Oxford Review of Education
Vol. 35, No. 2, April 2009, pp. 205–226

Routledge
Taylor & Francis Group

Comparability of examination standards between subjects: an international perspective

Iasonas Lamprianou*
University of Manchester, UK

Heated discussions about the comparability of standards between examination subjects have kept Qualification Authorities, Testing Services, independent researchers and academics around the world busy for many years. As a result, many countries have adopted statistical techniques which aspire to make aggregated scores based on different subjects comparable. This paper presents eight international case studies and highlights the criticisms against the statistical comparability methods. The side-effects (problems) they may have sparked in various countries are discussed in length. The author maintains the view that the Gordian knot in this case is not statistical: it is one of perceived unfairness. The Qualification Authorities and the Testing Services should step forward and explain to the stakeholders what can be done and what cannot be done using statistical techniques. Once the stakeholders accept the limitations of any statistical technique, it will be easier for them to strike a balance between the desire for comparability and the need to democratically allow students to choose their own combination of subjects to study. The psychometricians should not be forced to apply statistical solutions to complex problems caused by political decisions; they are more likely to fail than to succeed.

Introduction

Heated discussions about the comparability of standards between examination subjects have kept Qualification Authorities, Testing Services, independent researchers and academics around the world busy for many years. This is probably the result of the anxiety felt by citizens of these countries when they face issues of perceived unfairness regarding high stakes public examinations.

These examinations are usually established under the assumption of trust and support from society. In many countries the citizens are involved in a 'social contract' where examinations are generally accepted by society as the filter that controls the distribution of scarce educational resources (and as a result affects social mobility).

*Agrafon 22A, Strovolos 2027, Nicosia, Cyprus. Email: iasonas.lamprianou @manchester.ac.uk

ISSN 0305-4985 (print)/ISSN 1465-3915 (online)/09/020205–22

DOI: 10.1080/03054980802649360

In these countries, people's moral and/or political obligation to accept the output of high-stakes public examinations depends upon a perceived 'contract' or 'agreement' between the citizens to control entrance to tertiary education. It is therefore reasonable that the perceived fairness of high stakes public examination systems around the world frequently receives its fair amount of political and educational discussions and debates.

In high stakes examination systems (such as the A-Level GCE examinations in England or the university admission examinations in Australia), many countries allow students to choose their examination subjects in order to compete for a place in further education. For example, Rodeiro (2006) quotes a large list of A-Level subjects offered to students in England between 2001 and 2005. The subjects are split into a number of categories such as sciences/mathematics, social science/humanities, arts etc. Each category includes tens of examination subjects. Therefore a huge variety of subject combinations is available to all students to choose from (e.g. Greek to computing) in order to achieve the qualifications they need to enter university.

Bell *et al.* (2007) have recently provided a clearer account of the current situation in England regarding how students choose their examination subjects. There are groups of subjects that have not seen much change in their rate of uptake in recent years. However, there are indications that students are driven away from the (perceived) 'more difficult' subjects like mathematics and physics. Students probably consider the perceived 'difficulty' of those subjects as a barrier in their quest for access to tertiary education. Similarly, Cyprus and Australia have seen the uptake of other subjects such as chemistry falling dramatically, because students thought that chemistry reduced their aggregated examination result and thus reduced their chances of gaining access to further education (Lamprianou, 2007).

Students who compete for the same places in universities may end up with similar grades by taking different examination subjects. Where societies value pluralism and free choice in education highly, they usually allow their students to take different examination subjects, but there are also public concerns about the comparability of the aggregated scores. In many such cases statistical models are used in order to reassure all stakeholders that choosing between subjects does not affect a candidate's chances to proceed to further education.

It is interesting to observe the viewpoint of the Curriculum Council (2006) in New South Wales (Australia) when making its case for its own statistical model to aggregate student scores on different examination subjects:

> The scaling system is in place to ensure that students are free to make educationally sensible subject selections without endangering their chances of university admission. Students cannot get higher Scaled Scores by choosing subjects that are scaled up, or by choosing subjects that are perceived as easy. (p. 59)

The same argument has been frequently voiced by prominent educationalists, bureaucrats or politicians in other places. For example, the Singapore Examinations and Assessment Board (2006) explains that:

> Computing the Aggregate Score [using a specific scaling method] underscores the equal importance placed on pupil performance in each … subject.

This statement was made in a context where a candidate's mother had sent a letter complaining about—as she perceived it—the unequal weight of subjects of different difficulty for the computation of the Aggregate Score. Therefore, the argument of the Singapore Examinations and Assessment Board capitalized on issues of fairness to support their decision for their statistical model.

Needless to say, officials in Cyprus, the Fiji islands and other places where statistical methods are formally used to ensure the comparability of standards between examination subjects, also make very similar arguments to support their policies. As the late Minister of Education in Cyprus put it (Cyprus Minister of Education, 2006), the statistical processing (scaling) of the raw scores of the students on different subjects is a '*necessary, though undesirable*' means to ensure the fairness of the aggregated scores. According to the Cyprus Ministry of Education, aggregating the unadjusted raw scores on different subjects would only lead to worse side-effects and problems with the possible consequence of raising feelings of unfairness among parents, students and other stakeholders.

Still, the debate about the fairness or the effectiveness of using statistical methods for this purpose has gone on fiercely around the world, for many years. Heated debates (in many countries) in Parliaments and in the Press frequently cast doubts over the specific statistical methods employed. In some cases, debates cast doubt on the concept of statistical comparability between subjects, with the extreme case of the Fiji islands where the Human Rights Commission was reported in the summer of 2006 to be investigating alleged issues of human rights violations.

The need for an international review of the current (statistical) comparability methods became more obvious when the English Qualifications and Curriculum Authority (QCA) organised a conference to study this thorny problem. QCA, at the time, seemed to have at the back of its mind the identification of plausible comparability methods that might be used in the English (and probably the Welsh) context. The result was the publication (Newton *et al.*, 2007) of a book covering many aspects of the problem of the comparability of examination subjects, but with limited references to the international perspective.

England and Wales make a very interesting case study, in the sense that the students are free to choose democratically any combination of subjects to gain access to tertiary education, as long as the subjects are somewhat related to the intended course of study. However, no formal adjustment is made to accommodate for the perceived 'difficulty' of various examination subjects. This does not mean that the issue has not been raised frequently in the past: much discussion has been going on about issues of comparability between examination subjects. For example, Newton (1997) discussed the issue of comparability between examination subjects and mainly approached the issue from a statistical perspective. Newton focused on one of the possible statistical techniques, the Subject-Pair Analysis and demonstrated that reliance on the statistical comparison of standards between subjects was misplaced. His comments about statistical comparability probably echoed the concerns of many

other people at the time. However, allegations about 'easy' and 'difficult' examination subjects are still frequently published by the media in England.

