland, South Africa, New Zealand and Australia, for example. Typically, the difficult subjects are said to be the sciences, mathematics and languages. In other jurisdictions there can be similar perceptions, for example in France it is perceived that the scientific pathway is the most challenging, with the economics and social sciences pathway viewed as the least challenging.

That raises a question about whether any statistical adjustments are really appropriate, as perhaps the calculated differences in subject difficulty are 'right'. The variety of international practices described in this working paper certainly suggests there is no widely agreed best route that GCSEs and A levels might follow.

References

Fiji Times (2015) *Scaling concern*. Available at: www.fijitimes.com/story.aspx?id=295649 (accessed 17th April 2015).

Hyland, A. (2011) *Entry to Higher Education in Ireland in the 21st Century.* Available at: www.transition.ie/files/Entry_to_Higher_Education_in_Ireland_in_the_21st_Century%20.pdf (accessed 20th March 2015).

Johnston, Dr. L. (2014) *Why is China reducing the importance of English among Gaokao exams?* Sinograduate. Available at: www.sinograduate.com/comment/articles/why-china-reducing-importance-english-among-gaokao-exams (accessed 10th April 2014).

Jones, B.E., Philips, D. and van Krieken, R. (2005) *INTER-SUBJECT STANDARDS: AN INSOLUBLE PROBLEM?* Manchester, Assessment and Qualifications Alliance.

Lamprianou, I. (2007) *Comparability methods and public distrust: An international perspective*. In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards*, pp. 368–371. London, the Qualifications and Curriculum Authority.

Lamprianou, I. (2012) *Unintended consequences of forced policy-making in high stakes examinations: the case of the Republic of Cyprus,* in *Principles, Policy & Practice Volume 19, Issue 1, 2012 Special Issue: High-stakes testing - value, fairness and consequences.* Assessment in Education.

Ofqual (2015b) *Inter-Subject Comparability: A Review of the Technical Literature: ISC Working Paper 2.* Coventry, the Office of Qualifications and Examinations Regulation.

Pitt, D.G.W. (2015) *On the scaling of NSW HSC marks in mathematics and encouraging higher participation in calculus-based courses*. Australian Journal of Education 2015, Vol. 59(1) pp. 65 – 81.

Prime Minister's Office (2013) Prime Minister Lee Hsien Loong's National Day Rally 2013. Available at: www.pmo.gov.sg/mediacentre/prime-minister-lee-hsien-loongs-national-day-rally-2013-speech-english (accessed 17th April 2015).

Sydney Morning Herald (2015) *HSC maths: students studying advanced maths stung with lower marks in ATAR*. Available at: www.smh.com.au/national/education/hsc-maths-students-studying-advanced-maths-stung-with-lower-marks-in-atar-20150519-gh45ox.html (accessed 12th June 2015).

Providers of the assessment systems selected

Alberta, Canada: Alberta Education, www.education.alberta.ca

Brazil: National Institute of Educational Studies and Research, portal.inep.gov.br/web/enem

China: Ministry of Education of the People's Republic of China, www.moe.edu.cn

Cyprus: Ministry of Education and Culture, www.moec.gov.cy/ypexams/en

Fiji: Ministry of Education, www.education.gov.fj

Finland: Finnish National Board of Education, www.oph.fi/english/education/overview_of_the_education_system

France: Ministry of National Education, www.education.gouv.fr

Germany: Federal Ministry of Education and Research, www.bmbf.de/en

Ghana, Liberia, Nigeria, Sierra Leone and The Gambia: The West African Examinations Council, www.waecnigeria.org/Home.aspx

Greece: Ministry of Education and Religious Affairs, www.minedu.gov.gr

Hong Kong: Hong Kong Examinations and Assessment Authority, www.hkeaa.edu.hk/DocLibrary/HKCEE/Grading_and_Marking_SRR/booklet_srr.pdf

International Baccalaureate Organisation, www.ibo.org

Ireland: Department of Education and Skills, www.education.ie/en

Israel: National Institute for Testing & Evaluation, https://nite.org.il/index.php/en

Japan: Ministry of Education, Culture, Sports, Science and Technology (MEXT), www.mext.go.jp/english

Kazakhstan: National Testing Center, http://testcenter.kz/en/entrants/ent

The Netherlands: Ministry of Education, Culture and Science, www.government.nl; National Institute for Curriculum Guidance www.slo.nl

New South Wales, Australia: NSW Students Online, http://studentsonline.bos.nsw.edu.au/go/seniorstudy/hsc_rules_and_procedures

New Zealand: Ministry of Education, www.minedu.govt.nz; New Zealand Qualifications Authority, www.nzqa.govt.nz

Poland: Central Examination Board, http://apl-bud.home.pl/pdfs/edusystem_and_ext_asessment_Poland.pdf

Russia: Federal Service for Supervision in Education and Science, http://government.ru/en/department/35

Scotland, UK: Scottish Qualifications Authority, www.sqa.org.uk

Singapore: Ministry of Education, Singapore, www.moe.gov.sg

South Africa: Umalusi, www.umalusi.org.za

Switzerland: Swiss Conference of Cantonal Ministers of Education (EDK), www.edk.ch/dyn/11553.php

Taiwan: The Ministry of Education, Republic of China (Taiwan), www.wes.org/ewenr/10may/feature.htm

Tasmania, Australia: The Office of Tasmanian Assessment, Standards and Certification, www.tqa.tas.gov.au/1906

Thailand: The National Institute of Educational Testing Service, www.niets.or.th/upload-files/uploadfile/5/5113f2fc40d9b7ccbf26972226c1a536.pdf

USA: The College Board, https://sat.collegeboard.org/home; ACT, www.act.org/products/k-12-act-test

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346