The author of this paper strongly believes that discussing the statistical comparability of subjects around the world could help governments, researchers and other stakeholders worldwide to make up their minds on how to solve similar issues at home. The author does not take a specific view either towards or against statistical comparability methods. His view is that local societies are free to think for themselves, but he warns against any attempt to introduce statistical comparability methods in any education system without first reaching a consensus by all stakeholders (or at least a broad consensus indeed). However, the author believes that statistical models have only rarely solved complex educational problems. Finding simple and transparent political solutions to the issue might be the way forward.

Aims and methodology

This research presents a number of international case studies where the comparability of examination standards between subjects is pursued using statistical models. The technical details of the statistical models are not of interest in this study; the focus lies on the side-effects and the social unrest caused by the application of the statistical methods in order to achieve comparability of examination subjects.

Case studies 1–2: Two Australian methods: The New South Wales Average Marks Scaling method and the Tasmanian Qualifications Authority method (using Rasch models);
Case study 3: The Fiji scaling system;
Case studies 4–5: The Singapore T-score and the Cyprus z-score methods;
Case study 6: The New Zealand scaling system (recently abolished);
Case study 7: The Scottish Qualification Authority's National Ratings (not formally applied on students' reported scores);
Case study 8 (counter-example): The Pan-Hellenic University Entrance examinations in Greece.

Case study 8 is a discussion about the Greek examination system (as a counter-example to the previous case studies), where no between-subject comparability is pursued. This is compared with the 'statistical comparability' paradigm of the first seven examples.

This article does not focus on technical issues, and avoids elaborate presentations of equations and models that could scare away less statistics-savvy readers. The aim of the paper is to review some international efforts to achieve comparability between examination subjects and to elaborate on the problems that have been encountered. The author maintains the view that the problems of comparability between examination subjects are not merely technical (i.e. statistical) in the sense that all statistical models convey specific statistical, educational and philosophical assumptions. These assumptions probably reflect the culture and the idiosyncrasies of local societies, and their own acceptable trade-off between fairness and practical solutions to educational problems. In that sense, the problems that have been encountered internationally are

mainly problems that have to do with the fairness of the examination system as perceived by various countries and the desire for a practical solution that will allow the aggregation of scores for purposes of selection, sorting etc. Trying to strike this balance is not easy—thus the problems and the heated debates.

It is not desirable in this paper to elaborate on purely statistical or technical issues, because technical arguments and counter-arguments abound in other papers which will be referenced in later sections. It is possible that a purely technical discussion will obscure the issue and carry the reader away from the real problem which is to approach the issue from the perspective of the policy-maker and the stakeholder in general (e.g. parents, students, teachers).

In order to facilitate the flow of the story, each case study will be discussed briefly, citing only the main characteristics of the statistical method used with limited references to technical aspects. More space will be devoted to a general discussion, with references to the English context, below.

It is acknowledged that local perception of the comparability methods employed in various countries would be better investigated if questionnaires or interviews could be held locally, drawing on representative samples of all major stakeholder groups: parents, students, state officials, examination officials, the Press, politicians etc. However, such an endeavour would be prohibitively costly and time consuming, and it would need a very strong commitment by groups of researchers in all countries mentioned in this research. The most convenient solution was to gather published evidence through academic journals, state publications, the Press and the Internet. The evidence (data) collection process is thoroughly discussed in the next section.

Evidence (data) collection

The researcher spent a full academic year (from September 2006 to August 2007) collecting evidence from each of the case study countries. It was deemed necessary by the researcher to keep the evidence collection on-going throughout a whole academic year. It has often been observed that references to examination issues intensify just before and just after the examinations. A few weeks after the publication of the results the debate usually flattens off, and then intensifies again when the examination period approaches.

The evidence collected could be categorised as: Published academic papers (both printed and electronic), Press articles and releases (both printed and electronic), internet publications and reputable blog accounts, published Minutes from Parliaments and other Governmental publications or memos. Hundreds of pieces of material were reviewed during this year. As a result 109 documents were evaluated as relevant. All of this material was studied and was used to generate this report; however, only a third was included in the references for reasons explained below.

One reason is that much of this material was referencing other material: in these cases every effort was made to track the initial source of the information. Whenever this was possible, only the original (initial) source of the information is cited. In addition to this, material that does not contribute decisively to the discussion is not

presented here so as not to distract the reader with too much information and to allow the text to flow more smoothly. Finally, material that was not deemed 'reliable' (i.e. unsigned documents, internet postings of unclear origin etc.) was initially collected and studied (in case it led to other material) and then discarded (therefore not presented in the references).

Every care was taken to collect evidence both 'for' and 'against' the comparability methods used. In other words, if there were publications supporting the use of a statistical method then this material was studied and was included in the article (if that would facilitate the discussion). The official account of the organisation responsible for each of the comparability methods is clearly presented when this helps the discussion.

The researcher does not have illusions that he managed to exhaust all possible sources of information regarding these eight examples. It is, however, hoped that this study could spark local research and local researchers would surely be more successful in uncovering relevant material or arranging for interviews with major local stakeholders. The author of this paper hopes that this will be the beginning of a more cross-country comparative study of the issue of comparability between examination subjects. It is argued that by understanding the diversity of 'solutions' that have been given to the problem by various nations it will be easier for other countries, such as England, to understand their own problem and develop solutions.

Case studies 1–2: Australian New South Wales and the Tasmanian Qualifications Authority

The first example to mention here is Tasmania, a sovereign Australian state. Since the year 2000, the Tasmanian Qualifications Authority (TQA) has used an analysis of the relative difficulties of different subjects to establish the lower and upper limits of the score range for each award for each subject. The method used to estimate the degree of relative difficulty is the Rasch model, and is commonly referred to in the state as 'scaling' (TQA, 2007a).

In the context of the Rasch model, the 'test items' are the subject assessments and the underlying characteristic that is being estimated is the 'general academic ability' or 'merit to enter university' of students. The assumption is that once subject assessments (the examination subjects) are placed on a common scale (that is, the Rasch scale), their difficulties are directly comparable. And without first placing the subject assessments on a common scale, so the official story goes, they are not comparable and therefore should not be used to aggregate scores for each student.

The use of Rasch models to establish comparability of examination standards between subjects was pioneered many years ago by Jim Tognolini. Although the author of this paper can recall reading a long PhD thesis authored by Jim Tognolini many years ago, it was impossible to locate the exact references to this original work. However, Tognolini and Andrich (1996) provided a discussion of this work where they offered the first published account of why it is reasonable to use a Rasch model to scale examination subjects. However, both Rasch practitioners and the public

wonder about the practicality and the transparency of using such a complex model for so high stakes purposes, especially if the statistical requirements of the model are not always met.

However, because of the data-model fit requirements of the Rasch model, TQA and the University of Tasmanian are using an elaborate method to carry out the analysis. They generally monitor for 'misfitting' data and they remove them from the analysis. They calibrate the rest of the data by fixing ('anchoring') the parameters and then they re-introduce the misfitting data.

Describing the technical details regarding the use of the Rasch model in the context of the 'comparability between examination subjects' is beyond the scope of this paper. It suffices to present the basic equation of the Rasch model (as used by the TQA) and to discuss some basic properties and assumptions.

Tognolini and Andrich introduced the simple Rasch model in the context of subject comparability as follows,

$$\Pr(\mathrm{X_{ni}} = \mathrm{x}) = \frac{1}{\gamma_{\mathrm{ni}}} \exp\{\mathrm{x}(\beta_{\mathrm{n}} - \delta_{\mathrm{i}})\},$$

where β_n is the latent ability of person n, δ_i is the difficulty of item i, X_{ni}, $x \in \{0,1\}$ is a dichotomous variable and $\gamma_{ni} = 1 + \exp(\beta_n - \delta_i)$ is a normalising factor that ensures that $\Pr\{0\}+\Pr\{1\}=1$. They then explained how this could be extended to the Extended Rasch model to cover the case where students' performance on a subject is scored on a scale. In that case, the difficulty of the item becomes the difficulty of a subject, and instead of a dichotomous scoring, a partial credit scoring is applied.

The Rasch model demands that the subjects included in the Rasch analysis tap into the same psychological dimension or construct. Tognolini and Andrich (1996) responded that the requirement of the Rasch model for unidimensionality is not an absolute concept but relative: unidimensionality in empirical datasets holds to some degree, and this degree can be monitored and evaluated at any stage. Furthermore, to their critics who doubted the validity of aggregating 'raw scores' in order to derive the Rasch 'ability' estimate, Tognolini and Andrich responded that the 'raw scores' on the different subjects were already aggregated in Australia in order to derive a university admission index. Further discussion on the issue may be found in Tognolini & Andrich (1996), Coe (2007) and TQA (2007b).

New South Wales in Australia uses a different statistical scaling method to achieve the comparability between examination subjects. Average Marks Scaling is described by Coe (2007), who offers a long account of the model, along with the relevant equations. Average Marks Scaling is based on the so-called 'equal achievement' principle which says that when the same group of students ('common candidature') takes two (or more) same subjects, then the average performance of the group on the two (or more) subjects should be roughly the same. Using the 'equal achievement' principle, scaling compares the results of each group of students (the common candidature) in every possible pair of subjects. Scaling adjusts the raw scores in all subjects so that the scaled scores in the different subjects will be comparable.

The actual mathematics of the scaling process (as well as for the Tasmanian Rasch model) are rather cumbersome, so will not be presented here. Coe (2007), Partis (1997) and Seneta (1987) present all the necessary technical details, as well as the assumptions and the peculiarities of the method. What is interesting to discuss in this paper is how the public perceives the fairness and the effectiveness of complex statistical methods.

There is much discussion in the media, and much concern from parents and students who do not understand the details of complex statistical methods. This adds to their worries and suspicions that issues of unfairness may emerge. A simple search on the Internet will reveal a very large number of students in Australia 'blogging' (chatting on web blogs) or discussing in web forums the question of whether certain subjects are consistently scaled down. Students are also trying to devise methods to 'work' the system by taking specific (favourable) combinations of subjects. The issue is important and governmental agencies have published a number of leaflets trying to persuade parents and students that there are no issues regarding the comparability between subjects (e.g. Universities Admissions Centre, 2006) and to explain the mathematical process, though without much success. A simple search on the Internet will reveal various criticisms like Whitfield (2007), a respectable blogger who, after reading some relevant published material, maintained that 'the cynic in me sees some of the processes involved as pure mumbo-jumbo, and deeply contradictory as well'. This blog was characterised as 'A great resource for all students and teachers …' by M. Frances on the English Teachers Association Bulletin Board, 25 March 2005.

The Australian Universities Admissions Centre, in one of its publications (Universities Admissions Centre, 2002), urges students not to try to 'work the system' because they would get it wrong—obviously because students often do try to identify 'easy' subjects (that are said to be scaled down) in order to avoid them. It is interesting to observe that there are a number of 'frequently asked questions' that come up in governmental leaflets and discussions on the media, such as:

- Are there subjects that are always scaled down?
- Can I 'work the system' by choosing a specific combination of subjects?
- Is it true that if I study this course I can't get a high scaled score, no matter what my raw score is?

During one of the recent heated discussions in the Parliament about the issue of scaling (House of Representatives, 2006, pp. 5–6) the following discussion took place:

Member A of the committee:	… we have had quite a bit of evidence to say that students who are studying agriculture at an agricultural high school set up like [anonymised organisation], for example, their marks are downgraded—because it is an agricultural subject—when it comes to getting entrance marks to universities.
Chair of the committee:	The [scaling method is named].
Member B of the committee:	It is often not so much the downgrading of agriculture; it is often more that students do not have the opportunity or the

right advice to study science subjects. Realistically, agriculture is underpinned by science. We do have problems, in that students do not come in with sufficient chemistry, in particular, so we have a number of bridging courses and special courses to assist those students who do not have that background. We do our very best.

Member A of the committee: We have … had evidence that if you are studying chemistry you are likely to have your marks downgraded as well.

Member B of the committee: It depends on what you mean by downgrading.

Member A of the committee: You might get 100 per cent but, because of the marks of English, you might be scaled down to 92 or something. You would not get 100 per cent in chemistry, I would not think, but they are downgraded on that basis. So chemistry is actually one of the subjects that was very similar to agriculture, and that was downgraded.

Chair of the committee: We have certainly had significant evidence that there is deep concern about the [scaling] formula. I do not know what they call it in other states. It is resulting in what I could only refer to as a dumbing-down of agriculture, particularly for those students who, for whatever reason—intellectual capacity or whatever—are unable to get to that level. Because of the way in which their efforts are being treated through the system, they are getting a negative attitude about the possibilities of a future in agriculture.

…

Chair of the committee: There is certainly a very deep feeling out there on the issue.

In this context, the scaling method has been blamed for frightening the students away from specific subjects and the Technical Committee on Scaling (2002) has commented that 'each year brings its own myths and conspiracy theories'. It is likely that the conspiracy theories are totally unfounded—others say they are not. The fact is that the whole issue attracts too much attention and there seems to be a general feeling of unease among the public, parents and students as well as politicians.

Case study 3: the Fiji scaling system

The second case study of public distrust comes from the Fiji islands where a statistical method is used to scale the marks of students in different subjects in order to award a single aggregate score. The usual arguments, as in all countries, are often stated by officials in order to support existing policy:

> The Ministry of Education's scaling system is based on sound educational assessment principles used universally in ranking students for selection into tertiary education, scholarships and the like.
>
> In a statement, the Education Chief Executive Officer (CEO) [anonymised] said the scaling of marks is essential if they were to give students the opportunity to choose the subjects they wish to appear for in an external examination.

> …
>
> Scaling is brought in to achieve subject to equivalence or comparability based on a set complex formula, which is universally recognized and includes an overall mean and standard deviation. (Fiji Ministry of Education 2006, p. 1)

The Ministry of Education has tried hard to convince the public about the fairness of the method, giving examples of its widespread use abroad; alas, with little success. The minutes from a 2004 session of the Parliament read:

> Currently, there is a strong public debate on the scaling of marks and the general feeling is that it is a bad practice, politically designed to disadvantage students who perform well academically. […] as a concerned stakeholder, I had a workshop conducted the previous Saturday that was attended by 61 heads of schools and school managers respectively. The session on the scaling of marks by the Ministry of Education's Exam Office was very revealing, I must state. It had a formula that made little sense, factors no one knew how they were derived, except the raw marks gained by the students.
>
> The interesting part of scaling marks, apart from being a political exercise, is to bring about comparability in subjects and accountable environmental and social factors. This however, did not convince the participants that scaling was the means of bringing about the desired results. The participants strongly felt that the exercise was deceitful and concocted to provide a false picture for political expediency.
>
> […]
>
> We all want to give equal credit to all subjects, but not through deceitful means such as scaling the marks, which does not expose the actual performance of the students. The Ministry is urged to review the scaling process, to be fair to the students. (Parliament of Fiji, 2004, p. 14).

The heated discussions in the Parliament went on (Parliament of Fiji, 2005) with politically blatant and offensive language (e.g. 'wrapped mind', 'selective memory', 'bigoted and condescending mannerisms') being exchanged between the members of the Parliament. The Fiji Human Rights Commission was reported in the summer of 2006 to be investigating alleged breaches of human rights when students' external examination marks were scaled. According to Radio New Zealand International (2006), 'the Fiji Times reports that this follows complaints that the mark scaling system is unfair, non-transparent and violates the Bill of Rights in the Constitution and the Fiji Human Rights Commission Act'.

There is no reason to present here the technical details of the statistical model used in order to achieve the comparability between examination subjects. The interested reader may contact the Fiji Ministry of Education for more details. The important aspect of the discussion is that the Ministry of Education tried several times to explain the usefulness of the method in order to achieve comparability between examination subjects. It even cited the use of such models by educationally advanced countries (e.g. Australia) to persuade the public that nothing wrong is being done in Fiji. Public distrust increased, however, and the whole issue turned dangerously political with allegations of violations of human rights. One interesting observation is that issues of

comparability of examination subjects have turned political in all the case study countries mentioned in this study.

Case studies 4–5: Singapore's T-score and Cyprus's z-score methods

In Singapore, pupils are ranked according to their performance in all the subjects at the PSLE examination. According to the web page of the Ministry of Education (http://www.moe.gov.sg/forum/2005/20051018a.htm) one way of doing this could be to add up all the raw marks obtained by a pupil. However, as it is argued, this would not be desirable as the subject with the largest variance of raw marks would have excessive influence on the rank order of pupils. According to the Ministry of Education, to solve this problem, the raw marks in a subject are converted to T-scores so that the mean and standard deviation for all subjects becomes 50 and 10 respectively. The T-score for all the subjects are then added up to give a pupil's Aggregate Score.

Singapore, like other countries, has seen a stream of queries and worries from stakeholders concerning their scaling system. It is indicative that the Ministry of Education of Singapore posted on their web page an answer to the queries of a mother who was worrying that the existing scaling method could lead to different subjects having different weightage in pupils' aggregate score. On the other hand, the Ministry of Education explained that, 'Computing the Aggregate Score in [the manner explained above] underscores the equal importance placed on pupil performance in each PSLE subject'.

The same problem occurs in Cyprus, only to a greater extent, where word of mouth, the newspapers, television and heated discussions in Parliament have convinced parents and students that there are specific subjects which are always 'scaled down' and should be avoided.

The Cyprus Testing Service (CTS) uses a z-score transformation to convert the raw scores of the candidates taking external university entrance examinations at the end of the 12th year of schooling. Every pupil's scores within a subject are z-transformed and then rescaled with a standard deviation of three and mean of 10. All the rescaled scores of every candidate (one for every subject that had been examined) are aggregated to form the final score that is used for sorting and selection purposes.

Scary stories of how scaling in Cyprus affects unfairly the prospects of bright students reaching university hit the media every year in the summer; just before and after the examinations. There is not much scope here to give exact references since everything is written in Greek, but for the Greek-speaking readers of this study, a simple search on the internet using the keywords 'Παγκύπριες εξετάσεις, αναγωγή' (Pancyprian examinations, scaling) will reveal a wealth of references. However, everybody that speaks Greek and lives in Cyprus must already be fully aware of the periodically heated discussion.

No matter whether the criticisms are well founded or not, no matter how hard the CTS is trying to alleviate these criticisms, and no matter whether CTS has introduced

improvements to its scaling methods recently, the side effects of public distrust are devastating. One fact, for example, is that chemistry, a subject notorious among parents and students for 'consistently' being scaled down, has seen its enrolment dropping by 70% from 2001 to 2006 (source: The Cyprus Ministry of Education web page, www.moec.gov.cy).

The root of the problem mainly lies in the type of pupils that actually choose to be examined on each subject. Biology is taken by more students wishing to study in Gymnastic Academies (usually students with lower general academic performance) but is also taken by many bright students with excellent general academic performance wishing to study medicine. Those who wish to become doctors prefer to take biology than chemistry because they are more likely to excel—the large number of low achieving students wishing to study in Gymnastic Academies makes biology look very 'difficult'—so all grades are scaled up and especially those students who excel academically.

No scaling or common examine method would manage to take account of the very different effects that appear when two groups so evidently different take the same exams—except perhaps if different scaling were used for different candidates; but how could this be politically supported? The CTS has problems in persuading the public and politicians about the fairness of the existing method (which one MP has recently characterised as 'complicated formulae … that no one will understand'), so how might somebody try to convince them that different methods should be used for different students?

To make things even worse, CTS does not have established mechanisms to prepare tests on different subjects that have exactly similar distribution of raw scores, owing to restrictions in pre-testing (for security reasons) and to the low number of candidates in certain subjects. Thus, the z-score transformation could, in some cases, indeed cause (and may have caused) atypical conversions from raw to scaled scores, producing a feeling of unfairness in some candidates and parents. This is probably what Singapore's Ministry of Education meant when they warned that '… for this method of computing Aggregate Score to work well, the mean and standard deviation of the raw marks for a given subject should not be too extreme. They should neither be too high nor too low'. It is possibly the case that ordinary people (with minimal technical knowledge), in Cyprus as well as in Singapore, have already figured this out.

Case study 6: New Zealand's scaling (now abolished) system

The previous examples have stressed the lack of public trust and the negative effects of scaling methods on the examination system of some countries. The international arena has examples of the opposite effect as well: there are situations where abolishing statistical scaling caused a new cycle of discussions and complaints by the stakeholders. This is the situation in New Zealand, where statistical scaling had faced criticisms in the past but when it was abolished the new system was received with many complaints and unpleasant surprises.

During the last 40 to 50 years up to 1990, students in Years 11–13 (the final three years of secondary school) had generally been taking national qualifications in the form of national end-of-year examinations (as well as moderated school-based assessments or a combination of both). According to Crooks (2002):

> All of [the] national examinations were normative in nature, with percentage or numerical quotas for passes or awards and with inter-subject statistical scaling designed to ensure that the cluster of candidates attempting any pair of subjects gained similar distributions of marks in both subjects. This scaling produced very different mark distributions for different subjects, but was established because otherwise students who took the most academically elite subjects (such as Latin) generally gained lower marks in that subject than in their less selective subjects. (p. 248)

By 1990, however, things changed significantly. The national examinations still existed and the inter-subject scaling was duly conducted; however, students received grades for each paper separately, with no aggregate requirement to pass as a whole. On the other hand, Year 13 students were offered the University Bursaries and Scholarships examination and a subject aggregate was needed to gain a tertiary study bursary.

The transformation from the raw marks to the scaled marks was difficult to explain to the general public, and there was much debate among the various stakeholders. One of the most notable and interesting changes, as a result, was the initiative of the then Minister of Education, Wyatt Creech, who stated that a new system was under consideration where:

> There [would] be no inter-subject scaling of the external assessment. This [had] been a source of dissatisfaction for some time because of the distortions that occur[ed] between raw and final marks. Such scaling [had been] necessary in order to aggregate marks for the award of A and B bursaries. Meaningful comparison between subjects [would] still be possible using percentile scores. (National Office, 1999, p. 12)

The new system was in place for some time until, in 2004, criticisms of very large inter-subject (and inter-year) variability in results sparked a governmental inquiry into both NCEA and New Zealand Scholarship. The result of the enquiry by the State Services Commission was critical of New Zealand Qualification Authority's (NZQA) performance. The report recommended specific 'safety nets' which could help to avoid excessive variation between subjects without a suitable explanation (a measure that was felt by some to be a step back to norm-referencing) (Wikipedia, 2007).

The result of the mess that was created in New Zealand was described in a nutshell by Nash (2005):

> …when in early 2005 NZQA published Scholarship results markedly inconsistent between subjects, its initial response was to defend them as an accurate reflection of the standards achieved, and defend the integrity of the system. This position was only abandoned in the face of political and public pressure sustained over a period of several weeks. The situation proved to be so unacceptable, however, that Cabinet itself sanctioned the issue of revised grades and, after official investigations, both the Chief Executive and the Chair of the Board resigned their positions. (p. 103)

The main story of this case study is that New Zealand, under the pressure of criticisms, abolished its scaling system in a context where students were examined on different subjects. Instead, a criterion-referenced 'standards' system was established aiming at abandoning the 'norm-referenced' paradigm once and for all. The end result was that there was so much variability between subjects' perceived 'difficulty' that the whole country was discussing the issue for many weeks. How hot the social and political discussion was may only be revealed by the following extract from the proceedings of the New Zealand Parliament:

Member of Commons:	Is the Minister now saying that all the statements he made—both inside the House and outside—about not knowing about [subject-specific 'difficulty'] variability until 13 January were wrong, and that, in fact, he was discussing this issue on 15 December 2004; and why do he and his fellow Minister insist on tawdry cover-ups and half-truths all the time, instead of just owning up?
Minister of Education:	The discussions on the 13th were around overall pass rates, not about variability.
Another Member of Commons:	What's the difference?
Minister of Education:	The member asked what the difference is—that just shows how thick he is.

(New Zealand's House of Representatives, 10 May 2005)

One might be tempted to say that scaling all subjects to have the same pass rates might be a solution to the issue. This, however, could be simplistic and would again spark a new round of discussions and debates, suggesting an unacceptable retreat towards heavy-handed normative methods.

Up to now this study has presented cases both against and in favour of scaling methods. The next case study (Scotland) is a neutral example of the use of statistical scaling methods to establish the comparability of examination subjects, but without directly affecting the reported scores of the students.

Case study 7: the Scottish Qualification Authority's national ratings

The problems of Cyprus and New Zealand where specific subjects (e.g. chemistry) are said to have suffered because of perceived scaling biases are not new. In accord with previous research (e.g. Fitz-Gibon & Vincent, 1994), Kent (1996) suggested that the relative 'difficulty' of modern foreign languages deterred students in Scotland from studying those subjects. He suggested that:

> The impression of the students is that it is more difficult to obtain good grades in foreign languages than in other subject areas. Objective evidence, provided in the national ratings produced by the Scottish Examination Board … show modern foreign languages to have a higher 'difficulty' level than many other subjects.

Indeed, national statistics showed (at least at the time) that the number of students taking languages was dropping. The national ratings published by the Scottish Examination Board may have played their role, since languages indeed

appeared to be among the most difficult subjects. Kent mentioned that 'Latin, German and French topped the table with comparability indices of –0.62, –0.50 and –0.49 respectively'. The meaning of these negative indices cited in the previous sentence will become clearer by explaining the concept of 'national ratings'. Recognising the issue of differential subject difficulty in Scotland, a scaling method has been employed to facilitate the comparability between examination subjects.

Each year national ratings (as the subject difficulty measures are more elegantly termed) are published for each subject offered for examination. The national rating for a subject provides an indication of the measure of the average performance of students nationally in that subject, in comparison with the average grade achieved by the same students across their other subjects.

The value of a national rating in Scotland is expressed in terms of grades (this refers to the Curriculum levels). Thus, a subject with a national rating of +0.50 indicates that students have scored half a grade higher than the average of their other subjects—thus, an easier subject. For a subject with a national rating of –0.50, students have scored half a grade lower than the average for their other subjects—thus a more difficult subject.

Having explained the above, it becomes obvious that in 1993, Latin, French and German were at least half a grade more difficult than other subjects. Is it possible that the differential difficulty as demonstrated by the national ratings caused the reduction in the number of students taking languages?

In the same context, four years after Kent, Sparkes (2000) concluded that differential subject 'difficulty' was not an important factor in determining students' choice of subject options. As a matter of fact, the statistics prepared by Sparkes (2000) tell the same story as the statistics prepared by Kent (1996). Sparkes suggested that:

> There is confirmation in this investigation of the findings of Fitz-Gibbon and Vincent; the relative 'difficulty' of the different subjects is similar to both A levels [English qualifications] and Highers [Scottish qualifications] and candidates of equivalent ability do get worse grades in modern foreign languages (with the exception of Spanish) and the mathematical and science subjects than the other non-science subjects. (p. 188)

However, Sparkes (2000) maintained that the decline in the number of candidates for certain subjects has more to do with the career paths of the students as dictated by the general economic environment than with their 'difficulty' index. Which of the two stories to believe is up to anyone to decide, although the situation may be more complex than it seems. In Cyprus, in the case study mentioned above, changes in the career paths of the students may have actually contributed to the huge decline in students enrolling for chemistry, although there are no official reports on the issue. Was this the case in Scotland as well?

In any case, the Scottish paradigm instructs that a whole 'value added' and 'league tables' industry has evolved on the foundations of the concept of the national ratings of examination subjects. Concepts like 'gender ratings' (if subjects are more or less

difficult for boys or girls), 'school ratings' (schools being more or less effective) or based on the concept of national ratings, and league tables are regularly produced. Although opposition in Scotland to these concepts is nowhere close to the public unrest in other countries, the concept of national ratings (and its derivatives) is frequently receiving its fair deal of criticism.

The fact is that Scotland may have managed to strike a balance between exhibiting the differential (statistical) difficulty of examination subjects, without directly affecting reported student scores, thereby alleviating the criticism and the distrust. What university admission officers do with those national ratings may be an altogether different story, but the final effect on the actual value of the qualifications of the students may be comparable to what would happen if Scotland had adapted a more direct statistical inter-subject moderation—only with more criticism and public unrest.

Case study 8: the counter-example of the Pan-Hellenic university entrance exams

An extreme case of conformity might be that all students take roughly the same tests on the same subjects at the end of the year in order to compete for access to higher education. There is much discussion in Greece every year about the 'difficult' and the 'easy' subjects for the Pan-Hellenic university entrance exams (Pan-Hellenic = for the whole Greek nation; Greeks are also called Hellenes and Greece is also called Hellas). A large number of examination subjects are considered as 'very difficult' whereas other subjects are considered to be 'softer'.

However, the problem of comparability between subjects has been solved in a very practical way: large groups of students study the same subjects at school, and are therefore examined on the same subjects. For example, students aspiring to study medicine normally take the same examination subjects in the Pan-Hellenic exams. Similarly, those aspiring to become engineers normally take the same examination subjects, and so on. Therefore, the Ministry of Education compares like with like, and the need to re-scale different subjects for comparability purposes is decreased.

Although the actual examination system in Greece is a lot more complex than presented here for communication purposes, it is a useful simplification to say that the Hellenic society and the Ministry of Education chose to 'fix' the examinable subjects, so that students with very diverse subjects would not compete for the same university places. To the degree that the Hellenes (= the Greeks) are happy, one wonders whether it might be better to offer students the opportunity to be examined on different subjects, and give space to comparability worries arising at the same time.

Could anyone complain about the lack of pluralism or democracy in Hellenic schools? Some people do complain, however, when it comes to university entrance issues, since the Greeks are very serious about issues of fairness and equal opportunities. The Greek society decided that a simple and transparent system which does

not raise issues of fairness is more important than a 'perceived pluralism' in selecting examination subjects.

Discussion and conclusions

Every society usually seeks to hit a balance between democracy, pluralism (i.e. allowing students to choose among different subjects) and practicality (i.e. aggregating scores) in order to make educational decisions.

The example of Greece shows that once societies strike a 'consensus' on a working model of examinations (and usually the simpler the examination system the better), they usually stick with it, even though some criticisms may arise from time to time.

Part of the problem in some countries may be caused by the brave and over-optimistic comments of some politicians or policy makers about their preferred scaling method. An example of such a brave (and very recent) comment is the New Zealand Education and Scholarship Trust's statement (on their web page) that they '… ensure comparability between subjects …' (Farrel, 2007); but anyone who has worked in the area of between-subject comparability knows that this is a very courageous statement to make—especially in the New Zealand context.

In the English context, Henry (2006) cites the comments of the director of the Joint Council for Qualifications (JCQ) saying, 'There is no such thing as an easier subject. The comparability of subjects is strictly controlled and there are ongoing studies to look at the degree of difficulty.' Whether the comparability of subjects is currently strictly controlled in England or not, this is of course debateable. The public receives these statements well, but then the Press publishes statistics that make people feel that they have been deceived. Obviously, raising the expectations of the public is a doomed attempt, as history has shown.

It is the view of the author of this commentary that comparability is a matter of perceived fairness on behalf of the society; it is hugely a political issue, not a statistical nor a theoretical one. Societies know that taking tests in different subjects is surely a different and non-comparable experience. There needs, however, to be a balance between solving practical problems (i.e. aggregation of scores for selection purposes) and perceived fairness.

When the patterns are clear in the data, many of the statistical methods may give practically similar results. For example, Figure 1 shows the agreement between the method used by the Scottish Qualification Authority to compute the National Ratings and the Partial Credit Rasch model used by the Tasmanian Qualification Authority to compute the relative difficulty of subjects. Both methods were applied by the author on year 2006 data from a European country. Each star is an examination subject.

The main idea of this paper is that we need to be brave and step forward, explaining to the public that statistical comparability is not an exact science, and that it may have some side-effects. Complex methods may be more difficult to justify and explain. For example, Manly (1988) argued that 'there may be a good reason to make different

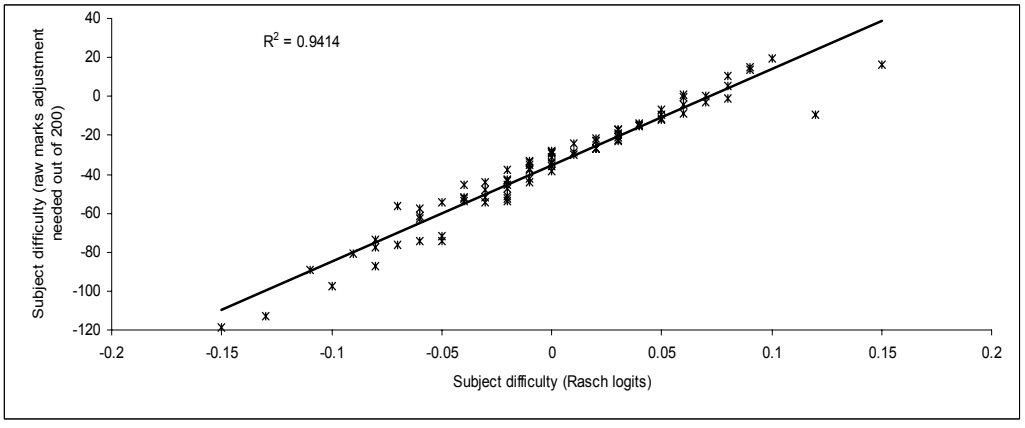

Figure 1. The Scottish (vertical axis) and the Tasmanian (horizontal axis) methods agree

types of scalings with different subjects' (p. 389). But to what extent would such a radical approach be supported by the public?

Education, including examinations, is a purely political act. This paper claims that any type of scaling (or even not scaling at all) will do, as long as the decision has the support of the major stakeholders. But being able to explain the decision clearly to parents and students will make it easier for everyone to understand (and also understand the possible shortcomings) and will possibly alleviate opposition.

There is a pattern in the case studies above: (a) the comparability methods in routine use are difficult for the layman to understand; and (b) there are beliefs for allegedly unfair effects. But is there a relationship between these two issues and the three conceptions of comparability suggested by Coe (2007)?

It is difficult to see how students or parents could argue in favour of the 'phenomenal' conception: being examined on different subjects, for instance dance or chemistry, is obviously a non-comparable experience. It is very hard to identify in the press or in international literature arguments advocating that 'phenomenal' comparability between such different subjects is possible to achieve. Such a claim would probably tap into a complex web of specifications and assumptions regarding each one of the subjects. If we adapt a 'phenomenal' conception of comparability, we would probably need to explain a lot about the intellectual demand made by each examination subject, therefore hitting both of the main problems mentioned in this study: our comparability method would be difficult for anyone to understand and to defend.

Having said this, however, the stakeholders in the case studies mentioned above seem to agree that there is something common between all subjects—albeit 'phenomenally' different—so that one is justified in aggregating scores on different subjects for ranking purposes. This has been regularly done in many countries around the world. For example, in the USA, many people use the Average Grade Point (an average of a student's scores in many subjects) as an indication of academic capacity.

It seems that when the public is faced with the need to make direct comparisons between candidates to allocate scarce educational resources, the 'construct' conception (which assumes that a similar 'academic ability' student dimension practically underpins all subjects) is easier to endorse. The problem, though, with the construct conception of comparability is that it is best (though not solely) served by complex statistical techniques like the Rasch model (used by the Tasmanian Qualifications Authority since 2001) or the AMS method. Complex methods are incomprehensible to the layman and it is frequently the case that extreme examples of 'unfair' scaling results may be put forward (e.g. the chemistry issue mentioned above), that will erode public trust in the system. Both the Rasch model and the AMS method, for example, also imply that the success rate will not be the same in different subjects, therefore generating allegations of unfairness, tapping the concept of 'statistical' comparability.

The statistical conception of comparability initially looks deceptively easier for parents and students to endorse. This, however, was a hard lesson learned in the case of New Zealand. Intuitively the public will question the results when the pass rates on different subjects are widely different; this is what happened in New Zealand and there was much public unrest and much political debate—sometimes less than polite.

So would keeping the grade distribution similar on each subject solve the problem of perceived unfairness? Such an approach to comparability could probably involve simple statistics so the first problem (that the layman does not understand the statistics) could be solved. However, the public would probably fail such an approach on the grounds of 'case study examples', i.e. bright students being awarded very different grades depending on the popularity of the subject and the competition within it. As has been found in the case of chemistry in Cyprus, in more popular subjects the competition for higher grades will be fierce whereas students may find it easier to achieve high grades on less popular subjects. Therefore, such an approach would probably pass the test of simplicity, but would fail the 'perceived fairness' criterion.

To sum up, solving the Gordian knot of comparability of examination subjects is very difficult. The public seems to have a dual approach: they endorse the construct approach in general, while they hesitate to drop the statistical approach altogether. Drawing on the experience of other countries, England and other interested countries must be very careful before proceeding with further (and probably formal) use of comparability methods. Though pragmatism should prevail when attempting to solve practical social/educational problems, it is important to get a general consensus from the major stakeholders before making the next step.

It is also important to make explicit to the public that any statistical scaling will probably have specific—usually undesirable—side effects. It is important to 'educate' the public that whenever anyone 'tampers' with the raw scores of individuals, it is very likely that 'unfairness' issues may emerge. Courageous statements that raise the expectations of the public do not usually help.

While discussing the issue of lack of comparability between examination subjects in England, Newton (1997) argues:

> The recommendation is that we should learn to accept and adapt to the unintelligible enigma of comparability between subjects. This seems all the more sensible when it is realised how unclear it is that adjustment would be desirable even if 'genuine' differences could actually be found. (p. 448)

In conclusion, the author of this study finds it very difficult to take a clear position either for or against the use of statistical methods in order to achieve comparability between subjects. The decision should always be taken with care and after a lot of consultation between the responsible bodies or stakeholders. Before any specific statistical model is routinely applied, it is necessary to study the possible side-effects on empirical data from previous years.

It is arguably better to establish clear-cut rules about access to tertiary education, than to try to achieve comparability between subjects using opaque statistical methods. The case of Greece is instructive: they reduced the perceived 'pluralism' but they increased the perceived 'fairness'. It is up to local societies to decide for themselves. All we, the researchers, can do is to provide the relevant information in our papers to help them decide. After all, as ancient Greek philosophers used to say, education is a purely political act, and there are no clear-cut rights and wrongs.

Notes on contributor

Iasonas (Jason) Lamprianou has taught advanced statistical modelling (Rasch analysis) at postgraduate level at the University of Manchester, Faculty of Education, and has given lectures at the University of Malta, at the Agha Khan University Examination Board (Pakistan) and at the Pedagogical Institute of Cyprus. He is currently supervising PhD students in the area of educational measurement. He has participated in research projects involving educational assessment and measurement; test equating, Item Banking, Computerized Adaptive Tests. In the last years he has offered his services to many organisations dealing with assessment. He has published in diverse education journals.

References

Bell, J. F., Malacova, E., Rodeiro C. L. V. & Shannon, M. (2007) A-level uptake: 'Crunchier subjects' and the 'Cracker effect', *Research Matters,* 3, 19–25.

Coe, R. (2007) Common examinee methods, in: P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds) *Techniques for monitoring the comparability of examination standards* (London, QCA).

Crooks, T. J. (2002) Educational assessment in New Zealand schools, *Assessment in Education,* 9(2), 237–253.

Curriculum Council (2006) *Syllabus manual volume I. General information 2006–2008* (Section 10, pp. 57–66). Available online at: http://www.curriculum.wa.edu.au/pages/syllabus_manuals/volumes/II_english/syllabus_manuals_0608/pdf/Cover%20sheet.pdf http://www.curriculum.wa.edu.au (accessed 13 June 2007).

Cyprus Minister of Education. (2006) Press release, 14 July. Available online at: http://www.cyprus.gov.cy/moi/pio/pio.nsf/All/F2277BE6883F716EC22571AB003B058D?Opendocument&print (accessed 12 July 2007).

Farrel, M. (2007) *The aural component in language examinations.* Available online at: www.nzest.ac.nz/academicCom.html (accessed 13 February 2007).

Fiji Ministry of Education (2006) Ministry defends education scaling system. Available online at: www.fiji.gov.fj/publish/page_7352.shtml (accessed 1 September 2006).

Fitz-Gibbon, C. T. & Vincent, L. (1994) *Candidates' performance in public examinations in mathematics and science* (SCAA Report) (Newcastle upon Tyne, CEM).

Henry, J. (2006) *A-levels have not got easier, says exam chief.* Available online at: http://www.telegraph.co.uk/news/1526213/A-levels-have-not-got-easier-says-exam-chief.html?mobile=true (accessed 20 March 2007).

House of Representatives (2005) *Questions for oral answers, section: New Zealand Certificate of Educational Achievement—scholarship.* New Zealand's House of Representatives. Available online at: http://www.hansard.parliament.govt.nz/Documents/20050510.pdf (accessed 13 February 2007).

House of Representatives (2006) *Standing Committee on Agriculture, Fisheries and Forestry* (Commonwealth of Australia, House of Representatives, Armidale, Australia).

Kent, D. M. (1996) *An investigation into the factors influencing the learning of foreign languages in S5 and S6 Scottish schools.* Available online at: http://www.scre.ac.uk/scotresearch/kentinves/index.html (accessed 14 April 2007).

Lamprianou, I. (2007) Commentary on Chapter 9, pp. 385–388, in: P. Newton, J.-A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds) *Techniques for monitoring the comparability of examination standards* (London, QCA).

Manly, B. F. J. (1988) The comparison and scaling of student assessment marks in several subjects, *Applied Statistics,* 37(3), 358–395.

Nash, R. (2005) A change of direction for NCEA: on re-marking, scaling and norm-referencing, *New Zealand Journal of Teachers' Work,* 2(2), 100–106.

National Office (1999) *Achievement 2001, letter from the Secretary for Education to the school Principals.* Available online at: http://www.minedu.govt.nz/web/downloadable/dl4800_v1/sfe-sept99.pdf (accessed 3 September 2006).

Newton, P.E. (1997) Measuring comparability of standards between subjects: why our statistical techniques do not make the grade, *British Educational Research Journal,* 23(4), 433–449.

Newton, P., Baird, J., Goldstein, H., Patrick H. & Tymms P. (2007) *Techniques for monitoring the comparability of examination standards* (London, QCA).

Parliament of Fiji (2004) *Parliamentary debates, 29 July. House of Representatives Daily Hansard.* Available online at: www.parliament.gov.fj (accessed 13 November 2007).

Parliament of Fiji (2005) *Parliamentary debates, 21 November. House of Representatives Daily Hansard.* Available online at: www.parliament.gov.fj (accessed 13 November 2007).

Partis, M.T. (1997) *Scaling of tertiary entrance marks in Western Australia* (Osbourne Park, WA, Western Australia Curriculum Council). Avaliable at http://www.curriculum.wa.edu.au/files/pdf/114537_1.pdf

Radio New Zealand International (2006) *Fiji Human Rights Commission to investigate exam scaling.* Available online at: http://www.rnzi.com/pages/news.php?op=read&id=26508 (accessed 3 March 2007).

Rodeiro C. L. V. (2006) *Uptake of GCE A-level subjects in England 2001–2005.* Statistics Report Series No. 3, Research Division – Statistics Group Assessment Research and Development, Cambridge Assessment, Cambridge. Available online at: www.cambridgeassessment.org.uk/ca/digitalAssets/113439_A-level_uptake_2001-2005.pdf (accessed 10 October 2007).

Seneta, E. (1987) *The University of Sydney scaling system for the New South Wales higher school certificate: a manual* (Sydney, Department of Mathematical Statistics, University of Sydney).

Singapore Examinations and Assessment Board (2006) *Forum Letter Replies.* Available online at: http://www.moe.gov.sg/media/forum/2005/20051018a.htm

Scottish Qualifications Authority (2007) *Standard tables and charts.* Available online at: http://www.scotland.gov.uk/library3/education/stac-16.asp (accessed 13 February 2007).

Sparkes, B. (2000) Subject comparisons—a Scottish perspective, *Oxford Review of Education*, 26(2), 175–189.

Tasmanian Qualifications Authority (2006) *An introduction to Rasch Modelling and the TCE*. Available online at: http://www.tqa.tas.gov.au/4DCGI/_WWW_doc/003675/RND01/Rasch_intro.pdf (accessed 7 December 2006).

Tasmanian Qualifications Authority (2007a) *How the scaled awards are calculated and used to determine the tertiary entrance score*. Available online at: www.tqa.tas.gov.au/0477 (accessed 13 February 2007).

Tasmanian Qualifications Authority (2007b) *Frequently asked questions about the TCE*. Available online at: http://www.tqa.tas.gov.au/1278 http://www.tqa.tas.gov.au/1278#scaling (accessed 12 February 2007).

Technical Committee on Scaling (2002) *Report of calculation of universities admission index 2001, for Committee of Chairs*. Available online at: http://www.secretariat.unsw.edu.au/acboard/committee_chairs/tcsrep302a.pdf (accessed 12 March 2007).

Tognolini, J. & Andrich, D. (1996) Analysis of profiles of students applying for entrance to universities, *Applied Measurement in Education*, 9(4), 323–353.

Universities Admissions Centre (2002) *The universities admissions index 2001*. Available online at: http://www.secretariat.unsw.education.au/acboard/committee_chairs/tcsrep302b.pdf (accessed 2 June 2008).

Universities Admissions Centre (2006) *You and your UAI: a booklet for 2006 NSW HSC students*. Available online at: http://www.uac.edu.au/pubs/pdf/uaibook_2006.pdf (accessed 25 April 2007).

Whitfield, N. (7 January 2007) The mysteries and injustices (?) of the NSW UAI ranking. Blog transcript. Available online at: http://neilwhitfield.wordpress.com/2007/01/07/the-mysteries-and-injustices-of-the-nsw-uai-ranking/ (accessed 13 February 2007).

Wikipedia, National Certificate of Educational Achievement. Available online at: http://en.wikipedia.org/wiki/National_Certificate_of_Educational_Achievement, (accessed 2 August 2007).